

MATH 1260: Significant Statistics

MATH 1260: Significant Statistics

An Introduction to Statistics

*ADAPTED BY JOHN MORGAN RUSSELL; FROM BARBARA ILLOWSKY
AND SUSAN DEAN, DAVID DIEZ, MINE CETINKAYA-RUNDEL AND
CHRISTOPHER D. BARR; AND JULIE VU AND DAVID HARRINGTON*

DEPARTMENT OF STATISTICS IN AFFILIATION WITH THE UNIVERSITY LIBRARIES AT VIRGINIA TECH
BLACKSBURG, VA



MATH 1260: Significant Statistics by John Morgan Russell, OpenStaxCollege, OpenIntro is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/), except where otherwise noted.

© 2021 OpenStax, OpenIntro, and John Morgan Russell. Textbook content produced by OpenStax is licensed under a [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by-sa/4.0/) license. You can access the original version of this textbook here: <https://openstax.org/details/books/introductory-statistics>; Introductory Statistics: OpenStax. Content from OpenIntro is licensed under a [Creative Commons Attribution ShareAlike License 3.0](https://creativecommons.org/licenses/by-sa/4.0/). The overall license of this book is [Creative Commons Attribution ShareAlike License 3.0](https://creativecommons.org/licenses/by-sa/4.0/).

Under this license, any user of this textbook or the textbook contents herein must provide proper attribution as follows:

The OpenStax name, OpenStax logo, OpenStax book covers, OpenStax CNX name, and OpenStax CNX logo are not subject to the creative commons license and may not be reproduced without the prior and express written consent of Rice University. For questions regarding this license, please contact support@openstax.org.

- If you use this textbook as a bibliographic reference, then you should cite it as follows:
OpenStax, Introductory Statistics. OpenStax CNX. Jun 17, 2019 <http://cnx.org/contents/30189442-6998-4686-ac05-ed152b91b9de@23.31>.
- If you redistribute this textbook in a print format, then you must include on every physical page the following attribution:
Download for free at <http://cnx.org/contents/30189442-6998-4686-ac05-ed152b91b9de@23.31>.
- If you redistribute part of this textbook, then you must retain in every digital format page view (including but not limited to EPUB, PDF, and HTML) and on every physical printed page the following attribution:
Download for free at <http://cnx.org/contents/30189442-6998-4686-ac05-ed152b91b9de@23.31>.

Contents

Introduction	xvii
<i>Letter To Students</i>	xvii
<i>Learning Objectives</i>	xvii
<i>Coverage and scope</i>	xviii
For Instructors	xix
<i>Letter to Instructors</i>	xix
<i>Supplemental Material</i>	xx
<i>Learning Objectives</i>	xx
<i>Coverage and scope</i>	xx
Attribution	xxii
<i>Sources and Workflow</i>	xxii
<i>About the Adapter</i>	xxii
<i>About the Editorial Team</i>	xxiii
<i>Project Funding</i>	xxiii
Chapter 0 Wrap-Up	xxiv
<i>Concept Check</i>	xxiv
<i>Section Reviews</i>	xxiv
<i>Key Terms</i>	xxiv
<i>Extra Practice</i>	xxv

Chapter 1: Sampling and Data

1.1 Introduction to Statistics and Key Terms	3
<i>Introduction</i>	3
<i>The Study of Statistics</i>	4
<i>Probability</i>	5
<i>Key Terms</i>	5

1.2 Data Basics	8
Types of Data	8
Levels of Measurement	10
Variation in Data	11
Data Analysis	12
1.3 Data Collection and Observational Studies	14
Data Collection Methods	15
Anecdotal Evidence	15
Observational Studies	15
1.4 Designed Experiments	18
Observational Studies vs. Experiments	18
Designed Experiments	18
More Experimental Design	21
1.5 Sampling Techniques and Ethics	25
Sampling	25
Simple Random Sampling	25
Other Sampling Techniques	26
Sampling and Replacement	28
Bias in Samples	29
Variation in Samples	29
Size of a Sample	30
Critical Evaluation	30
Ethics	31
Chapter 1 Wrap Up	34
Concept Check	34
Section Reviews	34
Key Terms	36
Extra Practice	38
References	66

Chapter 2: Descriptive Statistics

2.1 Introduction to Descriptive Statistics and Frequency Tables	71
Descriptive Statistics	72
Frequency Tables	73
Relative Frequencies	74
2.2 Displaying and Describing Categorical Data	79
Descriptive Statistics for Categorical Data	79
Graphical Methods for Categorical Data	79
Pie Charts	79
Bar Graphs	80
Pie vs. Bar Charts	81
Describing Categorical Data	88
Mode	88
Variability	90
2.3 Displaying Quantitative Data	93
Descriptive Statistics for Quantitative Data	93
Graphical Methods for Quantitative Data	93
Stem-And-Leaf Plots	93
Line Graphs	96
Dot Plots	98
Histograms	98
Frequency Polygons	101
Time Series Plots	103
2.4 Describing Quantitative Distributions	107
Key Aspects of Quantitative Data	107
2.5 Measures of Location and Outliers	114
Percentiles	114
Quartiles	117
Five Number Summary	120
Interquartile Range	121
Fence Rule	121
Box Plots	122

2.6 Measures of Center	126
The Mean	126
The Median	127
The Mode	128
Order Relationship of Measures of Center	129
Calculating the Mean of Grouped Frequency Tables	134
2.7 Measures of Spread	138
The Interquartile Range	138
The Standard Deviation	138
The Standard Deviation in Context	142
Z-scores	145
Identifying Unusual Values with the Standard Deviation	146
Chapter 2 Wrap Up	148
Concept Check	148
Section Reviews	148
Key Terms	151
Extra Practice	153
References	226

Chapter 3: Basics of Probability

3.1 Introduction to Probability and Terminology	233
Probability	234
Solving Probability Problems	241
3.2 Visualizing Probabilities	243
Contingency Tables	243
Tree Diagrams	245
Venn Diagram	249
3.3 Compound Events	253
Finding Probabilities of Unions	253
Applying the Addition Rule to Multiple Events	258
Finding Probabilities of Intersections	259
Finding a Conditional Probability	260

Chapter 3 Wrap Up	263
Concept Check	263
Section Reviews	263
Key Terms	264
Extra Practice	266
References	327

Chapter 4: Discrete Random Variables

4.1 Introduction to Discrete Random Variables and Notation	333
<i>Random Variables</i>	334
<i>Discrete Random Variables</i>	334
4.2 Measures of General DRVs	339
<i>The Expected Value (Mean) of a Discrete Random Variable</i>	339
<i>The Variance and Standard Deviation of a Discrete Random Variable</i>	341
4.3 The Binomial Distribution	345
<i>The Binomial Setting</i>	345
<i>Notation for the Binomial</i>	348
<i>Binomial Probability Function</i>	348
Chapter 4 Wrap Up	353
Concept Check	353
Section Reviews	353
Key Terms	354
Extra Practice	355
References	384

Chapter 5: Continuous Random Variables

5.1 Introduction to Continuous Random Variables and The Uniform Distribution	387
<i>Properties of Continuous Probability Distributions</i>	389
<i>Some Continuous Distributions</i>	389
<i>Probability Density Functions</i>	390
<i>The Uniform Distribution</i>	390

5.2 The Normal Distribution	397
The Empirical Rule	398
Finding Normal Probabilities	400
The Standard Normal Distribution	400
The standardizing process	403
The “un-standardizing” process	404
5.3 The Normal Approximation to the Binomial	407
Binomial Approximation Conditions	408
The Continuity Correction	409
Chapter 5 Wrap Up	411
Concept Check	411
Section Reviews	411
Key Terms	413
Extra Practice	414
References	429

Chapter 6: Foundations of Inference

6.1 Point Estimation and Sampling Distributions	433
Statistical Inference	434
Point Estimation	434
Sampling Distributions	435
Unbiased Estimation	435
6.2 The Sampling Distribution of the Sample Mean (σ Known)	438
The Central Limit Theorem for a Sample Mean	438
Using the CLT	441
6.3 Introduction to Confidence Intervals	444
Confidence Intervals	444
Calculating the Confidence Interval	446
Changing the Confidence Level	447
Finding the Critical Value	448
Calculating the Margin of Error (MoE)	449
Constructing the Confidence Interval	450
Writing the Interpretation	450

6.4 The Behavior of Confidence Intervals	453
Changing the Confidence Level or Sample Size	453
Alternative Interpretation	454
Working Backwards to Find the Error Bound or Sample Mean	455
Calculating the Sample Size needed.	456
6.5 Introduction to Hypothesis Tests	459
Defining your hypotheses	460
Calculating a Test Statistic	462
Making a Decision	462
Decision and conclusion	464
6.6 Hypothesis Tests In-Depth	467
Errors in Hypothesis Tests	467
Statistical Significance Versus Practical Significance	469
Chapter 6 Wrap Up	470
Concept Check	470
Section Reviews	470
Key Terms	473
Extra Practice	475
References	512

Chapter 7: Inference for One Sample

7.1 The Sampling Distribution of the Sample Mean (σ Un-known)	517
Student's t Distribution	518
Finding T Distribution Probabilities	521
7.2 Inference for the Mean in Practice	523
Confidence Intervals for the Mean (σ Unknown)	523
Hypothesis Tests for the Mean (σ Unknown)	525
Summary of Assumptions	526
7.3 The Sampling Distribution of the Sample Proportion	527
Understanding the Variability of a Proportion	527
Conditions for the CLT for p	528

7.4 Inference for a Proportion	531
Hypothesis Tests for p	531
Confidence Intervals for p	532
Constructing Confidence Intervals for p	533
7.5 Behavior of Confidence Intervals for a Proportion	535
Calculating the Sample Size n	535
“Plus Four” Confidence Interval for p .	536
Chapter 7 Wrap Up	538
Concept Check	538
Section Reviews	538
Key Terms	541
Extra Practice	542
References	570

Chapter 8: Inference for Two Samples

8.1 Inference for Two Dependent Samples (Matched Pairs)	575
Two Dependent Samples (Matched Pairs)	577
Hypothesis Tests for the Mean difference	577
Confidence Intervals for the Mean difference	579
8.2 Inference for Two Independent Sample Means	584
Both Population Standard Deviations Known (Z)	584
Both Population Standard Deviations UnKnown (t)	585
Hypothesis Tests for the Difference in Two Independent Sample Means	585
Confidence Intervals for the Difference in Two Independent Sample Means	586
8.3 Inference for Two Sample Proportions	587
Sampling Distribution of the Difference in Two Proportions	587
Hypothesis Test for the Difference in Two Proportions	587
Confidence Intervals for the Difference in Two Proportions	589

Chapter 8 Wrap Up	590
Concept Check	590
Section Reviews	590
Key Terms	592
Extra Practice	593
References	639
 <u>Chapter 9: Simple Linear Regression</u>	
9.1 Introduction to Bivariate Data and Scatterplots	643
Bivariate Data	644
Scatterplots	644
9.2 Measures of Association	651
The Correlation Coefficient, r	651
The Coefficient of Determination, r^2	654
9.3 Modeling Linear Relationships	656
Linear Regression	656
Understanding Slope	659
Understanding the Y-Intercept	659
Prediction	660
9.4 Cautions about Regression	663
Linearity	663
Correlation Does Not Imply Causation	663
Extrapolation	663
Outliers and Influential Points	664
9.5 Inference for Regression	674
Inference for regression Assumptions:	674
Regression Standard Errors	675
Inference on the slope	675

Chapter Wrap Up	677
Concept Check	677
Section Reviews	677
Key Terms	678
Extra Practice	680
References	697

Class Group Activities

Normal Distribution (Lap Times)	701
Normal Distribution (Pinkie Length)	703
Central Limit Theorem (Pocket Change)	705
Central Limit Theorem (Cookie Recipes)	708
Confidence Interval (Home Costs)	712
Confidence Interval (Place of Birth)	714
Confidence Interval (Women's Heights)	716
Continuous Distribution	718
Data Collection Experiment	721
Sampling Experiment	723
Hypothesis Testing of a Single Mean and Single Proportion	726
Probability Topics	729
Discrete Distribution (Playing Card Experiment)	732
Discrete Distribution (Lucky Dice Experiment)	736
Regression (Distance from School)	739
Regression (Textbook Cost)	741
Regression (Fuel Efficiency)	743
Descriptive Statistics	745

Review Exercises (Ch 1-13)	747
Chapter 3	0
Chapter 4	0
Chapter 5	0
Chapter 6	0
Chapter 7	0
Chapter 8	0
Chapter 9	0
Chapter 10	0
Chapter 11	0
Chapter 12	0
Practice Tests (1-4) and Final Exams	748
Practice Test 1	0
Practice Test 1 Solutions	0
Practice Test 2	0
Practice Test 2 Solutions	0
Practice Test 3	0
Practice Test 3 Solutions	0
Practice Test 4	0
Practice Test 4 Solutions	0
Data Sets	749
Lap Times	749
Stock Prices	750
Group and Partner Projects	752
Solution Sheets	761
Mathematical Phrases, Symbols, and Formulas	765
Notes for the TI-83, 83+, 84, 84+ Calculators	774
Tables	788
Glossary	789

Introduction

Letter To Students

Dear Reader,

Welcome to **Significant Statistics: An Introduction to Statistics**. This textbook was written to increase student access to high-quality learning materials at no cost. These types of materials available under Creative Commons licenses are often called Open Educational Resources (OER).

Statistics is about separating the signal from the noise, deciphering what is actually significant versus what is just happening due to random chance. In addition to demonstrating the basic concepts needed to do that, this book attempts to focus on what is significant and eliminate some of the noise that may commonly be found in many introductory statistics texts.

In this book I have “remixed” sections from two of the most widely used OER texts in the introductory statistics space and sprinkled in some thoughts of my own. This book does not focus or lean on any specific technology, but rather the concepts.

Most sections feature worked examples with solutions, some of which have interactive features. Then “Your Turn” problems are included to encourage readers to try similar problems independently. Each end of chapter “wrap-up” includes summaries of each section, key terms, and more practice problems.

Thanks for reading this book. I hope it proves useful!

Sincerely,

John Morgan Russell

Learning Objectives

Significant Statistics: An Introduction to Statistics is intended as a one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. In addition to end of section practice and homework sets, examples of each topic are explained step-by-step throughout the text and followed by a Your Turn problem designed as extra practice for students.

Having successfully completed the course the student should be able to:

1. Identify and critique the use of statistical reasoning in science, industry, and public discourse
2. Identify appropriate data to be gathered to answer research questions
3. Assign appropriate data collection methods
4. Apply appropriate methods of data visualization to explore data from a variety of disciplines
5. Analyze data provided and use relevant technology when needed

6. Appropriately interpret results of data exploration and statistical tests
7. Employ critical thinking to make decisions
8. Apply ethical reasoning and principles to scientific research

Coverage and scope

Chapter 1 Sampling and Data

Chapter 2 Descriptive Statistics

Chapter 3 Probability Topics

Chapter 4 Discrete Random Variables

Chapter 5 Continuous Random Variables

Chapter 6 Introduction to Inference

Chapter 7 One Sample Inference

Chapter 8 Two Sample Inference

Chapter 9 Simple Linear Regression

For Instructors

Letter to Instructors

Dear Colleague,

We all know the issues with the price of higher education. One small, but still significant, aspect over which instructors may have some level of control is the materials used. The use of Open Educational Resources (OER) is a growing trend that I hope will continue to catch on.

Statistics is about separating the signal from the noise, deciphering what is actually significant versus what is just happening due to random chance. In addition to demonstrating the basic concepts needed to do that, this book attempts to focus on what is significant and eliminate some of the noise that may commonly be found in many introductory statistics texts. In the realm of introductory statistics there are many OER options available, the most complete being [Introductory Statistics from OpenStax](#) and [OpenIntro Statistics](#). While these are both adequate options, the beauty of OER is that you can customize material to the needs of your course and students which is what I have tried to do here. Specific to this book I have “remixed” sections from the aforementioned texts and sprinkled in some thoughts of my own.

The intended audience for **Significant Statistics: An Introduction to Statistics** includes students who may not have completed a calculus prerequisite. In contrast to similar introductory statistics texts, this text has an emphasis on data collection, but does not deeply delve into probability topics. The text also takes a repetitive approach to one-sample inference to drive those basic concepts home. It includes introductory level material on two-sample inference (Ch 8) and omits advanced inference topics such as Chi Square or ANOVA which are infrequently covered in an introductory-level statistics course. The regression chapter (Ch 9) could be moved anywhere in the book by omitting the section on inference for regression (9.5). Please see the Attribution section for more details.

I’ve also tried to make this book technology agnostic in order to accommodate a wide variety of technology preferences. It is my belief that if one understands the concepts, they can figure out how to use any technology to accomplish the task at hand. However, if learners lean on technology when first being introduced to the concepts, true understanding may not be achieved.

I hope you will consider using this text!

Sincerely,

John Morgan Russell

P.S: The Chapter Wrap-ups and Glossary functionalities are still a work in progress!

Instructors reviewing, adopting, or adapting this textbook please help us understand your use by filling out this form: <https://bit.ly/stat-interest>

If you locate an error in the book, please report it here: <https://bit.ly/feedback-stat>. Errata reported thus far are available at: <http://bit.ly/stat-errata>

Supplemental Material

Supplemental multimedia material aligned with this textbook including videos, audio-only versions of the videos in podcast format, and PowerPoint lecture notes which I have branded “Significant Statistics can be found at:

- [Significant Statistics Website](#)
- [Significant Statistics YouTube Channel](#)
- [Significant Statistics Podcast Channel](#)

Learning Objectives

Significant Statistics: An Introduction to Statistics is intended for the one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. In addition to end of section practice and homework sets, examples of each topic are explained step-by-step throughout the text and followed by a Your Turn problem designed as extra practice for students.

Having successfully completed the course the student should be able to:

1. Identify and critique the use of statistical reasoning in science, industry, and public discourse
2. Identify appropriate data to be gathered to answer research questions
3. Assign appropriate data collection methods
4. Apply appropriate methods of data visualization to explore data from a variety of disciplines
5. Analyze data provided and use relevant technology when needed
6. Appropriately interpret results of data exploration and statistical tests
7. Employ critical thinking to make decisions
8. Apply ethical reasoning and principles to scientific research

Coverage and scope

Chapter 1 Sampling and Data

Chapter 2 Descriptive Statistics

Chapter 3 Probability Topics

Chapter 4 Discrete Random Variables

Chapter 5 Continuous Random Variables

Chapter 6 Introduction to Inference

Chapter 7 One Sample Inference
Chapter 8 Two Sample Inference
Chapter 9 Simple Linear Regression

Attribution

Sources and Workflow

Significant Statistics: An Introduction to Statistics is adapted by John Morgan Russell from open textbooks from OpenStax and OpenIntro. These source books are released under open licenses which allow reuse and remix at no cost with attribution. Additional topics, examples, and innovations in terminology and practical applications have been added, all with a goal of increasing relevance and accessibility for students.

Content for this book was gathered and adapted from multiple openly-licensed sources including:

- [OpenStax Introductory Statistics by Barbara Illowsky and Susan Dean](#), which is licensed with a [Creative Commons Attribution 4.0 \(CC BY 4.0\) license](#),
- [OpenIntro Statistics by David Diez, Mine Çetinkaya-Rundel, and Christopher D Barr](#) which is licensed with a [Creative Commons Attribution Share-Alike 3.0 \(CC BY SA 3.0\) license](#), and
- [Introductory Statistics for the Life and Biomedical Sciences by Julie Vu and David Harrington](#) which is licensed with a [Creative Commons Attribution Share-Alike 3.0 \(CC BY SA 3.0\) license](#)

The base of the book is from OpenStax, much of which was reworded and reorganized. The main reorganizations were streamlining the probability chapter (CH3), removing ancillary discrete (CH4) and continuous (CH5) distribution sections, introducing the normal distribution (5.3) in the continuous distributions chapter and removing the Chi-Square and ANOVA Chapters.

Additional content from the OpenIntro texts was then added to fill in gaps. This included adding more detailed information on data collection (1.3 & 1.4), Normal approximation (5.3), and inferential techniques applied to proportions (7.3-7.5, 8.3). Several figures were also adapted from the OpenIntro texts.

About the Adapter

John Morgan Russell teaches various introductory statistics courses at Virginia Tech, and previously taught at George Mason University and Old Dominion University. He earned a BS in Mathematics from Christopher Newport University, an MS in Statistical Science from George Mason University, and an Ed.S. in Instructional design and Technology from Virginia Tech. His interests include statistics education, instructional design, and open educational resources.

About the Editorial Team

Editorial support was provided for this project by Managing Editor, Anita Walz and Graphic Design Specialist, Kindred Grey.

Kindred Grey is a recent graduate of Virginia Tech with majors in Statistics and Psychology. Apart from her studies, she enjoys graphic design and color theory. Her contributions to this textbook include review and identification of elements that would make the book more relevant and accessible to today's students: development of the color scheme, original graphics, and selection of images to form a more cohesive and lucid textbook, and authoring of alt-text to enable equivalent access to graphics by readers who rely on screen reader software.

Anita Walz is Associate Professor, and Assistant Director of Open Education and Scholarly Communication Librarian in the University Libraries at Virginia Tech. She received her MS in Library and Information Science from the University of Illinois at Urbana-Champaign and has worked in university, government, school, and international libraries for eighteen years. She is the founder of the Open Education Initiative at Virginia Tech and the managing editor of several open textbooks adapted or created at Virginia Tech, many of which may be found here: <https://vtechworks.lib.vt.edu/handle/10919/70959>. She provided overall planning, project coordination, day-to-day supervision, and oversight.

Project Funding

Development of this book was made possible in part through a grant from [the University Libraries at Virginia Tech's Open Education Initiative](#).

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=20#h5p-1>

Section Reviews

Reviews of each section will appear here!

Key Terms

Key terminology for each section will appear here!

Extra Practice

Extra Practice problems for each section will appear here!

CHAPTER 1: SAMPLING AND DATA

1.1 Introduction to Statistics and Key Terms

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms dealing with statistics
- Identify different types of data
- Identify data collection methods and study designs
- Apply various types of sampling methods to data collection

Introduction

We encounter statistics in our daily lives more often than we probably realize in many different contexts such as the news, the weather, the lab, and the classroom.

“Statistics’ ultimate goal is translating data into knowledge” – Alan Agresti & Christine Franklin

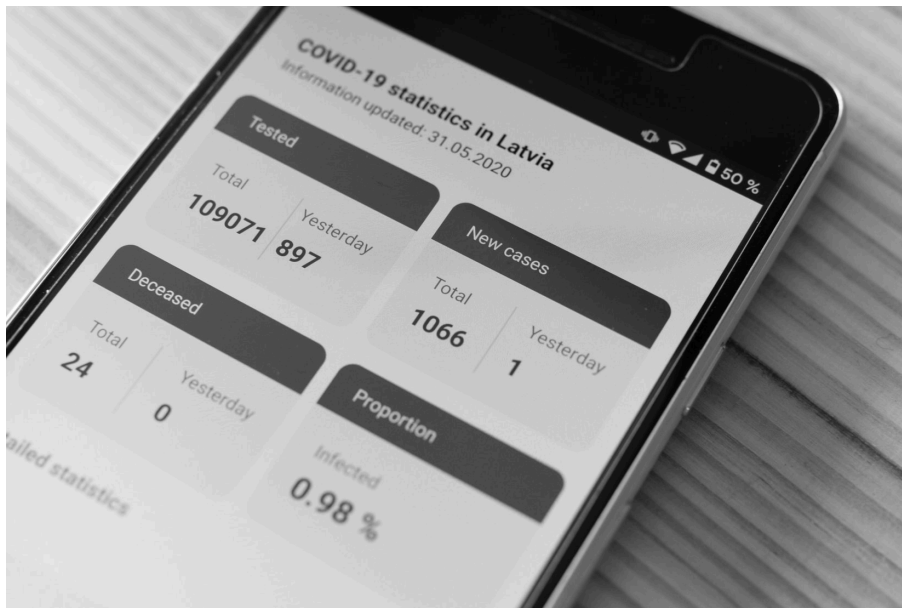


Figure 1.1: Smartphone Display of COVID-19 Statistics. Since the Coronavirus is novel, statisticians must collect and translate data into digestible information to give to the public.

You are probably asking yourself the question, “When and where will I use statistics?” If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or “fact.” Statistical methods can help you make the “best educated guess.”

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and how “good” data can be distinguished from “bad.”

The Study of Statistics

We see and use data in our everyday lives. The science of statistics deals with the collection, analysis, interpretation, and presentation of data. This is reflected in the **data analysis process** which we will expand on in the next section.

You will first learn how to organize and summarize data. Organizing, summarizing, and presenting data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing useful conclusions from data while filtering out the noise. The formal methods are called **inferential statistics**.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination. You will encounter what will seem to be a lot of mathematical formulas to make calculations. Keep in mind, the goal of statistics is not to perform numerous calculations using the formulas, but to interpret data to gain an understanding. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a medical test incorrectly diagnosing the presence of a disease. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will utilize the power of mathematics and probability to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a sample. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. **Parameters** are numbers that describe a characteristic of the population.

Because it may take a lot of time and resources (time, money, manpower, etc.) to examine an entire

population, we often study only a subset of that population. Taking a **sample** is a very practical technique to accomplish this. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the information we collect in our sample, we can calculate a **statistic**. A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A parameter is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

Individuals are the units we are collecting information about. This could be a person, animal, item, thing, or place. A **variable**, usually notated by capital letters such as X or Y, is a specific characteristic or measurement that can be determined for each individual. The **values** of a variable are the possible observations of the variable. If there are multiple variables collected on an individual the entire row may be called a case or observational unit.

Data is the collection of the actual values of the variables of interest. They may be numbers or they may be words. We'll dive into data in the next section.

Example

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=23#h5p-2>

Your turn!

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Image References

Figure 1.1: Markus Winkler (2020). “Corona death and new cases stats.” Public domain. Retrieved from: <https://unsplash.com/photos/tUEnyweZjEU>

1.2 Data Basics

Types of Data

Data may come from a **population** or from a **sample**. Lowercase letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- **Qualitative (categorical)**
- **Quantitative (numerical)**

Qualitative, or categorical data come in many forms. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Categorical data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+.



Figure 1.2: Red Jaguar. Car type (in this case, Jaguar) can be considered categorical data since it is described using words.

Quantitative data are always numbers and is often called numerical data. Quantitative data are typically the result of counting or measuring attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called quantitative discrete data. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are made up not only of counting numbers, but of all possible values on an interval (the real numbers) are called quantitative continuous data. Continuous data are often the results of measurements like lengths, weights, or times. The length, in minutes, of a phone call would be quantitative continuous data.

If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some categories include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Example

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=26#h5p-3>

Your turn!



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=26#h5p-4>

Levels of Measurement

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a nominal scale is categorical data where the categories have no natural order. Colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful. Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an ordinal scale is similar to nominal scale data but there is a big difference. Ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data. Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the interval scale is similar to ordinal level data because it has a definite order. However, there is a meaningful difference between values of the data from an arbitrary starting point. Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, differences make sense, but 40° is equal to 100° minus 60° . But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0. Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the ratio scale takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded. The data can be put in order from lowest to highest: 20, 68, 80, 92. The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Note: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Your turn!



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=26#h5p-5>

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8, 16.1, 15.2, 14.8, 15.8 15.9, 16.0, 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers

regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range. Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Data Analysis

In this age of “Big Data” **data analysis** is an essential tool. Informally, it could be defined as the process of collecting, organizing, and analyzing your data. Formally, the process consists of 4 phases and associated questions to answer:

1. Identify the research objective.
 - What questions are to be answered?
 - What group should be studied?
 - Have attempts been made to answer it before?
2. Collect the information needed.
 - Is data already available?
 - Can you access the entire population?
 - How can you collect a good sample?
3. Organize and summarize the information
 - What visual descriptive techniques are appropriate?
 - What numerical descriptive techniques are appropriate?
 - What aspects of the data stick out?
4. Draw conclusions from the information.
 - What Inferential techniques are appropriate?
 - What conclusions can I draw?

We will answer all of these questions and more throughout the course.

Image References

Figure 1.2: Mateusz Delegacz (2017). “London Jaguar 2.” Public domain. Retrieved from <https://unsplash.com/photos/1Ah8CAwk3vM>

1.3 Data Collection and Observational Studies

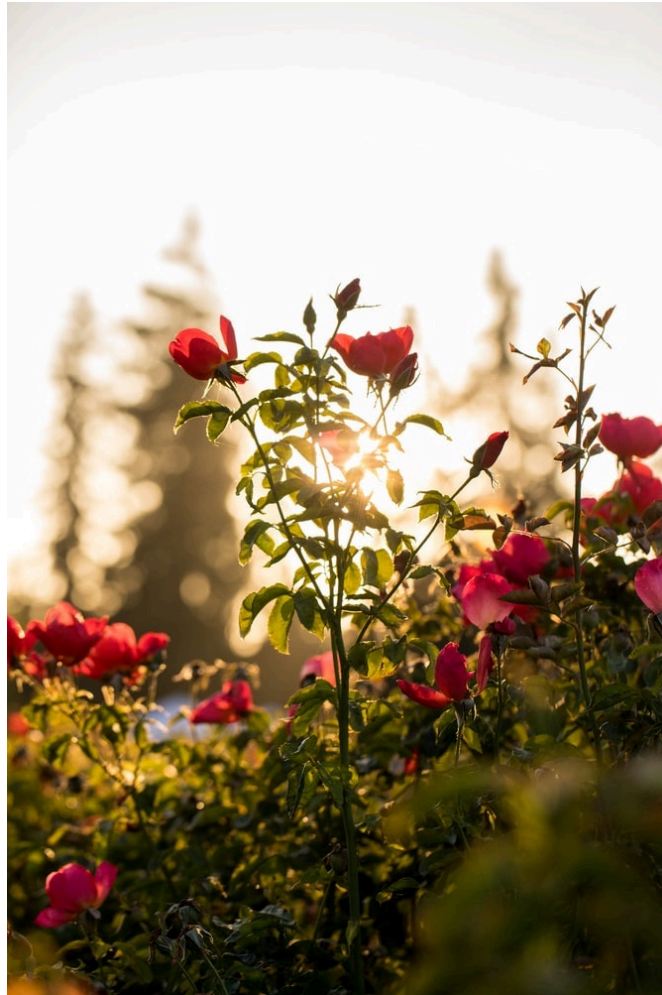


Figure 1.3: Flower Growth. Is one brand of fertilizer more effective at growing flowers than another? Statisticians can answer this question by determining what effect the explanatory variable (fertilizer brands) has on the response variable (flower growth).

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? When we are interested in the effect one variable may have on another, we call the first variable the **explanatory variable** and the second the **response variable**. Questions like these are answered using studies and experiments. Proper study design ensures the production of reliable, accurate data.

Data Collection Methods

There are many ways **data** is commonly collected, each with their own pros and cons. Some ways data may be collected are:

- **Anecdotal evidence**
- **Observational studies**
- **Designed (controlled) experiments**

The latter two options are more commonly accepted, but we will briefly describe the former first.

Anecdotal Evidence

Consider the following possible responses to the these research questions:

1. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
2. A man on the news had an adverse reaction to a vaccine, so it must be dangerous.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is technically based on data, however, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called anecdotal evidence. While such evidence may be true and verifiable, be careful of data collected in this way since it may only represent extraordinary or unusual cases. Often we are more likely to recall cases relying on anecdotal evidence based on their striking characteristics. For instance, in case #2 above, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Observational Studies

Researchers perform an observational study when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via a questionnaire or survey, review medical or company records, or follow a group of many similar individuals to form hypotheses about why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In

general, observational studies can provide evidence of naturally occurring **associations** between variables, but they cannot by themselves show a causal connection. Why not? Consider the following example:

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer? Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer? One important piece of information that is absent may be sun exposure.

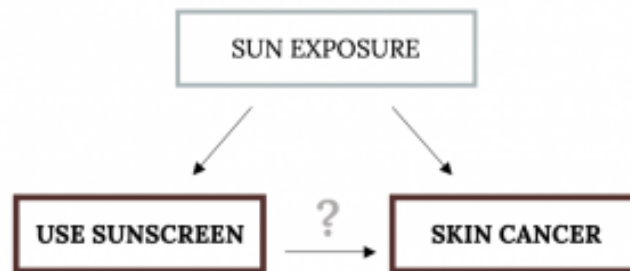


Figure 1.4: Association Between Sunscreen and Skin Cancer

Exposure to the sun is unaccounted for in this simple investigation since it stands to reason if someone is out in the sun all day, she is more likely to use sunscreen *but also* more likely to get skin cancer. Sun exposure here is an example of what we might call a **confounding (lurking, conditional) variable**, a variable that was not accounted for and may actually be important. Confounding variables can cause many misleading, counterintuitive or even humorous (spurious) correlations.

Observational studies come in two forms: **prospective** and **retrospective** studies. A prospective study identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of patients over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989. This prospective study recruits registered nurses and then collects data from them using questionnaires. Retrospective studies collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets may contain both prospectively- and retrospectively-collected variables.

There are other classifications of observational studies you may encounter, especially in life science and medical contexts: A **cohort study** is when we follow a group of many similar individuals over time often producing **longitudinal** data. A **cross-sectional study** indicates data collection on a population at one point in time (often prospective). A **case-control study** compares a group that has a certain characteristic to a group that does not, often a retrospective study for rare conditions.

Example

A researcher is studying the relationship between time spent studying in med school and depression rates among students. The researcher looks at graduated students' medical records to determine if they have ever seen a psychologist. He also sends out a questionnaire to the same students to ask how much time they spent studying in college. Identify which type of study this is.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=30#h5p-6>

Your turn!

A researcher is wondering if the same individual can contract COVID-19 more than once. She randomly selects 300 people who have tested positive for COVID-19. The participants fill out a self-report survey once a month to inform the researcher if they have tested positive again. What type of study is this?

Image References

Figure 1.3: Jason Leung (2018). “Selective focus photo of red peonies.” Pubic domain, Retrieved from <https://unsplash.com/photos/nonlZlChSZQ>

Figure 1.4: Kindred Grey (2020). “Sun Exposure Confounding Factors.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Sun_Exposure_Confounding_Factors.png

1.4 Designed Experiments

Observational Studies vs. Experiments

Ignoring anecdotal evidence, there are two primary types of data collection: **observational studies** and **controlled (designed) experiments**. Remember, we typically cannot make claims of causality from observation studies because of the potential presence of confounding factors. However, making causal conclusions based on experiments is often reasonable by controlling for those factors. Consider the following example:

Suppose you want to investigate the effectiveness of vitamin D in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin D. You notice that the subjects who take vitamin D exhibit better health on average than those who do not. Does this prove that vitamin D is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin D consumption. People who take vitamin D regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not necessarily prove that vitamin D is the key to disease prevention.

Experiments ultimately provide evidence to make decisions, so how could we narrow our focus and make claims of causality? In this section, you will learn important aspects of experimental design.

Designed Experiments

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable may be called **treatments**. An **experimental unit** is a single object or individual to be measured.

The main principles we want to follow in experimental design are:

1. Randomization
2. Replication
3. Control

Randomization

In order to provide evidence that the explanatory variable is indeed causing the changes in the response variable, it is necessary to isolate the explanatory variable. The researcher must design their experiment in

such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by randomization of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can show an apparent cause-and-effect connection between the explanatory and response variables.

Recall our previous example of investigating the effectiveness of vitamin D in preventing disease. Individuals in our trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: The control group (no treatment) and the second group receives extra doses of Vitamin D.

Replication

The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we replicate by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding. Having individuals experience a treatment more than once, called **repeated measures** is often helpful as well.

Control

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*¹

It is often difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind** experiment is one in which both the subjects and the researchers involved with the subjects are blinded.

Randomized experiments are an essential tool in research. The US Food and Drug Administration typically

1. McClung, M. Collins, D. “Because I know it will!”: placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

requires that a new drug can only be marketed after two independently conducted randomized trials confirm its safety and efficacy; the European Medicines Agency has a similar policy. Large randomized experiments in medicine have provided the basis for major public health initiatives. In 1954 approximately 750,000 children participated in a randomized study comparing the polio vaccine with a placebo. In the United States, the results of the study quickly led to the widespread and successful use of the vaccine for polio prevention.

Example

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=33#h5p-7>

Your turn!

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.





An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=33#h5p-8>

More Experimental Design

There are many different experimental designs from the most basic, a single treatment and control group, to some very complicated designs. In an experimental design setting, when working with more than one variable, or treatment, they are often called **factors**, especially if it is categorical. The values of factors are often called **levels**. When there are multiple factors, the combinations of each of the levels are called **treatment combinations**, or interactions. Some basic ones you may see are:

1. **Completely randomized**
2. **Block design**
3. **Matched pairs design**

Completely Randomized

While very important and an essential research tool, not much explanation is needed for this design. It involves figuring out how many treatments will be administered and randomly assigning participants to their respective groups.

Block Design

Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into blocks and then randomize cases within each block to the treatment groups. This strategy is often referred to as blocking. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and

the other half to the treatment group, as shown in the figure below. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

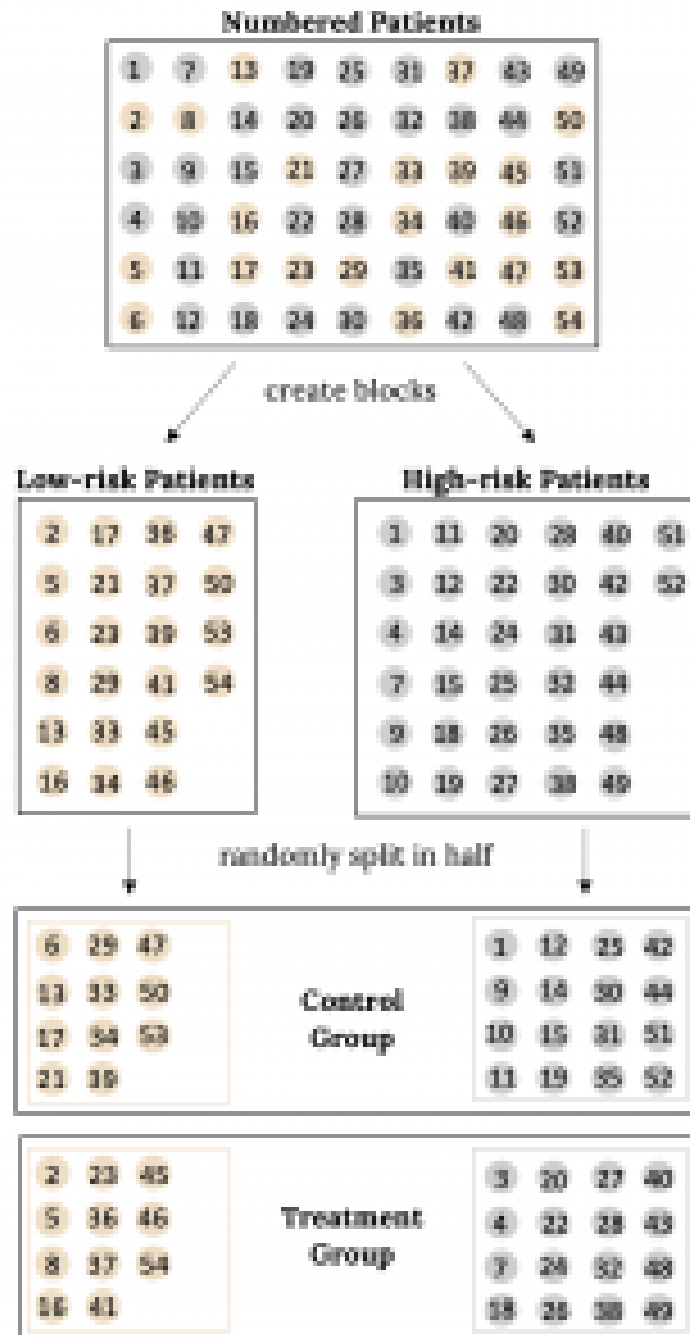


Figure 1.5: Block Design

Matched Pairs

A **matched pairs design** is one where we have very similar individuals (or even the same individual) receiving different two treatments (or treatment vs. control), then comparing their results. This design is very powerful, however, it can be hard to find many like individuals to match up. Some common ways of creating a matched pairs design are twin studies, before and after measurements, pre and post test situations, or crossover studies. Consider the following example:

In the 2000 Olympics, was the use of a new wetsuit design responsible for an observed increase in swim velocities? In a matched pairs study designed to investigate this question, twelve competitive swimmers swam 1500 meters at maximal speed, once wearing a wetsuit and once wearing a regular swimsuit. The order of wetsuit versus swimsuit was randomized for each of the 12 swimmers. Figure 1.6 shows the average velocity recorded for each swimmer, measured in meters per second (m/s).

Figure 1.6: Average Velocity of Swimmers

	swimmer.number	wet.suit.velocity	swim.suit.velocity	velocity.diff
1	1	1.57	1.49	0.08
2	2	1.47	1.37	0.10
3	3	1.42	1.35	0.07
4	4	1.35	1.27	0.08
5	5	1.22	1.12	0.10
6	6	1.75	1.64	0.11
7	7	1.64	1.59	0.05
8	8	1.57	1.52	0.05
9	9	1.56	1.50	0.06
10	10	1.53	1.45	0.08
11	11	1.49	1.44	0.05
12	12	1.51	1.41	0.10

Notice in this data, two sets of observations are uniquely paired so that an observation in one set matches an observation in the other; in this case, each swimmer has two measured velocities, one with a wetsuit and one with a swimsuit. A natural measure of the effect of the wetsuit on swim velocity is the difference between the measured maximum velocities ($\text{velocity.diff} = \text{wet.suit.velocity} - \text{swim.suit.velocity}$). Even though there are two measurements per individual, using the difference in observations as the variable of interest allows for the problem to be analyzed.

Example

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. What experiment design is being implemented here?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=33#h5p-9>

Your turn!

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after. What experiment design is being implemented here?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=33#h5p-9>

Image References

Figure 1.5: Kindred Grey (2020). “Block Design.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Block_Design.png

1.5 Sampling Techniques and Ethics

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we typically use a **sample** of the population which should have the same characteristics as the population it is representing. Statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of random sampling. In each form, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons.

Simple Random Sampling

The gold standard and maybe easiest method to describe is called a **simple random sample (SRS)**. Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in the figure below:

Figure 1.7: Lisa's Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360, 0.99832, 0.14669, 0.51470, 0.40581, 0.73381, 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

Other Sampling Techniques

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. Other well-known random sampling methods are:

- **Stratified sampling**
- **Cluster sampling**
- **Systematic sampling**

To choose a **stratified sample**, divide the population into groups called strata and then take a proportionate number from each stratum. For example, you could stratify (group) your college population by department and

then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

Example

A study is done to determine the average tuition that Virginia Tech undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=35#h5p-10>

Your turn!

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task. The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music. Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Sampling and Replacement

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. For any particular sample of 1,000, if you are sampling with replacement,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling without replacement,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions $\frac{999}{10,000}$ and $\frac{999}{9,999}$. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling

with replacement for any particular sample, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample without replacement, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions $\frac{9}{25}$ and $\frac{9}{24}$. To four decimal places, $\frac{9}{25} = 0.3600$ and $\frac{9}{24} = 0.3750$. To four decimal places, these numbers are not equivalent.

Bias in Samples

When you analyze data, it is important to be aware of sampling errors and non-sampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause non-sampling errors. A defective counting device can cause a non-sampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Variation in Samples

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This idea of **sampling variability** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include:

- **Convenience sampling:** A type of sampling that is non-random and involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.
- **Problems with samples:** A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- **Self-selected samples:** Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- **Sample size issues:** Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- **Undue influence:** collecting data or asking questions in a way that influences the response
- **Non-response or refusal of subject to participate:** The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- **Causality:** A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- **Self-funded or self-interest studies:** A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- **Misleading use of data:** improperly displayed graphs, incomplete data, or lack of context
- **Confounding:** When the effects of multiple factors on a response cannot be separated. Confounding

makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world’s top journals including *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

*Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. “It was a quest for aesthetics, for beauty—instead of the truth,” he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.*²

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel’s fraud states that, “statistical flaws frequently revealed a lack of familiarity with elementary statistics.”³ Many of Stapel’s co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don’t want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and

Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as Institutional Review Boards (IRB). All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give informed consent. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a [website](#) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

- a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
- b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
- c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the

neighborhood.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=35#h5p-11>

Your turn!

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

- a. The survey is commissioned by the seller of a popular brand of apple juice.
- b. There are only two types of juice included in the study: apple juice and cranberry juice.
- c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
- d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

Chapter 1 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-12>

Section Reviews

1.1 Introduction to Statistics and Key Terms

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

1.2 Data Basics

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement: nominal, ordinal, interval, and ratio.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

1.3 Data Collection and Observational Studies

In summary, making causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations or form hypotheses that we later check using controlled experiments which we will discuss in the next section.

1.4 Designed Experiments

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule,” Andrew Gelman. Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

1.5 Sampling Techniques and Ethics

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the

population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

1.1 Introduction to Statistics and Key Terms

- **Data analysis process**
- **Descriptive statistics**
- **Inferential statistics**
- **Probability**
- **Population**
- **Parameters**
- **Sample**
- **Statistic**
- **Individuals**
- **Variable**
- **Values**
- **Data**

1.2 Data Basics

- **Data**
- **Population**
- **Sample**
- **Qualitative (categorical)**
- **Quantitative (numerical)**
- **Discrete**
- **Continuous**
- **Nominal scale level**

- Ordinal scale level
- Interval scale level
- Ratio scale level
- Variation
- Data analysis

1.3 Data Collection and Observational Studies

- Explanatory variable
- Response variable
- Data
- Anecdotal evidence
- Observational studies
- Designed (controlled) experiment
- Associations
- Confounding (lurking, conditional) variable
- Prospective study
- Retrospective study
- Cohort study
- Longitudinal study
- Cross-sectional study
- Case-control study

1.4 Designed Experiments

- Observational study
- Controlled (designed) experiments
- Explanatory variable
- Response variable
- Treatments
- Experimental unit
- Repeated measures
- Control group
- Placebo
- Blinding
- Double-blind
- Factors
- Levels
- Treatment combinations (interactions)

- **Completely randomized**
- **Block design**
- **Matched pairs design**

1.5 Sampling Techniques and Ethics

- **Sample**
- **Simple random sample (SRS)**
- **Stratified sampling**
- **Cluster sampling**
- **Systematic sampling**
- **Sampling bias**
- **Sampling variability**
- **Convenience sampling**

Extra Practice

1.1 Introduction to Statistics and Key Terms

1. Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-2>

2. Determine what the key terms refer to in the following study. We want to know the average (mean) amount

of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

3. Determine what the key terms refer to in the following study.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-13>

4. As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Figure 1.8: Automobile safety

Speed at which Cars Crashed	Location of “driver” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.¹



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-14>

5. An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

1. The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-15>

6. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

Researcher B: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

Determine what the key terms refer to in the example for Researcher A.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-16>

7. For each of the following exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

a. A fitness center is interested in the mean amount of time a client exercises in the center each week.

- Solution: The population is all of the clients of the fitness center. A sample of the clients that use the fitness center for a given week. The average amount of time that all clients exercise in one week. The average amount of time that a sample of clients exercises in one week. The amount of time that a client exercises in one week. Examples are: 2 hours, 5 hours, and 7.5 hours

b. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

- Solution:

a. all children who take ski or snowboard lessons

- b. a group of these children
 - c. the population mean age of children who take their first snowboard lesson
 - d. the sample mean age of children who take their first snowboard lesson
 - e. X = the age of one child who takes his or her first ski or snowboard lesson
 - f. values for X , such as 3, 7, and so on
- c. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.
- Solution: the cardiologist's patients, a group of the cardiologist's patients, the mean recovery period of all of the cardiologist's patients, the mean recovery period of the group of the cardiologist's patients, X = the mean recovery period of one patient values for X , such as 10 days, 14 days, 20 days, and so on
- d. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.
- Solutions:
- a. the clients of the insurance companies
 - b. a group of the clients
 - c. the mean health costs of the clients
 - d. the mean health costs of the sample
 - e. X = the health costs of one client
 - f. values for X , such as 34, 9, 82, and so on
- e. A politician is interested in the proportion of voters in his district who think he is doing a good job.
- Solutions: all voters in the politician's district, a random selection of voters in the politician's district, the proportion of voters in this district who think this politician is doing a good job, the proportion of voters in this district who think this politician is doing a good job in the sample, X = the number of voters in the district who think this politician is doing a good job, Yes, he is doing a good job. No, he is not doing a good job.
- f. A marriage counselor is interested in the proportion of clients she counsels who stay married.
- Solutions:
- a. all the clients of this counselor
 - b. a group of clients of this marriage counselor
 - c. the proportion of all her clients who stay married
 - d. the proportion of the sample of the counselor's clients who stay married
 - e. X = the number of couples who stay married
 - f. yes, no
- g. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

- Solutions: all voters (in a certain geographic area), a random selection of all the voters, the proportion of voters who are interested in this particular cause, the proportion of voters who are interested in this particular cause in the sample, X = the number of voters who are interested in this particular cause, yes, no

h. A marketing company is interested in the proportion of people who will buy a particular product.

- Solutions:
 - a. all people (maybe in a certain geographic area, such as the United States)
 - b. a group of the people
 - c. the proportion of all people who will buy the product
 - d. the proportion of the sample who will buy the product
 - e. X = the number of people who will buy it
 - f. buy, not buy

8. A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

a. What is the population she is interested in?

- a. all Lake Tahoe Community College students
- b. all Lake Tahoe Community College English students
- c. all Lake Tahoe Community College students in her classes
- d. all Lake Tahoe Community College math students

- Solution: d

b. Consider the following: X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of a:

- a. variable.
- b. population.
- c. statistic.
- d. data.

- Solution: a

c. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

- a. parameter.
- b. data.

- c. statistic.
- d. variable.

- Solution: c

9. In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks.

a. The “average increase” for all NASDAQ stocks is the:

- a. population
- b. statistic
- c. parameter
- d. sample
- e. variable

b. All of the NASDAQ stocks are the:

- a. population
- b. statistics
- c. parameter
- d. sample
- e. variable

c. Nine percent is the:

- a. population
- b. statistics
- c. parameter
- d. sample
- e. variable

d. The 100 NASDAQ stocks in the survey are the:

- a. population
- b. statistic
- c. parameter
- d. sample
- e. variable

e. The percent increase for one stock in the survey is the:

- a. population

- b. statistic
- c. parameter
- d. sample
- e. variable

f. Would the data collected be qualitative, quantitative discrete, or quantitative continuous?

1.2 Data Basics

1. The data are the colors of backpacks. You sample five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. What type of data is this?

- Solution: qualitative (categorical) data
-

2. The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. What type of data are the numbers of books (three, four, two, and one)?

- Solution: quantitative discrete data
-

3. The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. What type of data is this?

- Solution: quantitative continuous data
-

4. The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

5. The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

6. The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

7. Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

1.3 Data Collection and Observational Studies + Designed Experiments

1. Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.²



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-17>

2. A researcher wants to study the effects of birth order on personality.



2. Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443> (accessed May 1, 2013)



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-18>

3. You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

Describe the explanatory and response variables in the study.

What are the treatments?

What should you consider when selecting participants?

Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?

Identify any lurking variables that could interfere with this study.

How can blinding be used in this study?

4. Identify any issues with the following studies

- a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
- b. A research study is designed to investigate a new children's allergy medication.
- c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-19>

1.5 Sampling Techniques and Ethics

1. Determine whether or not the following samples are representative.

- a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
- b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
- c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
- d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
- e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.
-

2. Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-20>

3. A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities. What type of sampling is used? (simple random, stratified, systematic, cluster, or convenience)



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-21>

4. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class. Use the random number generator to generate different types of samples from the data.

Figure 1.12: Quiz scores

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

- Create a stratified sample by column. Pick three quiz scores randomly from each column.
- Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
- Create a simple random sample of 15 quiz scores.
- Create a systematic sample of 12 quiz scores.

5. Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

\$128 \$87 \$173 \$116 \$130 \$204 \$147 \$189 \$93 \$153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

\$50 \$40 \$36 \$15 \$50 \$100 \$40 \$53 \$22 \$22

It is unlikely that any student is in both samples.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-22>



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-23>

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

\$180 \$50 \$150 \$85 \$260 \$75 \$180 \$200 \$200 \$150



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-24>

6. What type of data is this?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-25>

7. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in Norfolk, Virginia. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.



— An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-26>



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-27>



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-28>

The population is _____

8. The following figure contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.³

3. “Earthquake Information by Year,” U.S. Geological Survey. <http://earthquake.usgs.gov/earthquakes/eqarchives/year/> (accessed May 1, 2013)

Figure 1.13: Earthquake fatalities

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Use figure above to answer the following questions.

- What is the proportion of deaths between 2007 and 2012?
- What percent of deaths occurred before 2001?
- What is the percent of deaths that occurred in 2003 or after 2010?
- What is the fraction of deaths that happened before 2012?
- What kind of data is the number of deaths?
- Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
- What contributed to the large number of deaths in 2010? In 2004? Explain.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-29>

9. Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

- A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

- b. A market researcher polls every tenth person who walks into a store.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-30>

- c. The first 50 people who walk into a sporting event are polled on their television preferences.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-31>

- d. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-32>

10. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

Researcher B: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

- a. Determine what the key term data refers to in the above example for Researcher A.

- values for X, such as 3, 4, 11, and so on

b. List two reasons why the data may differ.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-33>

c. Can you tell if one researcher is correct and the other one is incorrect? Why?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-34>


d. Would you expect the data to be identical? Why or why not?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-35>

e. Suggest at least two methods the researchers might use to gather random data.

-  An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-36>

f. Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?



An interactive H5P element has been excluded from this version of the text. You can view

it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-37>

g. Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=37#h5p-38>

11. Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Figure 1.16: Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Figure 1.17: Researcher B

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

a. Give a reason why the data may differ.

- Solution: The researchers are studying different groups, so there will be some variation in the data.

b. Would the sample size be large enough if the population is the students in the school?

- Solution: Yes, the sample size of 150 would be large enough to reflect a population of one school.

c. Would the sample size be large enough if the population is school-aged children and young adults in the United States?

- Solution: There are many school-aged children and young adults in the United States, and the study was done at only one school, so the sample size is not large enough to reflect the population. ->

d. Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

- Solution: Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

e. As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

- Solution: Yes, people who play games more might be more likely to participate, since they would want the gift card more than a student who does not play video games. This would leave out many students who do not play games at all and skew the data.

12. A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in Figure 1.18. The second study collected the data in Figure 1.19.

Figure 1.18: First study

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Figure 1.19: Second study

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

- a. Given what you know, which study is correct?
- There is not enough information given to judge if either one is correct or incorrect.
- b. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?
- Solution: The second study is more reliable, because the company would be interested in showing results that favored a higher rate of improvement from patients using their software. The data may be skewed; however, the American Medical Association is not concerned with the success of the software and so should be objective. ->
- c. Both groups that performed the study concluded that the software works. Is this accurate?
- The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.
- d. The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?
- Solution: No, the data suggest the two are correlated, but more studies need to be done to prove that using the software causes improvement in stroke patients.
- e. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from the second study?
- Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.
- f. Is a sample size of 1,000 a reliable measure for a population of 5,000?

- Solution: Yes, 1,000 represents 20% of the population and should be representative, if the population of the sample is chosen at random.

g. Is a sample of 500 volunteers a reliable measure for a population of 2,500?

- No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

h. A question on a survey reads: “Do you prefer the delicious taste of Brand X or the taste of Brand Y?” Is this a fair question?

- Solution: No, the question is creating undue influence by adding the word “delicious” to describe Brand X. The wording may influence responses.

i. Is a sample size of two representative of a population of five?

- No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

j. Is it possible for two experiments to be well run with similar sample sizes to get different data?

- Solution: Yes, there will most likely be a degree of variation between any two studies, even if they are set up and run the same way. Each study may be affected differently by unknown factors such as location, mood of the subjects, or time of year.

13. For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

a. number of tickets sold to a concert

- Solution: quantitative discrete, 150

b. percent of body fat

- Solution: quantitative continuous, 19.2%

c. favorite baseball team

- Solution: qualitative, Oakland A's

d. time in line to buy groceries

- Solution: quantitative continuous, 7.2 minutes
 - e. number of students enrolled at Evergreen Valley College
 - Solution: quantitative discrete, 11,234 students
 - f. most-watched television show
 - Solution: qualitative, The Voice
 - g. brand of toothpaste
 - Solution: qualitative, Crest
 - h. distance to the closest movie theatre
 - Solution: quantitative continuous, 8.32 miles
 - i. age of executives in Fortune 500 companies
 - Solution: quantitative continuous, 47.3 years
 - j. number of competing computer spreadsheet software packages
 - Solution: quantitative discrete, three
-

14. A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in Norfolk. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

- a. “Number of times per week” is what type of data?
 - Solution: quantitative discrete
 - b. “Duration (amount of time)” is what type of data?
 - Solution: quantitative continuous
-

15. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six

flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.

- Solution:

- The survey was conducted using six similar flights.
- The survey would not be a true representation of the entire population of air travelers.
- Conducting the survey on a holiday weekend will not produce representative results.

b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

- Solution:

- Conduct the survey during different times of the year.
- Conduct the survey using flights to and from various locations.
- Conduct the survey on different days of the week.

16. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

- Solution: Answers will vary. Sample Answer: Randomly choose 15 colleges in the state. Use all statistics classes from each of the chosen colleges in the sample. This can be done by listing all the colleges together with a two-digit number starting with 00 then 01, etc. The list of colleges can be found on [Wikipedia](#). Use a random number generator to pick 15 colleges.

17. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

- Solution: Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

18. List some practical difficulties involved in getting accurate results from a telephone survey.

- Solution: Answers will vary. Sample Answer: Many people live in different areas than their area code. Many people hang up or do not respond to phone surveys.

19. List some practical difficulties involved in getting accurate results from a mailed survey.

- Solution: Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.
-

20. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. What type of sampling did she use?

- Solution: stratified sampling
-

21. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. What was the sampling method?

- Solution: systematic
-

22. Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.

- Solution: convenience

b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.

- Solution: cluster

c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.

- Solution: stratified

d. The librarian at a public library wants to determine what proportion of the library users are children.

The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

- Solution: systematic

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

- Solution: simple random
-

23. A “random survey” was conducted of 3,274 people of the “microprocessor generation” (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had \$2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?

- Solution: Yes, in polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done.

b. Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

- Solution: We do not have enough information to decide if this is a random sample from the U.S. population.

c. Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called “America's Smithsonian.” With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?

- Solution: No, this is a convenience sample taken from individuals who visited an exhibition in the Angeles Convention Center. This sample is not representative of the U.S. population.

d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

- Solution: It is possible that the two sample statistics, 48% and 66% are larger than the true parameters in the population at large. In any event, no conclusion about the population proportions can be inferred from this convenience sample.

24. The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.⁴

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?

- Solution: qualitative

b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?

- Solution: quantitative discrete

c. In the last seven days, on how many days did you exercise for 30 minutes or more?

- Solution: quantitative discrete

d. Do you have health insurance coverage?

- Solution: qualitative

25. In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.⁵

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.

- Solution: The country was in the middle of the Great Depression and many people could not afford these “luxury” items and therefore not able to be included in the survey.

4. Gallup-Healthways Well-Being Index. <http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx> (accessed May 1, 2013).

5. Dominic Lusinchi, “President’ Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?” *Social Science History* 36, no. 1: 23-54 (2012), <https://www.jstor.org/stable/41407095> (accessed January 26, 2021).

b. What effect does the low response rate have on the reliability of the sample?

- Solution: Samples that are too small can lead to sampling bias.

c. Are these problems examples of sampling error or nonsampling error?

- Solution: sampling error

d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called “quota sampling” to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?⁶

- Solution: stratified

26. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI’s *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.⁷

Which of the potential problems with samples discussed could explain this connection?

Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

27. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?”⁸

As of April 25, 11 people responded to this question. Each participant answered “NO!”

6. Data from <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

7. “United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at <http://www.disastercenter.com/crime/> (accessed May 2, 2013).

8. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: <http://www.youpolls.com/details.aspx?id=12328> (accessed May 1, 2013).

Which of the potential problems with samples discussed in this module could explain this connection?

- Solution: Self-Selected Samples: Only people who are interested in the topic are choosing to respond. Sample Size Issues: A sample with only 11 participants will not accurately represent the opinions of a nation. Undue Influence: The question is wording in a specific way to generate a specific response. Self-Funded or Self-Interest Studies: This question was generated to support one person's claim and it was designed to get the answer that the person desires.
-

28. A scholarly article about response rates begins with the following quote:

"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."⁹

The Pew Research Center for People and the Press admits:

"The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more."¹⁰

a. What are some reasons for the decline in response rate over the past decade?

- Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed

b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

- When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.
-

29. Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in Figure 1.20.

9. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD

Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), <http://poq.oxfordjournals.org/content/70/5/759.full> (<http://poq.oxfordjournals.org/content/70/5/759.full>) (accessed May 1, 2013).

10. Frequently Asked Questions, Pew Research Center for the People & the Press, <http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls> (accessed May 1, 2013).

Figure 1.20: LBCC distance learning survey results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

a. What percent of the students surveyed do not have a computer at home?

- Solution: 4%

b. About how many students in the survey live at least 16 miles from campus?

- Solution: 13%

c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

- Solution: Not necessarily. Long beach City is the seventh largest in California the college has an enrollment of approximately 27,000 students. On the other hand, Great Basin College has its campuses in rural northeastern Nevada, and its enrollment of about 3,500 students.

30. Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this

study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

- Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

References

Figures

Figure 1.9: Kindred Grey via Virginia Tech (2020). “Other Guy’s Investments.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Other_Guy%27s_Investments.png . Adaptation of Figure 1.14 from

OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/1-homework#fs-idp81996208>

Figure 1.10: Kindred Grey via Virginia Tech (2020). “Acme’s Investments.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Acme%27s_Investments.png . Adaptation of Figure 1.14 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/1-homework#fs-idp81996208>

Figure 1.11: Kindred Grey (2020). “Airline Complaints 2.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Airline_Complaints_2.png

Text

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, <http://www.hsph.harvard.edu/nutritionsource/vitamin-e/> (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com, <http://www.athleteinme.com/ArticleView.aspx?id=1053> (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443> (accessed May 1, 2013).

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html> (accessed May 1, 2013).

M.L. Jacscon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013), <http://www.ncbi.nlm.nih.gov/pubmed/22721550> (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. <http://earthquake.usgs.gov/earthquakes/eqarchives/year/> (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. <http://www-fars.nhtsa.dot.gov/Main/index.aspx> (accessed May 1, 2013).

Meier, Paul. “The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine.” *Statistics: a guide to the unknown*. San Francisco: Holden-Day (1972): 2-13.

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” <http://www.forbes.com/best-small-companies/list/> (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," econoclass.com, <http://www.econoclass.com/misleadingstats.html> (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, <http://cnx.org/content/m15555/latest/> (accessed May 1, 2013).

McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

Yudhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

"Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tilburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, <http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf> (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/default.asp> (accessed May 1, 2013).

Data from <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

"The Literary Digest Poll," Virtual Laboratories in Probability and Statistics <http://www.math.uah.edu/stat/data/LiteraryDigest.html> (accessed May 1, 2013).

"Gallup Presidential Election Trial-Heat Trends, 1936-2008," Gallup Politics <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4> (accessed May 1, 2013).

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html> (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, <http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus> (accessed May 1, 2013).

Data from San Jose Mercury News

CHAPTER 2: DESCRIPTIVE STATISTICS

2.1 Introduction to Descriptive Statistics and Frequency Tables

Learning Objectives

By the end of this chapter, the student should be able to:

- Display and interpret categorical data
- Display and interpret quantitative data
- Recognize, describe, and calculate the measures of the center of quantitative data
- Recognize, describe, and calculate the measures of the spread of quantitative data
- Recognize, describe, and calculate the measures of location of quantitative data
- Identify outliers in quantitative data



Figure 2.1: Voting Ballots. When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized.

Descriptive Statistics

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at numerical descriptions such as the average or median house price. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **descriptive statistics**. We will look at both **graphical** and **numerical** descriptive methods. You will learn how to construct and calculate, and even more importantly, how to interpret these measurements and graphs.

Numerical descriptors consist of summary statistics, typically calculated from a sample, that represent important aspects such as the central tendency and variability of a distribution, or relative standing of a single observation with regards to the rest of the distribution.

Graphical descriptive methods consist of chart, tables, and graphs. These are tools that help you learn about the **distribution**, or shape of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values.

Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

The type of graph you choose to use first depends on the type of data you are working with. Some of the types of graphs used to display Categorical data are pie charts and bar charts. Some graphs that are used to summarize and organize Quantitative data are the dot plot, the histogram, the stem-and-leaf plot, the frequency polygon, the box plot, and the time series plot in special cases. The emphasis will be on histograms and box plots.

We will start by looking at a graphical method that can display any type of data, the frequency table.

Frequency Tables

Frequency tables are a great starting place for summarizing and organizing your data. Once you have a set of data, you may first want to organize it to see the **frequency**, or how often each value occurs in the set.

Frequency tables can be used to show either quantitative or categorical data. Displaying categorical data in a frequency table is fairly straightforward since you already have clearly defined categories. For example if you polled 20 kindergarteners on their favorite colors you could construct the following simple frequency table:

Table 2.1: Frequency Table of Children's favorite colors

Color	FREQUENCY
Red	2
Orange	2
Yellow	1
Green	3
Blue	4
Purple	3
Pink	4
Clear with Sparkles	1
	Total = 20

Some quantitative data, especially discrete, may only contain a limited number of values and little thought would be needed in creating the frequency table. Some data may have a natural grouping. For example, if you had ages of adults from 20-69, it might make intuitive sense to group them as follows:

- 20 – 29
- 30 – 39
- 40 – 49
- 50 – 59

- 60 – 69

Consider the 30–39 class. 30 is known as the **lower class limit**, while 39 is the **upper class limit**. The **class width** is defined as the difference between consecutive lower class limits. For the class 30 – 39, the class width = $40 - 30 = 10$. The **class midpoint** is found by adding the lower limit and upper limit, then dividing by 2. For the class 30 – 39, the class midpoint = $(30 + 39)/2 = 34.5$.

Depending on the format and precision of the data reported, we may have to decide how best to group our data into intervals, sometimes called bins or classes. Grouping data may not always have an intuitive way to do it or work out cleanly. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 ($6.1 - 0.05 = 6.05$). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 ($1.5 - 0.005 = 1.495$). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 ($1.0 - 0.0005 = 0.9995$). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 ($2 - 0.5 = 1.5$). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

The next question may be how many bins should we use? Generally anywhere from 5–20 bins, since too few does not display distribution well, but too many can create strange effects. A good place to start is the square root of your number of observations (n). Some other basic guidelines are bins should not overlap, not have gaps between them, have the same width, and cover the entire range of the data. The class limits and width should be “reasonable” numbers such as whole numbers, 5s, 10s, etc... In the end it really just depends on the format of your data, but following these general guidelines should make sure our table is useful.

Relative Frequencies

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percents, or decimals. To find the relative frequency:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:

$$RF = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, $f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the figure below.

NOTES:

- The sum of all frequencies will add up to n , or your sample size.
- All relative frequencies should add up to one (pending rounding)
- The first entry of the cumulative relative frequency column will be the same as the first entry of the relative frequency column since there is nothing to accumulate.
- The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated

Example

The following table represents one way of grouping the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table 2.5: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	Total = 1.00	

In this sample, there are five players whose heights fall within the interval 59.95–61.95 inches, three players whose heights fall within the interval 61.95–63.95 inches, 15 players whose heights fall within the interval 63.95–65.95 inches, 40 players whose heights fall within the interval 65.95–67.95 inches, 17 players whose heights fall within the interval 67.95–69.95 inches, 12 players whose heights fall within the interval 69.95–71.95, seven players whose heights fall within the interval 71.95–73.95, and one player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

- a. From the figure above, find the percentage of heights that are less than 65.95 inches.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=41#h5p-39>

- b. Find the percentage of heights that fall between 61.95 and 65.95 inches.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=41#h5p-40>

c. Use the heights of the 100 male semiprofessional soccer players. Fill in the blanks and check your answers.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=41#h5p-41>

d.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=41#h5p-42>

e. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=41#h5p-43>

Remember, you count frequencies. To find the relative frequency, divide the frequency by the total

number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Your turn!

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3. Construct a bar graph that shows the registered voter population by district.

Construct an appropriate table including frequencies, relative frequencies, and cumulative relative frequencies.

Image Credits

Figure 2.1: U.S. Marine Corps photo by Staff Sgt. William Greeson (2009). "US Navy 090821-M-0440G-043 Voting ballots organized and arranged for counting by Afghan presidential election workers at a local school in the Nawa District." Public domain. Retrieved from: https://commons.wikimedia.org/wiki/File:US_Navy_090821-M-0440G-043_Voting_ballots_organized_and_arranged_for_counting_by_Afghan_presidential_election_workers_at_a_local_school_in_the_Nawa_District.jpg

2.2 Displaying and Describing Categorical Data

Descriptive Statistics for Categorical Data

Categorical data is typically more straightforward to work with. Recall descriptive statistics consists of visual and numerical methods. We usually start with visual methods and then move into numerical.

Graphical Methods for Categorical Data

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Figure 2.6: Full and Part Time Students

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display categorical data are pie charts and bar graphs.

Pie Charts

In a pie chart, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category. Suppose a statistics professor collects information about the

classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart below.

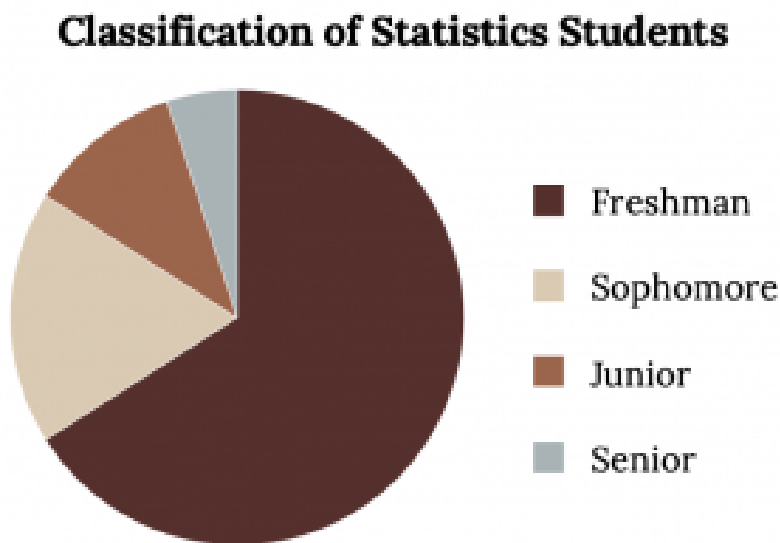


Figure 2.7: Classification of Statistics Students

Bar Graphs

Bar graphs consist of bars that are separated from each other. The length of the bar for each category is proportional to the number or percent of individuals in each category. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The bar graph shown in the figure below has age groups represented on the x -axis and proportions on the y -axis.

By the end of 2011, Facebook had over 146 million users in the United States. The figure below shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Figure 2.8: Facebook Users

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

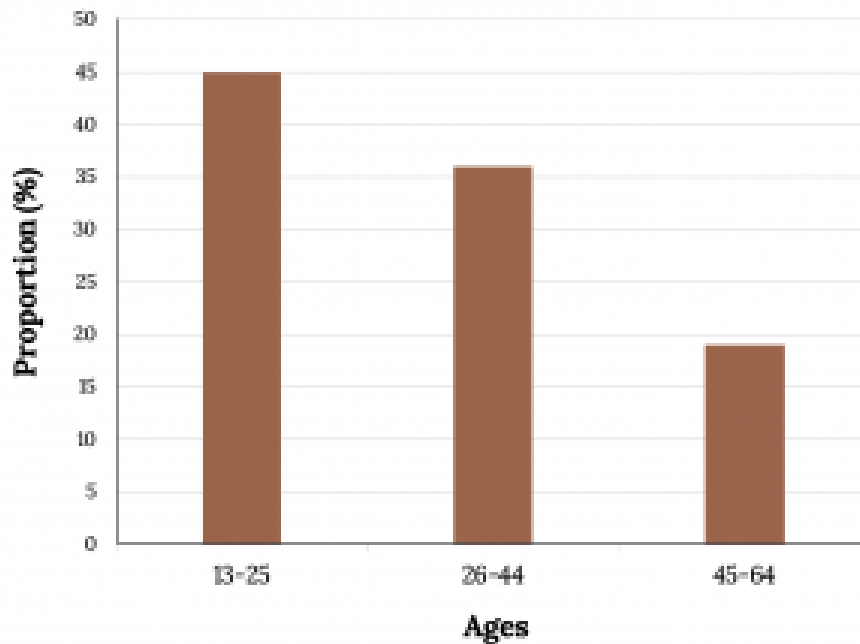


Figure 2.9: Facebook Users (Bar Graph)

Pie vs. Bar Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for. Look at the following plots (pie or bar) and think about which you think displays the comparisons better:

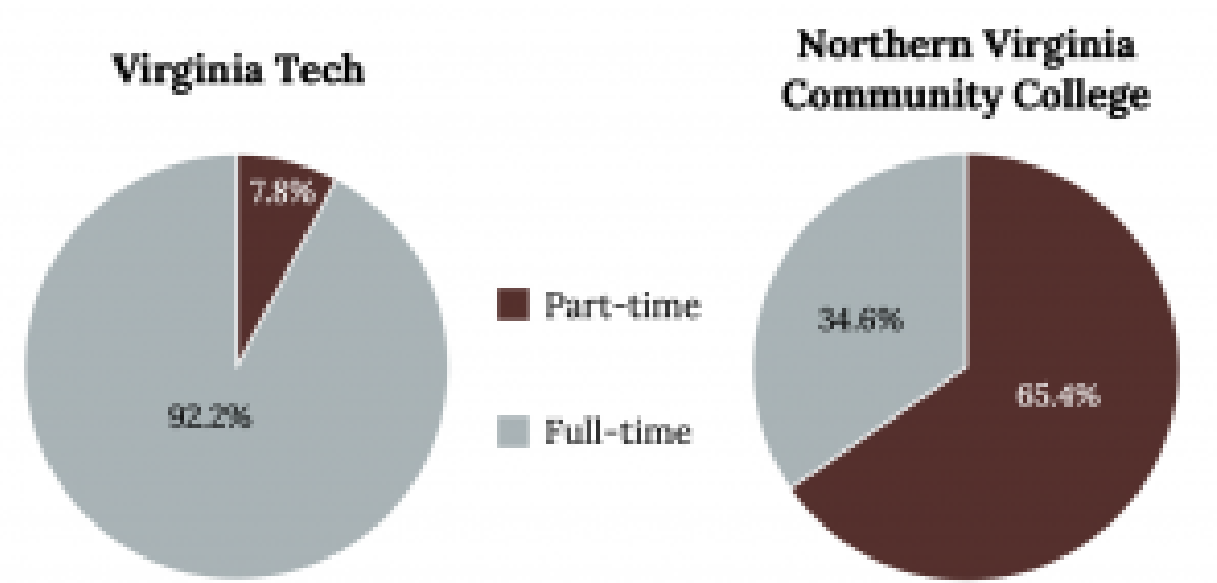


Figure 2.10: Full-time and Part-time Students at Virginia Tech and NVCC

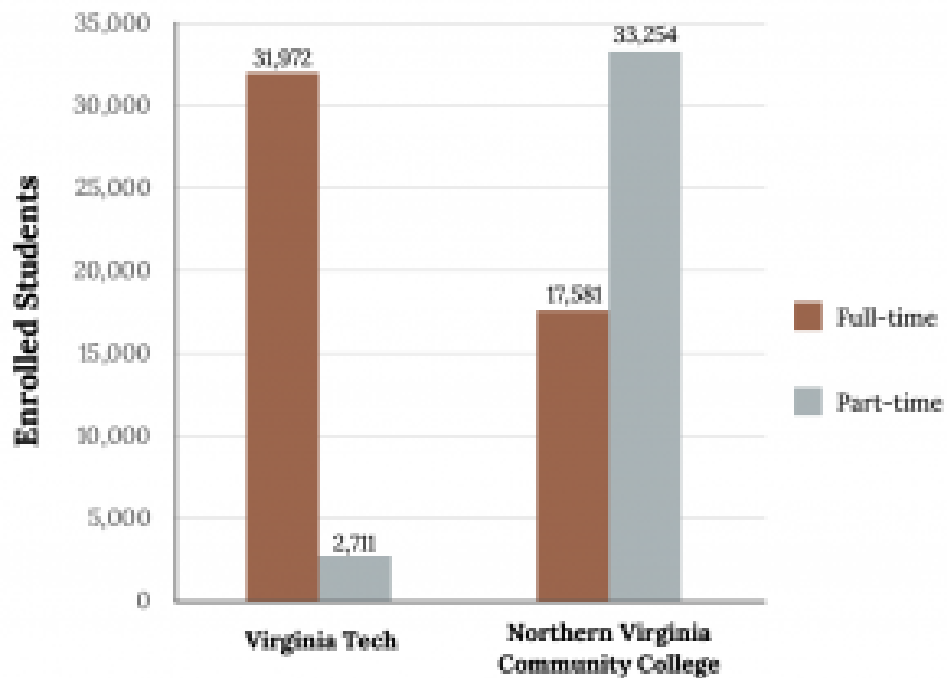


Figure 2.11: Full-time and Part-time Students at Virginia Tech and NVCC (Bar Graph)

Percentages That Add to More (or Less) than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Figure 2.12: De Anza College Data

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

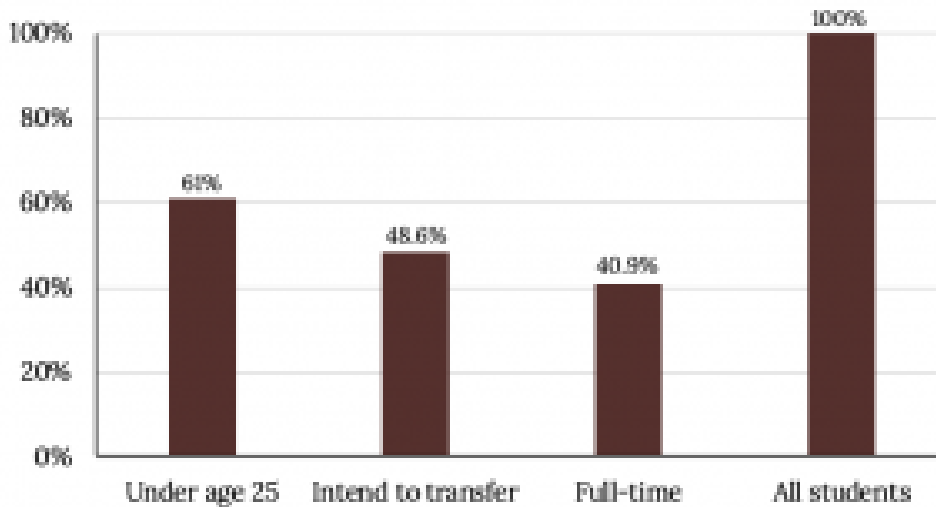


Figure 2.13: De Anza College Bar Graph

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the “Other/Unknown” category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Figure 2.14: Ethnicity of Students at De Anza College

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

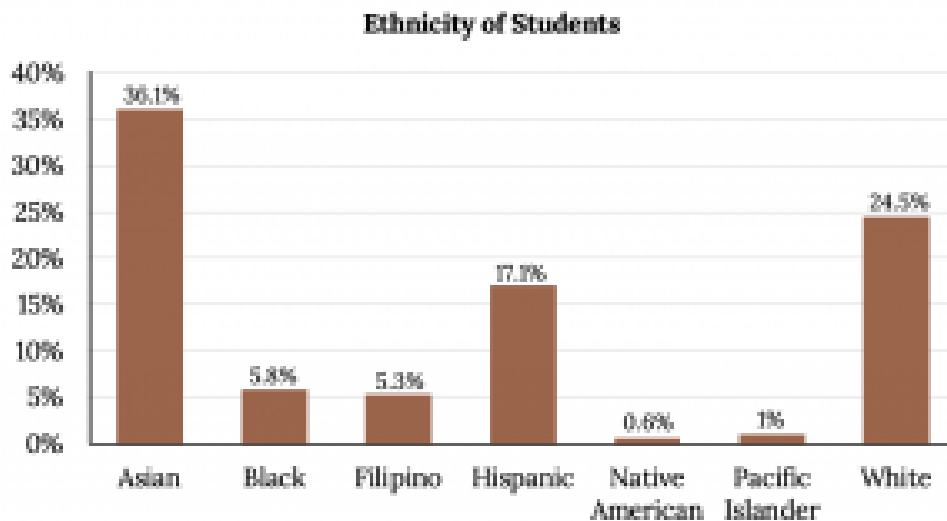


Figure 2.15: Ethnicity of Students at De Anza College (Bar Graph)

The following graph is the same as the previous graph but the “Other/Unknown” percent (9.6%) has been included. The “Other/Unknown” category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

Bar Graph with Other/Unknown Category

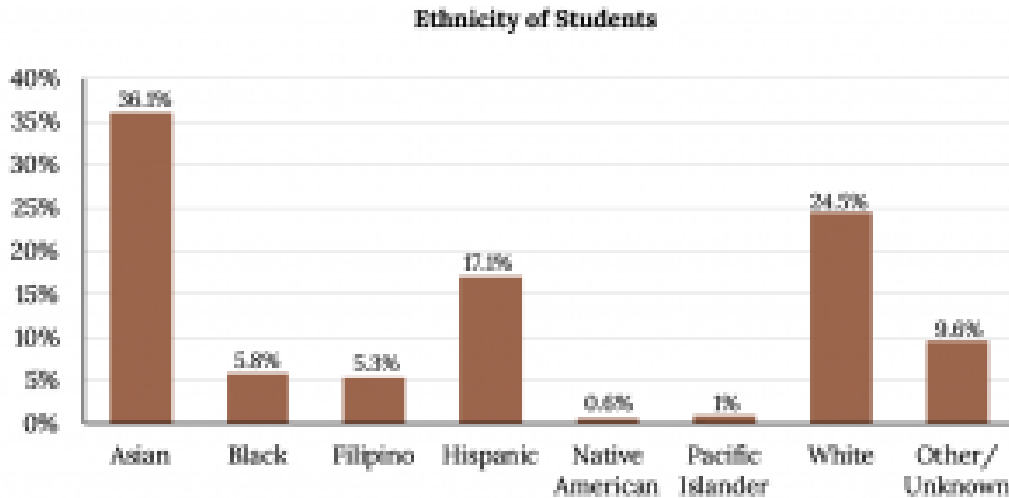


Figure 2.16: Ethnicity of Students at De Anza College (Bar Graph with 'Other' Category)

This particular bar graph could be difficult to understand visually at first glance. A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest). This Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

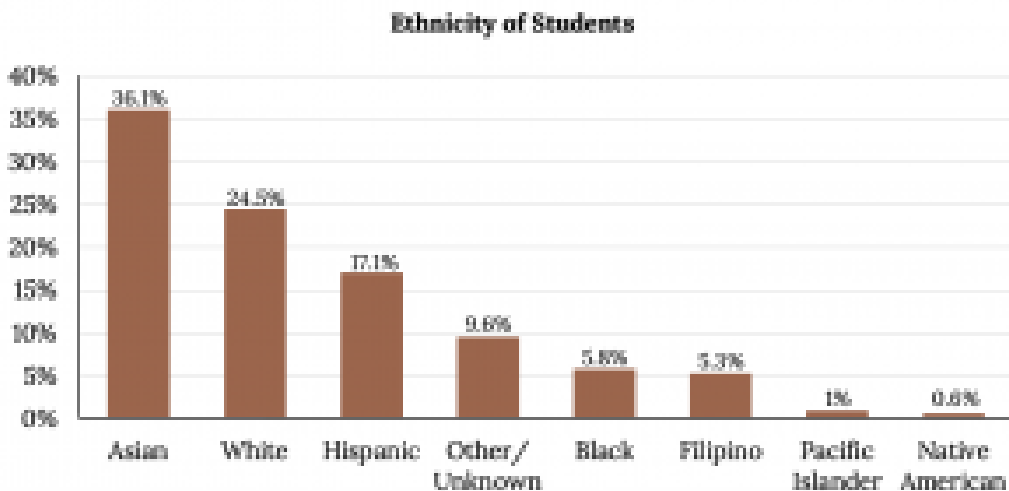


Figure 2.17: Ethnicity of Students at De Anza College (Bar Graph with 'Other' Category)

Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The second chart below is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in the first chart below.

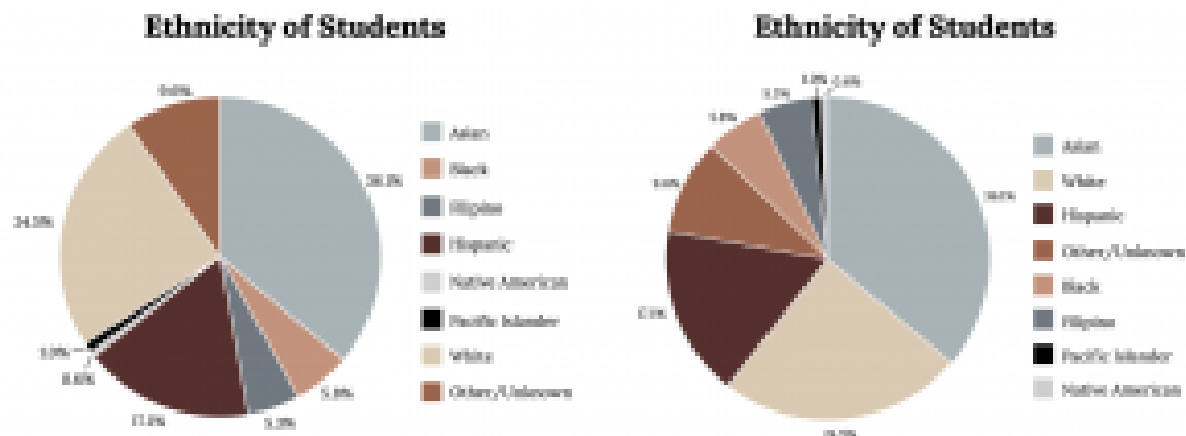


Figure 2.18: Pie Charts With No Missing Data

Example

The columns in the figure below contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the x -axis, and the Advanced Placement examinee population percentages on the y -axis.

Figure 2.19: AP Student Population

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution:

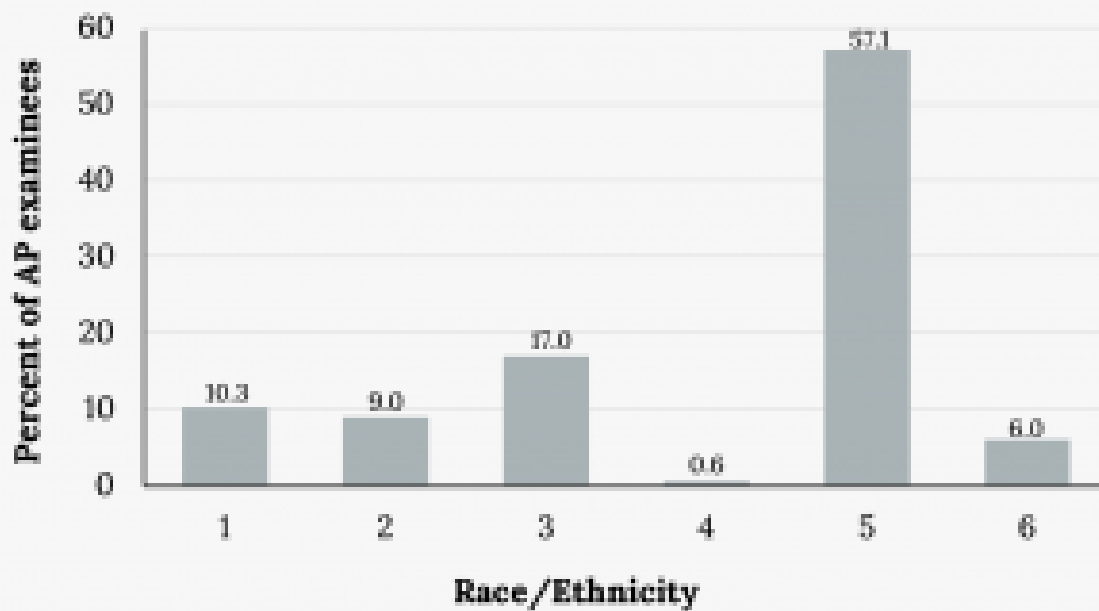


Figure 2.20: AP Student Population (Bar Graph)

Your turn!

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

Figure 2.21: Registered Voter Population by District

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Construct a bar graph that shows the registered voter population by district.

Describing Categorical Data

After we have displayed the data visually, we then want to follow up by describing it with numerical measures. Since Categorical Data does not lend itself to mathematical calculations by nature there are not many numerical descriptors we can use to describe it. However, we can describe a categorical distribution's "typical value" with the **mode**, and can also note its level of **variability**.

Mode

The Mode of a dataset is the most frequently occurring value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal, three modes – trimodal, multiple modes – multimodal, etc. In most cases the mode can easily be found as the largest piece of a pie chart, or largest bar in a bar chart. Looking at some previous examples:

Classification of Statistics Students

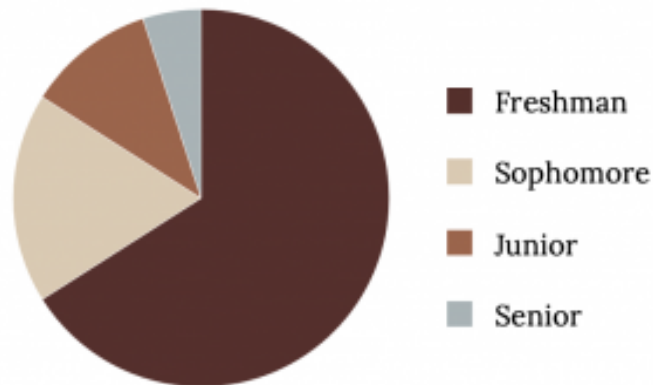


Figure 2.7 (repeat): Classification of Statistics Students

Ethnicity of Students

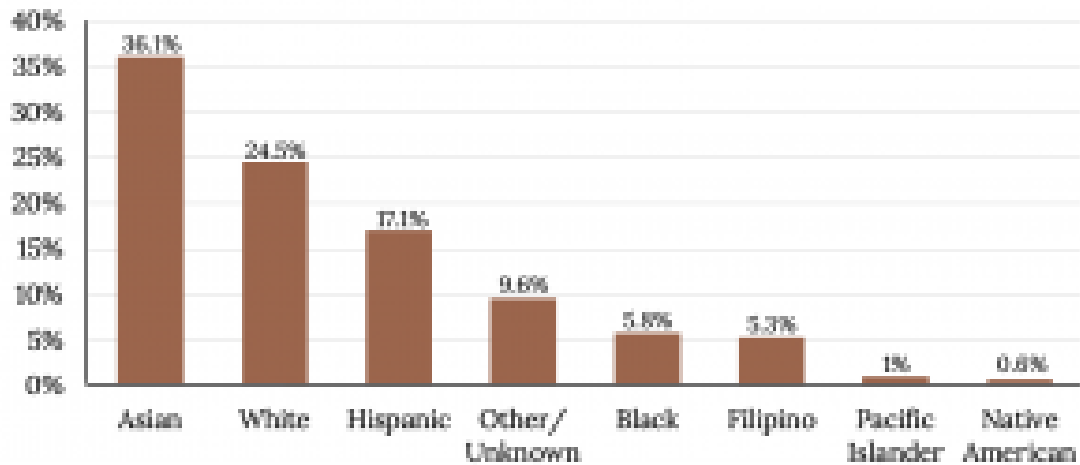


Figure 2.17 (repeat): Ethnicity of Students at De Anza College (Bar Graph with 'Other' Category)

The mode of the class of Statistics students is obviously Freshman. If any doubt remains a Pareto chart makes identifying the mode trivial, which is Asian in the previous example.

Variability

The best way to gauge variability in categorical data is by thinking about it as *diversity*. Although we will not calculate a numerical measure here, we can note it visually. A variable that has observations spread out fairly evenly over all categories shows high variability, while a variable where most observations are only in one or a handful of categories displays low variability. Consider the level of variability in the two pie charts below.

Example

Consider the level of variability in the two pie charts below. Which college has more variability?

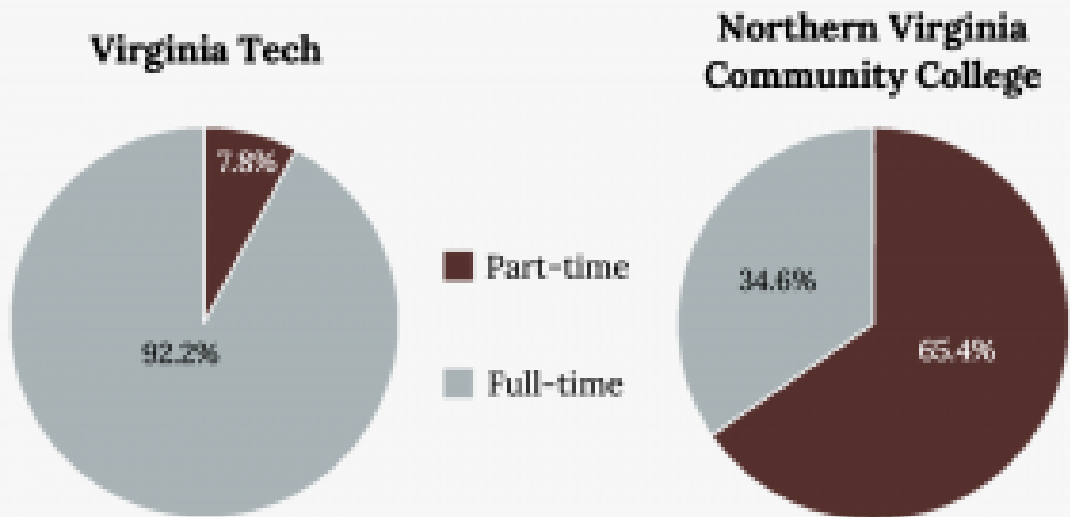


Figure 2.10 (repeat): Full-time and Part-time Students at Virginia Tech and NVCC



An interactive H5P element has been excluded from this version of the text. You can view it



online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=55#h5p-44>

Your turn!

Let's consider the variability in the following bar charts. Which bar chart shows greater variability?

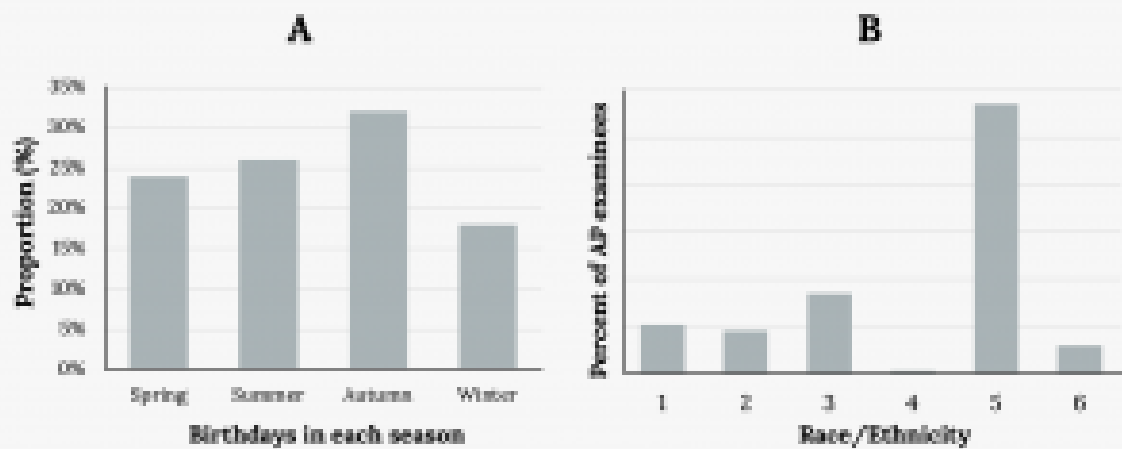


Figure 2.22: Variability Comparisons



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=55#h5p-45>

Image References

Figure 2.7: Kindred Grey (2020). “Classification of Statistics Students.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Classification_of_Statistics_Students.png

Figure 2.9: Kindred Grey (2020). “Ages and proportions.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Ages_and_proportions.png

Figure 2.10: Kindred Grey (2020). “Virginia Tech and NVCC stats.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Virginia_Tech_and_NVCC_stats.png

Figure 2.11: Kindred Grey (2020). “VT and NVCC chart.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:VT_and_NVCC_chart.png

Figure 2.13: Kindred Grey (2020). “Figure 2.13.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.13.png

Figure 2.15: Kindred Grey (2020). “Figure 2.15.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.15.png

Figure 2.16: Kindred Grey (2020). “Figure 2.16.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.16.png

Figure 2.17: Kindred Grey (2020). “Figure 2.17.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.17.png

Figure 2.18: Kindred Grey (2020). “Figure 2.18.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.18.png

Figure 2.20: Kindred Grey (2020). “Figure 2.20.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.20.png

Figure 2.22: Kindred Grey (2020). “Figure 2.22.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.22.png

2.3 Displaying Quantitative Data

Descriptive Statistics for Quantitative Data

Descriptive options for **quantitative data** are much more robust than for categorical. Recall descriptive statistics consists of visual and numerical methods. We usually start with visual methods and then move into numerical.

This section will expand on graphical methods while the next few sections will focus on numerical summaries of quantitative data.

Graphical Methods for Quantitative Data

The first thing we may do, especially for quantitative data, is to examine it in a frequency table. We have many more graphical options beyond that for quantitative data. Some of them we will discuss here are:

- Stem-and-leaf plots
- Dot plots
- Line graphs
- Histograms
- Frequency polygons
- Time series plots

Each of these methods comes with its own pros and cons.

Stem-And-Leaf Plots

One simple graph, the stem-and-leaf graph or stemplot, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a “stem” and a “leaf”. The leaf consists of a final significant digit. For example you could divide the number 23 into a stem two and a leaf of three. The number 432 could have a stem of 43 and leaf of two. The decimal 9.3 could have a stem of nine and leaf of three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

Figure 2.23: Exam 1 Scores

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% ($\frac{8}{31}$) were in the 90s or 100, a fairly high number of As.

The stemplot is a quick way to organize things and gives a good picture of the data. You can quickly and easily find basic summary statistics such as the Maximum, Minimum, range, etc. Also some measures we will explore in the future such as the median and quartiles. They can be good for seeing individual data points and mainly handle discrete or rounded continuous data.

Comparisons with Stem-and-Leaf Plots

Back-to-back or side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems.

Your turn!

The following two tables show the ages of U.S. presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Figure 2.24: Presidential Ages at Inauguration

President	A ge	Preside nt	A ge	Presiden t	A ge	Presiden t	A ge
Washington	57	Fillmore	50	McKinley	54	Nixon	56
J. Adams	61	Pierce	48	T. Roosevelt	42	Ford	61
Jefferson	57	Buchanan	56	Taft	51	Carter	52
Madison	57	Lincoln	52	Wilson	56	Reagan	69
Monroe	58	A. Johnson	56	Harding	55	G.H.W. Bush	64
J. Q. Adams	57	Grant	46	Coolidge	51	Clinton	74
Jackson	61	Hayes	45	Hoover	54	G. W. Bush	54
Van Buren	45	Garfield	49	F. Roosevelt	51	Obama	47
W. H. Harrison	68	Arthur	51	Truman	60	Trump	70
Tyler	51	Cleveland	47	Eisenhower	62	Biden	78
Polk	49	B. Harrison	55	Kennedy	34		
Taylor	64	Cleveland	55	L. Johnson	55		

Figure 2.25: Presidential Ages at Death

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58	G.H.W. Bush	94
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=62#h5p-46>

Line Graphs

Another type of graph that is useful for showing trends in specific data values (**discrete** data) is a line graph. In the particular line graph shown below, the x-axis (horizontal axis) consists of data values and the y-axis (vertical axis) consists of frequency points. The frequency points are connected using line segments.

Side Note: Line graphs could also be used with some **ordinal categorical data**.

Example

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to complete chores. The results are shown in the table and chart below.

Figure 2.26: Chore Reminder Data

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

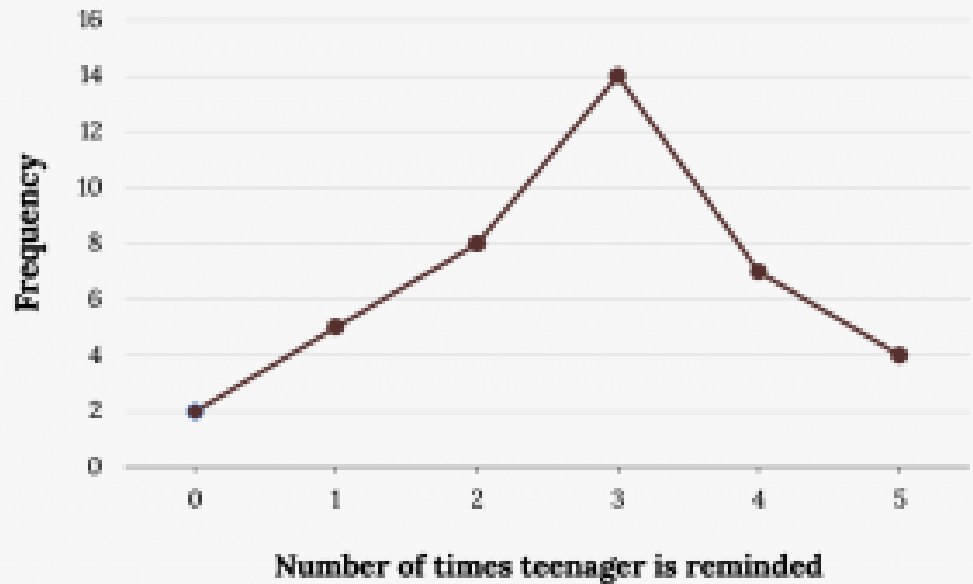


Figure 2.27: Chore Reminder (Line Graph)

Dot Plots

A dot plot consists of a number line and dots (or points) positioned above the number line.

Dot plots are very similar in functionality to stem-leaf-plots, but look a little bit cleaner. Look for an overall pattern and any outliers or extreme values. An outlier is an observation of data that does not fit the rest of the data. When graphed, an outlier will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to fully explain outliers; we will cover them in more detail later.

Example

Consider the following data dealing with the hours of sleep students get per night: 5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9

The dot plot for this data would be as follows:

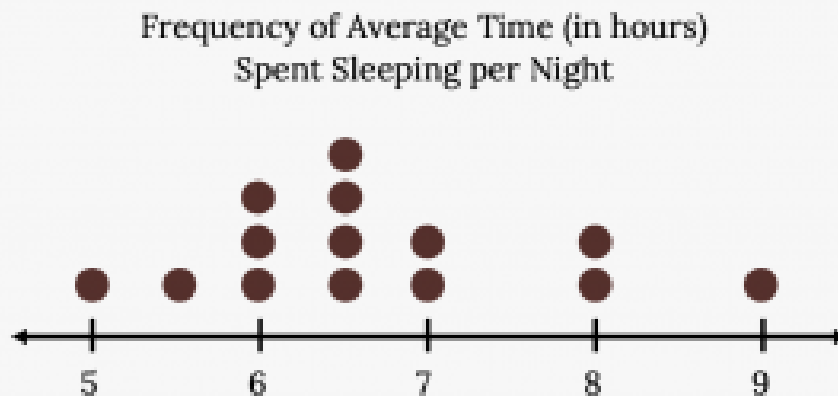


Figure 2.28: Student Sleep Hours

Histograms

For most of the work in this book, histograms will display the data. One advantage of a histogram is that it can

readily display large continuous data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram can give you a really good look at the overall shape of the data, the center, and the spread. However, you do lose individual data points.

A Histogram is essentially a 2-D Frequency table. To construct a histogram, you must first decide the size and number of bars, intervals, or classes, similarly to how you would with a frequency table.

Example

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are continuous data, since height is measured.

60, 60.5, 61, 61, 61.5, 63.5, 63.5, 63.5, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 68, 68, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69.5, 69.5, 69.5, 69.5, 69.5, 70, 70, 70, 70, 70, 70, 70.5, 70.5, 70.5, 71, 71, 71, 72, 72, 72, 72.5, 72.5, 73, 73.5, 74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

$60 - 0.05 = 59.95$ which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so $74 + 0.05 = 74.05$ is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76.$$

NOTE

We will round up to two and make each bar or class interval two units wide. Rounding up

to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the number of bars or class intervals is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

Some values in data sets might fall on boundaries for different intervals. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

The boundaries are:

- 59.95
- $59.95 + 2 = 61.95$
- $61.95 + 2 = 63.95$
- $63.95 + 2 = 65.95$
- $65.95 + 2 = 67.95$
- $67.95 + 2 = 69.95$
- $69.95 + 2 = 71.95$
- $71.95 + 2 = 73.95$
- $73.95 + 2 = 75.95$

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the x -axis and relative frequency on the y -axis.

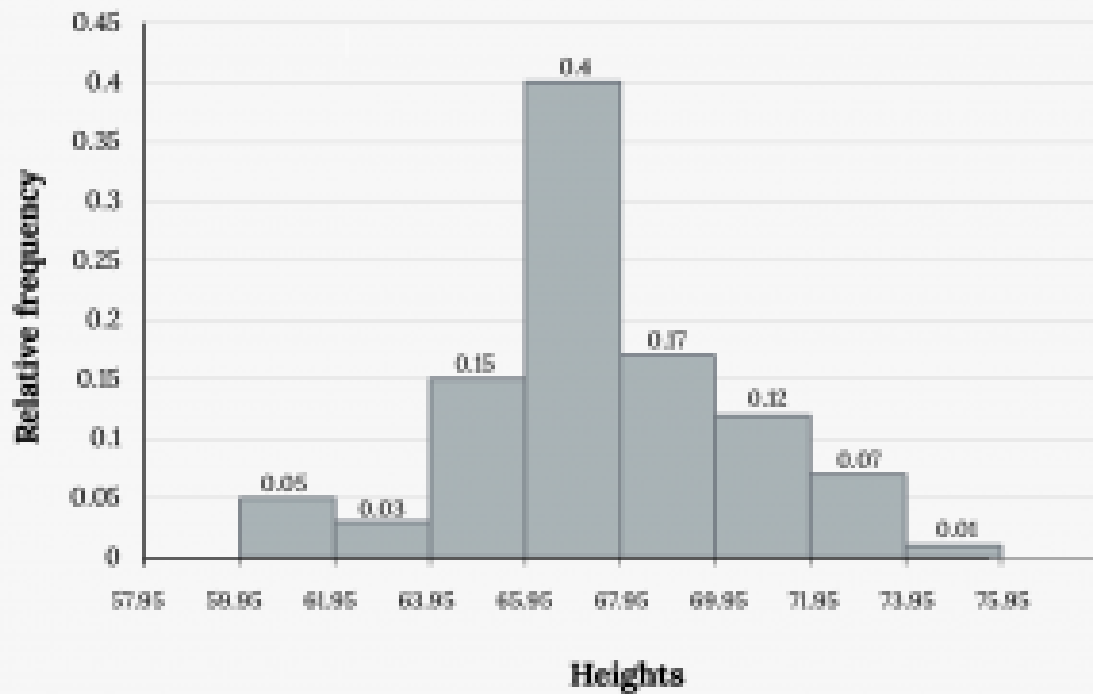


Figure 2.29: Soccer Player Heights

Frequency Polygons

Frequency polygons are analogous to line graphs, but instead utilize binning techniques to make continuous data visually easy to interpret. It is essentially a combination of a histogram and line graph.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Frequency polygons are sometimes more useful for comparing continuous distributions than histograms. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example

A frequency polygon was constructed from the frequency table below.

Figure 2.30: Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

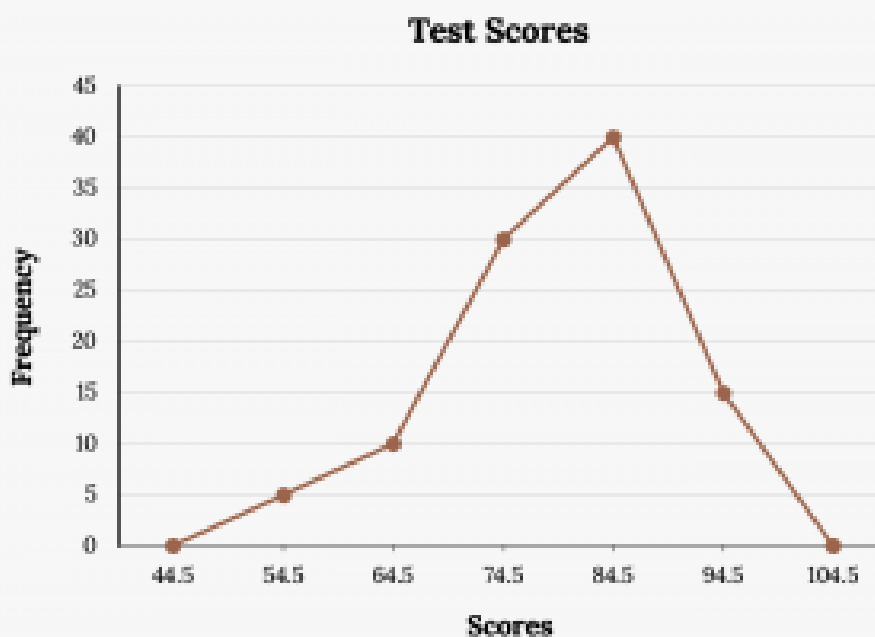


Figure 2.31: Calculus Final Test Scores (Frequency Polygon)

The first label on the x -axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the x -axis. The point labeled 54.5 represents the next interval, or the first “real” interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the x -axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Time Series Plots

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

To construct a time series graph, we must look at both pieces of our paired data set. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Figure 2.32: CPI Data

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.92 ₂
2012	226.655	227.663	229.392	230.085	229.815	229.478	229.104
2013	230.28 ₀	232.166	232.773	232.531	232.945	233.504	233.59 ₆
2014	233.916	234.781	236.293	237.072	237.900	238.343	238.25 ₀
2015	233.707	234.722	236.119	236.599	237.805	238.638	238.65 ₄
2016	236.916	237.111	238.132	239.261	240.236	241.038	240.64 ₇
2017	242.839	243.60 ₃	243.801	244.524	244.733	244.955	244.786
2018	247.867	248.991	249.554	250.546	251.588	251.989	252.00 ₆
2019	251.712	252.776	254.202	255.548	256.092	256.143	256.571

Year	Aug	Sep	Oct	Nov	Dec	Annual
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594
2013	233.877	234.149	233.546	233.069	233.049	232.957
2014	237.852	238.031	237.433	236.151	234.812	236.736
2015	238.316	237.945	237.838	237.336	236.525	237.017
2016	240.853	241.428	241.729	241.353	241.432	240.007
2017	245.519	246.819	246.663	246.669	246.524	245.120
2018	252.146	252.439	252.885	252.038	251.233	251.107
2019	256.558	256.759	257.346	257.208	256.974	255.657

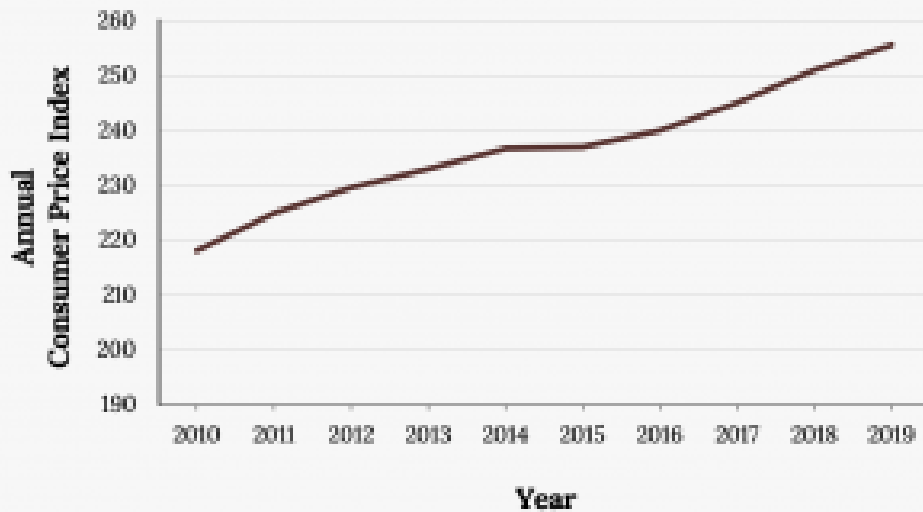


Figure 2.33: CPI Time Series Plot

Image References

Figure 2.27: Kindred Grey (2020). "Figure 2.27." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.27.png

Figure 2.28: Kindred Grey (2020). "Figure 2.28." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.28.png

Figure 2.29: Kindred Grey (2020). "Figure 2.29." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.29.png

Figure 2.31: Kindred Grey (2020). "Figure 2.31." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.31.png

Figure 2.32: Data retrieved from <https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/>

Figure 2.33: Kindred Grey (2020). "Figure 2.33." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.33.png

2.4 Describing Quantitative Distributions

Consider the following exercise:

Your classmates write down the average time (in hours, to the nearest half-hour) they sleep per night and then create a simple dot plot of the data. Suppose the resulting Dot Plot looked like this:

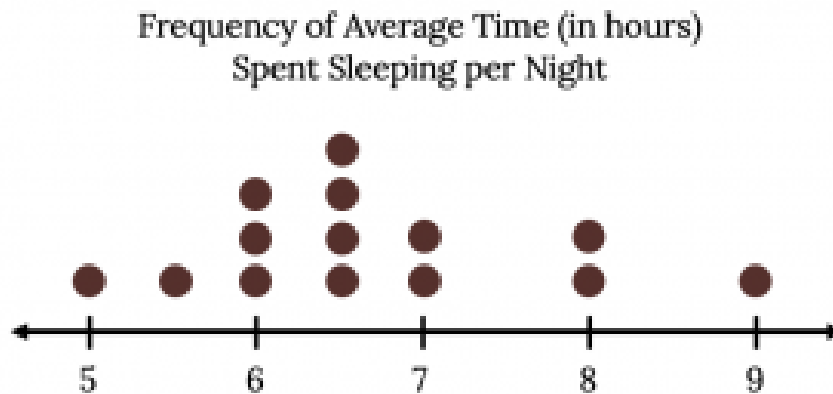


Figure 2.28 (repeat): Student Sleep Hours

How would you interpret or explain this distribution? Where do your data appear to cluster? How might you interpret the clustering? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

The questions above ask you to analyze and interpret your data. It isn't enough to just make graphs, we must be able to interpret it with a critical eye.

Key Aspects of Quantitative Data

When describing a Quantitative Distribution we want to at least note 4 things: The shape of the distribution, the presence of outliers, the center, and the spread. A helpful acronym to remember this is **SOCS**:

- **Shape**
- **Outliers**
- **Center**
- **Spread**

Shape is the main characteristic we can determine by looking at a graph. We are often able to identify potential

outliers visually as well. Center and spread can be roughly gauged visually, but are more numerical calculations for those last two aspects will be discussed in the following sections.

Shape

Shape is the main characteristic we can determine by looking at a graph. The shape of a distribution is the first thing we should note since it will often dictate how to proceed with the rest of our analysis. We have already seen most of our graphical methods can give us an idea the shape of a distribution, but the best in most situations is a properly formatted histogram. Consider the following:

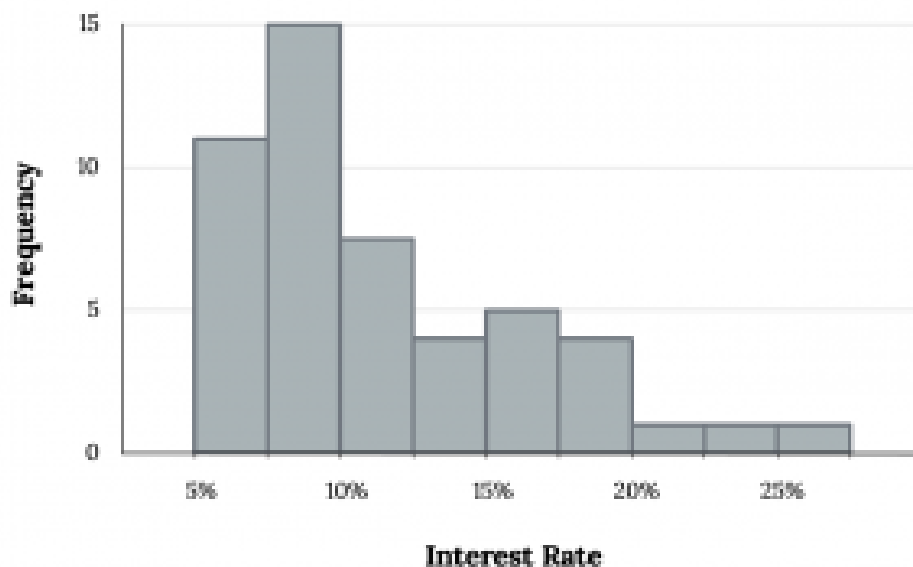


Figure 2.34: Interest Rates

Histograms are especially convenient for understanding the shape of the data distribution. The figure above suggests that most loans have rates under 15%, while only a handful of loans have rates above 20%. When data trail off to the right in this way and has a longer right tail, the shape is said to be right skewed.

Data sets with the reverse characteristic – a long, thinner tail to the left – are said to be left skewed. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off equally in both directions are called symmetric.

Modality

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify the **modality** of a distribution. A mode is represented by a prominent peak in the distribution. There is only one prominent peak in the histogram of loan amount. The definition of mode sometimes taught in math classes is the value with the most occurrences in the data set. However, for many real-world data sets, it is common to have no observations with the same value in a data set, making this definition impractical in data analysis. The figure below shows histograms that have one, two, or three prominent peaks.

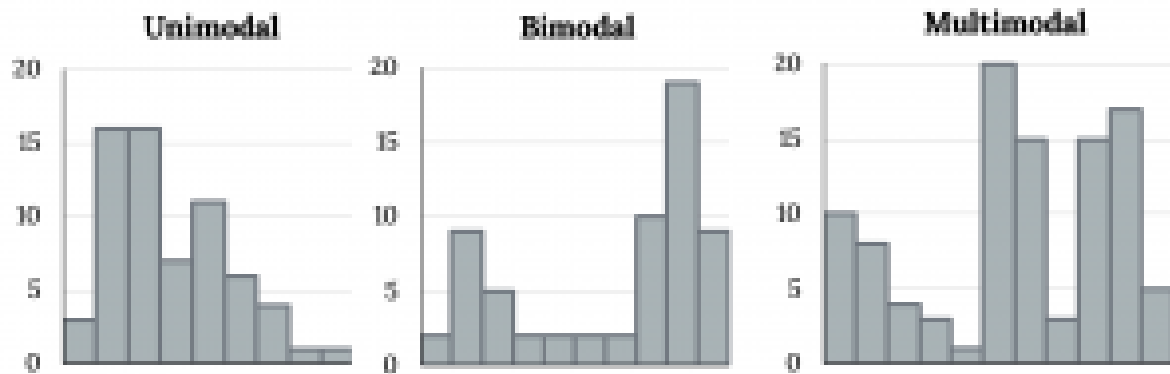


Figure 2.35: Unimodal, Bimodal, and Multimodal Distributions

Such distributions are called unimodal, bimodal, and multimodal, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.

Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why prominent is not rigorously defined in this book. The most important part of this examination is to better understand your data.

Outliers

Sometimes one or more data points that stick out visually. These extreme values could potentially be **outliers**. Sometimes they may be obvious to us as in the following histogram:

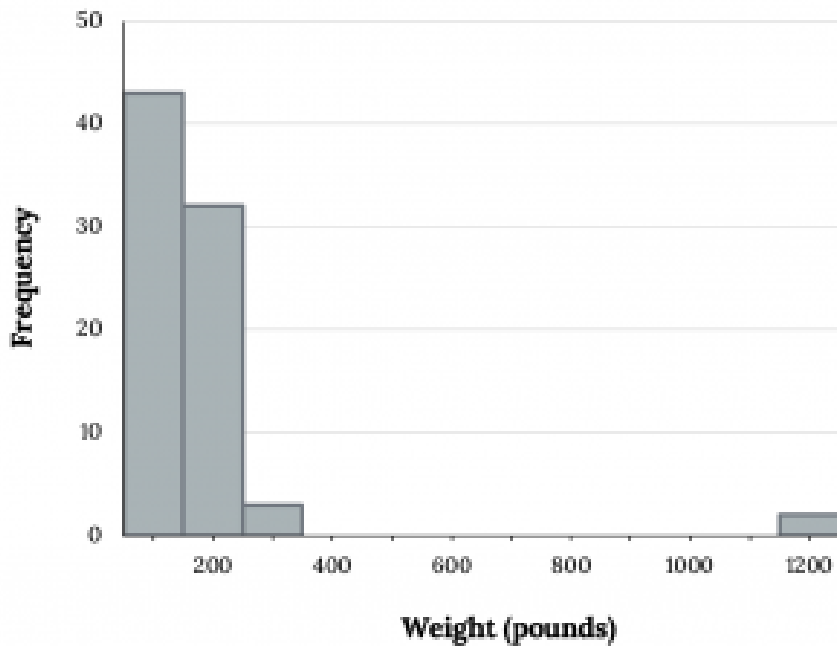


Figure 2.36: Outlier

Or they may not be as obvious and might only show up upon careful examination of a dot plot or other methods. Examining data for outliers serves many useful purposes, including:

1. Identifying skewness in the distribution.
2. Identifying possible data collection or data entry errors.
3. Providing insight into interesting properties of the data.

In subsequent sections we will see numerical methods to “officially” identify outliers and how to deal with them.

Center

We also want to make sure to describe a quantitative distribution’s “central tendency” or most “typical value”. We can simply estimate this visually but will see more robust and appropriate measures we can calculate in the future.

Spread

A rough measure of spread we can usually determine visually is the Range. Recall: $\text{Range} = \text{Maximum} - \text{Minimum}$. Again, we will see more robust and appropriate measures we can calculate in the future.

Example

Use the following graph to answer a-e.

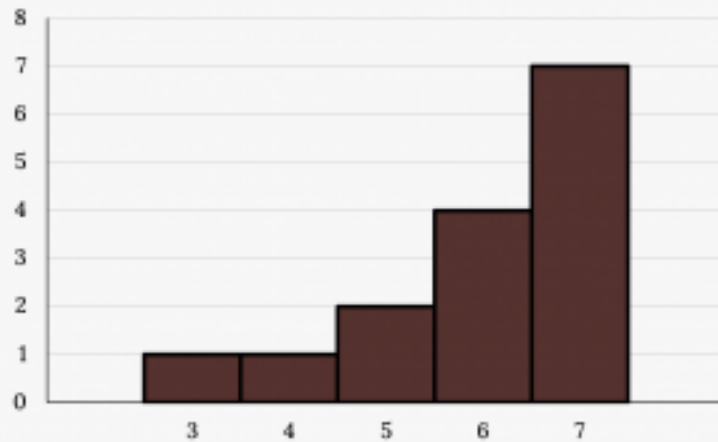


Figure 2.37: Distribution 1

a. Describe the shape of this distribution.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=70#h5p-47>

b. Describe the modality of the distribution.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=70#h5p-48>

c. Do you see any apparent outliers?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=70#h5p-49>

d. What does the center appear to be?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=70#h5p-50>

e. Provide a rough estimate of the spread.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=70#h5p-51>

Your turn!

Describe the shape of this distribution visually:

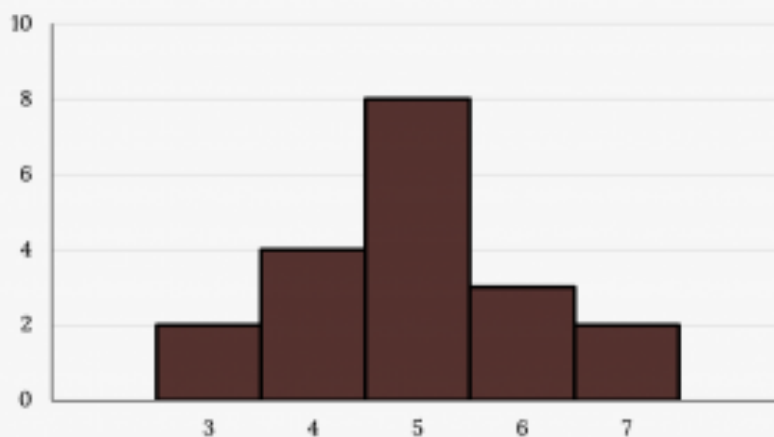


Figure 2.38: Distribution 2

- Are the data symmetric or skewed? If you see skewness in what direction?
- Describe the modality of the distribution.
- Do you see any apparent outliers?
- What does the center appear to be?
- Provide a rough estimate of the spread.

Image References

Figure 2.34: Kindred Grey (2020). "Figure 2.34." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.34.png

Figure 2.35: Kindred Grey (2020). "Figure 2.35." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.35.png

Figure 2.36: Kindred Grey (2020). "Figure 2.36." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.36.png

Figure 2.37: Kindred Grey (2020). "Figure 2.37." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.37.png

Figure 2.38: Kindred Grey (2020). "Figure 2.38." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.38.png

2.5 Measures of Location and Outliers

Let's keep working through our acronym, **SOCS**, to describe the key aspects of our data.

- Shape
- **Outliers**
- Center
- Spread

Measures of location are a tool used to quantify where an observation stands in relation to the rest of the distribution. They also provide the building blocks to formally identify outliers. Common measures of location are quartiles and percentiles. Quartiles divide ordered data into quarters while percentiles divide ordered data into hundredths.

Percentiles

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

There are two inverse ways you may work with percentiles; Finding the k th Percentile of a distribution, or finding the percentile of a given observation.

Finding the k th Percentile of a Distribution

Sometimes we may want to find the “ k th” percentile of a distribution. For instance what would you have to score on the SAT to be in the 90th percentile?

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

NOTE: You can calculate percentiles using calculators and computers. There are a variety of online calculators.

Example

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.
18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

- Find the 70th percentile.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=75#h5p-52>

- Find the 83rd percentile.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=75#h5p-53>

Your turn!

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Calculate the 20th percentile and the 55th percentile.

Finding the Percentile of a Value in a Data Set

To find the corresponding percentile of a given observation the process is as follows:

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

- Find the percentile for 58.



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=75#h5p-54>

- Find the percentile for 25.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=75#h5p-55>

Your turn!

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

Find the percentiles for 47 and 31.

Quartiles

Quartiles again deal with an ordered dataset and are really just special percentiles. The first quartile, Q_1 , is the same as the 25th percentile. The second quartile, Q_2 , is the same as the 50th percentile, and is also called the **Median**. and the third quartile, Q_3 , is the same as the 75th percentile.

The Median

The median is a number that measures the “halfway point” of the data. You can think of the median as the “middle value,” but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data: 1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1

Ordered from smallest to largest: 1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8+7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Depending on the context, the median could be both a measure of location and/or center. We'll discuss more on the Median and using it as a measure of center in the future.

Finding Quartiles

Quartiles can be found by either treating them as a percentile or in a similar fashion to the median. They may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data, and the third quartile, Q_3 , is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set: 1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The median or second quartile is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two. 1, 1, 2, 2, 4, 6, 6.8

The number two, which is part of the data, is the first quartile. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The third quartile, Q_3 , is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the p^{th} percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.” The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered “good;” in other contexts a high percentile might be considered “good”. In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

NOTE: When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile

Example

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Your turn!

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Five Number Summary

The Five Number summary is a simple, easy way to quickly summarize a data set. It consists of:

1. Minimum
2. Q_1
3. Median
4. Q_3
5. Maximum

Example

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes, 40 minutes, 60 minutes, 30 minutes, 60 minutes, 10 minutes, 45 minutes, 30 minutes, 300 minutes, 90 minutes, 30 minutes, 120 minutes, 60 minutes, 0 minutes, 20 minutes

Determine the following five values.

- Min = 0
- $Q_1 = 20$
- Med = 40
- $Q_3 = 60$
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the IQR is 40 minutes ($60 - 20 = 40$), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems

a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$$

Interquartile Range

The interquartile range is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The IQR is also helpful to determine potential **outliers**. It can also be used as a measure of spread and will be discussed further.

Fence Rule

Although points may often look like outliers on a graph, we establish **the upper and lower fences** to numerically decide if a value is an outlier. The lower fence is 1.5 times the **IQR** below the first quartile ($LF = Q_1 - 1.5 \cdot IQR$) while the upper fence is 1.5 times the **IQR** above the third quartile ($UF = Q_3 + 1.5 \cdot IQR$). If a value falls outside of these fences, i.e. less than the lower fence or greater than the upper fence, we will flag it as an outlier.

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data. Potential outliers always require further investigation.

Example

[Continued from Sharpe Middle School example above]

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- $Q_1 = 20$

- $Q_3 = 60$
- $\text{Max} = 120$

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

Box Plots

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from the five number summary. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box**. The “whiskers” extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

NOTE: You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values because they have been identified as outliers according to the fence rules.

Example

Consider, again, this dataset.

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.



Figure 2.39: Box Plot

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE: It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful

Example

The following data are the heights of 40 students in a statistics class.

59, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68, 68, 69, 70, 70, 70, 70, 70, 71, 71, 72, 72, 73, 74, 74, 75, 77

Construct a box plot with the following properties.

- Minimum value = 59
- Maximum value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median = 66
- Q3: Third quartile = 70

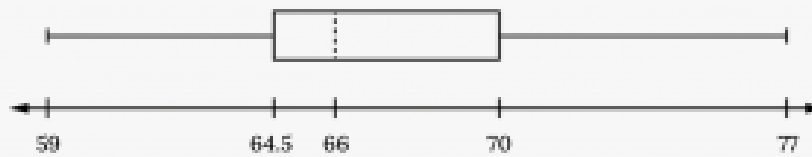


Figure 2.40: Student Heights (Box Plot)

- Each quarter has approximately 25% of the data.
- The spreads of the four quarters are $64.5 - 59 = 5.5$ (first quarter), $66 - 64.5 = 1.5$ (second quarter), $70 - 66 = 4$ (third quarter), and $77 - 70 = 7$ (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- Range = maximum value - the minimum value = $77 - 59 = 18$
- Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
- The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- The middle 50% (middle half) of the data has a range of 5.5 inches.

Your turn!

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136, 140, 178, 190, 205, 215, 217, 218, 232, 234, 240, 255, 270, 275, 290, 301, 303, 315, 317, 318, 326, 333, 343, 349, 360, 369, 377, 388, 391, 392, 398, 400, 402, 405, 408, 422, 429, 450, 475, 512

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:

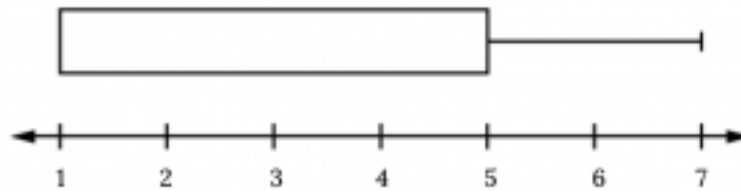


Figure 2.41: Box Plot With The Same Values

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Image References

Figure 2.39: Kindred Grey via Virginia Tech (2020). “Figure 2.39” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.39.png . Adaptation of Figure 2.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-4-box-plots>

Figure 2.40: Kindred Grey via Virginia Tech (2020). “Figure 2.40” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.40.png . Adaptation of Figure 2.12 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-4-box-plots>

Figure 2.41: Kindred Grey via Virginia Tech (2020). “Figure 2.41” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.41.png . Adaptation of Figure 2.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-4-box-plots>

2.6 Measures of Center

Continuing through our acronym, **SOCS**, to describe the key aspects of our data:

- Shape
- Outliers
- **Center**
- Spread

The “center”, is a way of describing “central tendency” or “typical value” of a data set. The two most widely used measures of the “center” of the data are the **mean (average)** and the **median**. Most people are familiar with the ideas of these two; To calculate the mean weight of 50 people, add the 50 weights together and divide by 50. To find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts.

However, some datasets may be better summarized by one or the other. The most “appropriate” measure of center depends on the shape of the distribution and presence of extreme values or potential outliers.

The Mean

The mean is the most common measure of the center. The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced “ x bar”): \bar{x} .

The Greek letter μ (pronounced “mew”) represents the **population mean**. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample taken to be truly random. We will often use the sample mean to estimate the population mean.

Example

Calculate the mean of the sample: 1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-56>

Your turn!

Calculate the mean of the sample: 7, 10, 14, 14, 15, 21, 38, 38, 38, 56.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-57>

The Median

The median is generally a better measure of the center when there are extreme values or outliers because it is more **robust**, or not affected by the precise numerical values of those outliers.

In many cases, especially for larger datasets you may choose to use the location function of the median rather than the traditional counting method. $\frac{n+1}{2}$.

Remember that this function simply tells you where to look for the median, not the actual value itself.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered.

For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The

median occurs midway between the 50th and 51st values. The location of the median and the value of the median are not the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest): 3, 4, 8, 8, 10, 11, 12, 13, 14, 15, 15, 16, 16, 17, 17, 18, 21, 22, 22, 24, 24, 25, 26, 26, 27, 27, 29, 29, 31, 32, 33, 33, 34, 34, 35, 37, 40, 44, 44, 47. Calculate the median.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-58>

Your turn!

Calculate the median of the sample: 7, 10, 14, 14, 15, 21, 38, 38, 38, 56.

The Mode

Another measure of the center is the **mode**. The mode is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal. For example, five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the “center”? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE: The mode can be calculated for categorical data as well as quantitative data but has different uses and interpretation. For example:

- If we had the categorical data set {red, red, red, green, green, yellow, purple, black, blue} the mode is red. This is useful to us
- If we had the quantitative data set {1.0, 2.1, 2.1, 5.0, 5.1, 5.5, 5.7, 6.1, 6.2, 6.4, 6.6, 7.1, 7.8, 8.1, 8.9} the numerical mode is 2.1, but does not do a good job of telling us about the actual **modality**, or where the data is clustered.

Example

Statistics exam scores for 20 students are as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

Find the mode.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-59>

Order Relationship of Measures of Center

Consider the following data set: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10.

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

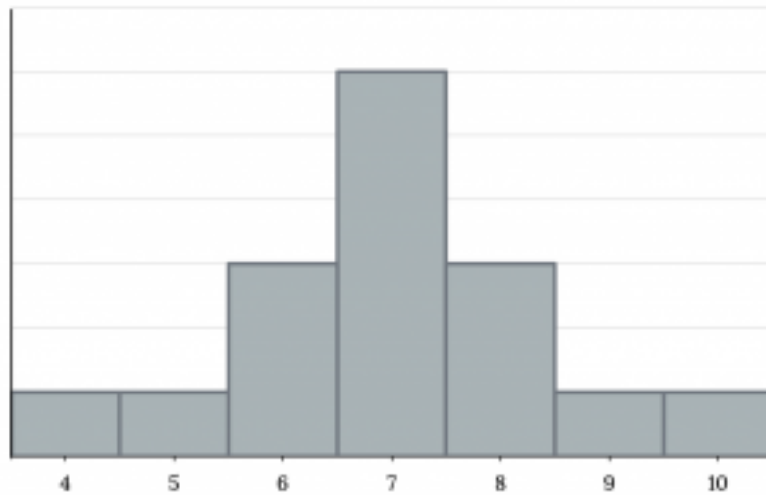


Figure 2.42: Symmetrical Distribution

The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data $\{4, 5, 6, 6, 6, 7, 7, 7, 7, 8\}$ is not symmetrical. The right-hand side seems “chopped off” compared to the left side. A distribution of this type is called skewed to the left because it is pulled out to the left.

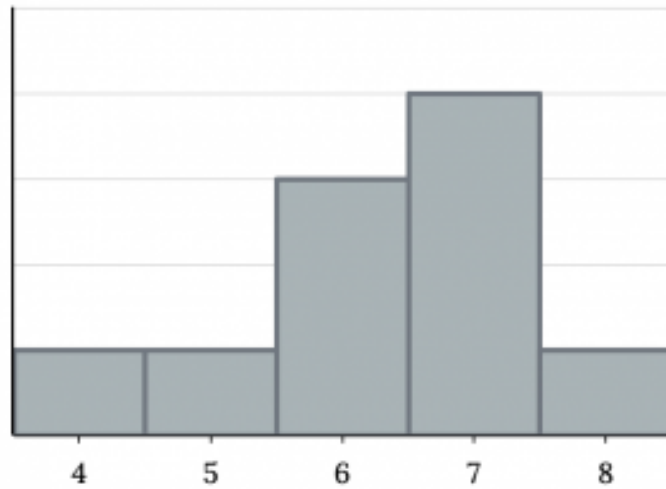


Figure 2.43: Skewed Left

The mean is 6.3, the median is 6.5, and the mode is seven. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data {6, 7, 7, 7, 7, 8, 8, 8, 9, 10} is also not symmetrical. It is skewed to the right.

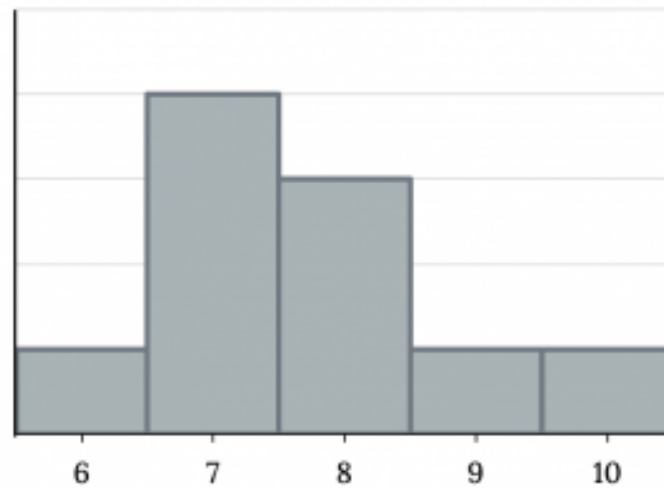


Figure 2.44: Skewed Right

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Darnell: 7, 9, 3, 3, 3, 4, 1, 3, 2, 2

Mary: 3, 3, 3, 4, 1, 4, 3, 2, 3, 1

Lee: 2, 3, 4, 4, 4, 6, 6, 6, 8, 3

a. Make a dot plot for the three authors and compare the shapes.

Darnell's distribution has a right (positive) skew.

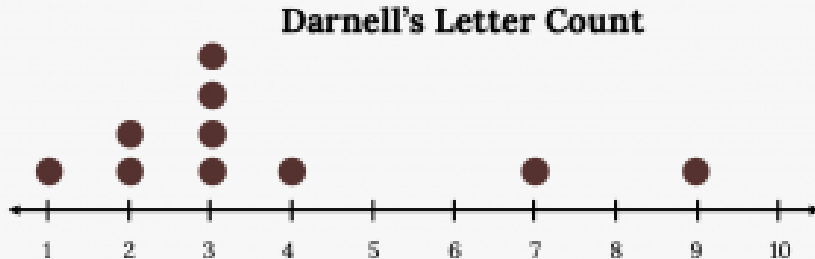
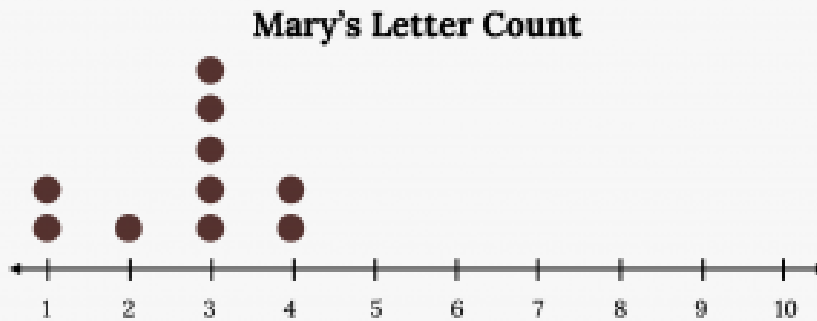


Figure 2.45: Darnell's Letter Count

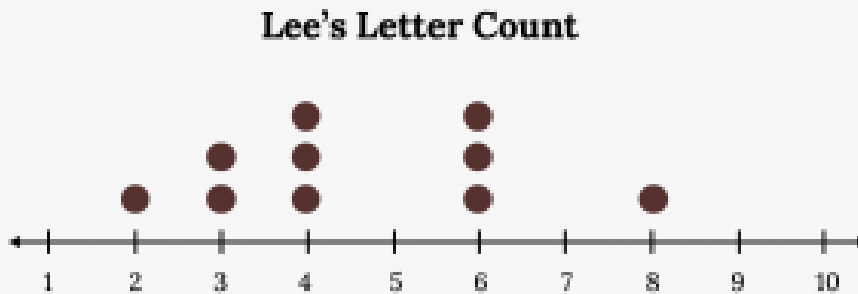
Mary's distribution has a left (negative) skew

Figure 2.46: Mary's Letter Count



Lee's distribution is symmetrically shaped.

Figure 2.47: Lee's Letter Count



b. Calculate the mean for each.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-60>

c. Calculate the median for each.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-61>

d. Describe any pattern you notice between the shape and the measures of center.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-62>

Your Turn!

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the “center”: the mean or the median?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-63>

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals

and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: $mean = \frac{\text{data sum}}{\text{number of data values}}$. We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{\text{lower boundary} + \text{upper boundary}}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example

A frequency table displaying professor Blount's last statistic test is shown.

- Find the best estimate of the class mean.

Figure 2.48: Blount's Statistics Test

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Find the midpoints for all intervals



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-64>

Calculate the sum of the product of each interval frequency.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-65>

Calculate the midpoint.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=83#h5p-66>

Your turn!

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Figure 2.49: Video Game Data

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Image References

Figure 2.42: Kindred Grey (2020). “Figure 2.42.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.42.png

Figure 2.43: Kindred Grey (2020). “Figure 2.43.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.43.png

Figure 2.44: Kindred Grey (2020). “Figure 2.44.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.44.png

Figure 2.45: Kindred Grey via Virginia Tech (2020). “Figure 2.45” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.45.png . Adaptation of Figure 2.21 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode>

Figure 2.46: Kindred Grey via Virginia Tech (2020). “Figure 2.46” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.46.png . Adaptation of Figure 2.22 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode>

Figure 2.47: Kindred Grey via Virginia Tech (2020). “Figure 2.47” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.47.png . Adaptation of Figure 2.23 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode>

2.7 Measures of Spread

We're now to the final key aspect of our acronym, **SOCS**:

- Shape
- Outliers
- Center
- **Spread**

A complement to the center of a distribution is the **variation, variability, or spread** of the data. In some data sets, the values are concentrated closely, while in others they are more spread out. Some rough measures of spread we have already seen are the range and IQR. The most common measure of spread is the standard deviation.

Similar to measures of center, the shape of the distribution and presence of extreme values can dictate what the most appropriate measure of spread is to describe the distribution.

The Interquartile Range

Recall the Interquartile Range (IQR):

$$\text{IQR} = Q_3 - Q_1.$$

In addition to helping us establish our fences and identify outliers, the IQR indicates the spread of the middle half or the middle 50% of the data. The IQR can be used as a somewhat rough but very robust measure of spread when outliers may be present. It is often used alongside the median to describe the center and spread of skewed distributions.

Simply showing the five number summary or a Box Plot can be a good way to get all of the information for a skewed dataset in one place

The Standard Deviation

The **standard deviation** is a measure of spread that measures how spread out values are from their mean. It is essentially the “average” deviation, or distance of each observation from the mean.

Not only does it provide a numerical measure of the overall amount of variation in a data set, it can also be used for other purposes

The lower case letter s represents the **sample** standard deviation and the lower case greek letter σ (sigma) represents the **population** standard deviation.

By extension, s^2 represents the sample **variance** and the lower case greek letter σ^2 represents the population variance. The variance is useful

The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. It must always greater than or equal to zero.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. the average wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes; at supermarket B the standard deviation for the wait time is four minutes.

Because supermarket B has a higher standard deviation, we know that there is more **variation** in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average.

Calculating the Standard Deviation

The procedure to calculate the standard deviation can be tedious and depends on whether the data are from the entire population or a sample. The calculations are similar, but not identical.

If x is a number, then the difference “ $x - \text{mean}$ ” is called its deviation. In a data set, there are as many deviations as there are items in the data set. The deviations can show how spread out the data are about the mean. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$. If you add the deviations, the sum is always zero, so you cannot simply add the deviations to get the spread of the data. You can fix this by squaring the deviations, making them positive numbers, therefore sum will also be positive.

The variance is the average of the squares of the deviations (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The variance, then, is the average squared deviation, which we use to get the standard deviation. The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N , the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample. Why not divide by n for a sample? The answer has to do with the population variance. The sample variance is an estimate of the population variance. Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

Formulas

The sample standard deviation

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

The population standard deviation

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

NOTES:

- The variance, population (σ^2) or sample (s^2), can be obtained if you do not apply the square root in their respective formulas
- In practice, we typically rely on technology to calculate the standard deviation. However please notice:
 - In the sample standard deviation formula, the denominator is **$n - 1$**
 - In the population standard deviation formula, the denominator is **N**
 - You may need to indicate on your technology of choice which form of the formula you want to use.
- We will often use the sample standard deviation or variance to estimate the population standard deviation or variance.

Example

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year: 9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11.5, 11.5.

First, try to find the mean and standard deviation by hand. Here is a table with the intermediate steps:

Figure 2.51: Fifth Grade Ages

X	Deviations	Deviations²
9	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$
9.5	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$
9.5	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
$\frac{10}{5}$	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
$\frac{10}{5}$	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
$\frac{10}{5}$	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
$\frac{10}{5}$	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
$\frac{11}{5}$	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$
$\frac{11}{5}$	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$
$\frac{11}{5}$	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$
-	-	The total is 9.7375

Verify your answers with your choice of technology.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=86#h5p-67>

Your turn!

On a baseball team, the ages of each of the players are as follows:

21, 21, 22, 23, 24, 24, 25, 25, 28, 29, 29, 31, 32, 33, 33, 34, 35, 36, 36, 36, 36, 38, 38, 38, 40

First, try to find the mean and standard deviation by hand. If you get stuck or want to check your work, plug it into your calculator or use your computer software.

The standard deviation, s or σ , is either zero or larger than zero. Describing the data with reference to the spread is called “variability”. The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The Standard Deviation in Context

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better “feel” for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, always graph your data. Display your data in a histogram or a box plot.

A number line may also help you understand standard deviation. If we were to put five and seven on a number

line, seven is to the right of five. We say, then, that seven is one standard deviation to the right of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is two standard deviations to the left of five because $5 + (-2)(2) = 1$.



Figure 2.50: Number Line

- In general, a value = mean + (#ofSTDEV)(standard deviation)
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is two standard deviations less than the mean of five because: $1 = 5 + (-2)(2)$.

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- Sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
- Population: $x = \mu + (\text{\#ofSTDEVs})(\sigma)$

Example

Suppose that Rosa and Binh both shop at supermarket A. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket A, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is two minutes longer than the average of five minutes.
- Rosa's wait time of seven minutes is one standard deviation above the average of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is four minutes less than the average of five minutes.
- Binh's wait time of one minute is two standard deviations below the average of five minutes.

Your turn!

Recall the previous example about the age of fifth grade students where $\bar{x} = 10.525$ and $s^2 = 0.7159$

b. Find the value that is one standard deviation above the mean. Find $(\bar{x} + 1s)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=86#h5p-68>

c. Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=86#h5p-69>

d. Find the values that are 1.5 standard deviations from (below and above) the mean.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=86#h5p-70>

Z-scores

The standard deviation can also be used to calculate a measure of location called a **z-score**. It represents the number of Standard deviations a given observation is away from it's mean (#ofSTDEVs above) is often denoted with just the letter z. In symbols, the formulas become:

Figure 2.52: Z-Score Formulas

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Not only are Z scores a useful measure of location for specific observations, they can also be used for other purposes. Suppose two data sets have different means and standard deviations, then comparing the data values directly can be misleading. However using Z scores, it is possible to put things on a level playing field to compare them.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\text{\#ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain relative to their means.

Example

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Figure 2.53: GPA Comparisons

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \text{\#ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=86#h5p-71>

Your turn!

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Figure 2.54: Swim Time Comparisons

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Identifying Unusual Values with the Standard Deviation

The following rules give more insight into how we can use the standard deviation to tell us about the distribution of the data.

Chebyshev's Rule

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.

“Unusual” Observations with Z scores

Recall we have already established our Fence Rules for identifying outliers. However for many distributions, anything outside of 2 standard deviations (a Z-score below -2 or greater than 2) is considered “unusual”. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate “rule of thumb” than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

Image References

Figure 2.50: Kindred Grey via Virginia Tech (2020). “Figure 2.50” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_2.50.png . Adaptation of Figure 2.26 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-7-measures-of-the-spread-of-the-data>

Chapter 2 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-72>

Section Reviews

2.1 Introduction to Descriptive Statistics and Frequency Tables

Descriptive statistics are ways of organizing summarizing and presenting data. There are two main types: visual and numerical. Usually we want to first examine a dataset visually then describe it numerically. Appropriate methods often depend on the type of data you are working with, however frequency tables are a quick easy way to organize any type of data.

2.2 Displaying and Describing Categorical Data

Two basic visual methods we have for displaying categorical statistics are:

- Pie charts
- Bar charts

When describing a categorical distribution we want to note:

- Mode
- Level of variability (diversity)

2.3 Displaying Quantitative Data

The following are common methods of displaying quantitative data

- Stem-and-leaf plots
- Dot plots
- Line graphs
- Histograms
- Frequency polygons
- Time series plots

Some work better to show certain aspects, or for different sample sizes than others.

2.4 Describing Quantitative Distributions

When describing a quantitative distribution we want to at least note 4 things: the shape of the distribution, the presence of outliers, the center, and the spread. A helpful acronym to remember this is **SOCS**:

- **Shape** – Can be identified visually, want to note symmetry or lack thereof (skewness) and modality
- **Outliers** – Extreme outliers can be seen visually
- **Center** – Central tendency can be estimated visually
- **Spread** – Dispersion can be estimated visually and roughly quantified with the range

2.5 Measures of Location and Outliers

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other observations in the set.

$$i = \left(\frac{k}{100} \right) (n+1)$$

Where:

- i = the ranking or position of a data value,

- k = the k th percentile,
- n = total number of data.

Expression for finding the percentile of a data value:

$$\left(\frac{x + 0.5y}{n} \right) (100)$$

Where:

- x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,
- y = the number of data values equal to the data value for which you want to find the percentile,
- n = total number of data

Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile, the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the 75th percentile.

The interquartile range, or IQR, is the range of the middle 50 percent of the data values. The IQR is found by subtracting Q_1 from Q_3 , and can help determine outliers by using the following fence rules.

- $Upper\ fence = Q_3 + IQR(1.5)$
- $Lower\ fence = Q_1 - IQR(1.5)$

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

2.6 Measures of Center

The mean and the median can be calculated to help you find the “center” of a data set. The mean may often be the best representation of the center of a dataset, but the median is often more appropriate when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set.

The mean of a dataset can be approximated from a frequency table by:

$$\mu = \frac{\sum fm}{\sum f}$$

Where:

- f = interval frequencies
- m = interval midpoints.

2.7 Measures of Spread

The variance and standard deviation are numerical measures of the spread or dispersion of a dataset. There are different equations to use if you are calculating the standard deviation of a sample or of a population. You find the sample and population standard deviations, respectively:

- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$
- $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

To find the standard deviation of a frequency table:

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \text{ where } \begin{matrix} s_x = \text{sample standard deviation} \\ \bar{x} = \text{sample mean} \end{matrix}$$

Z-scores are a measure of location that puts an observation in units of standard deviations relative to the mean. We can use these to compare things from different distributions.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

2.1 Introduction

- Descriptive statistics
- Graphical descriptive methods
- Numerical descriptive methods
- Distribution
- Frequency
- Relative frequency
- Cumulative relative frequency
- Lower class limit
- Upper class limit
- Class width
- Class midpoint

2.2 Displaying and Describing Categorical Data

- **Categorical data**
- **Mode**
- **Variability**

2.3 Displaying Quantitative Data

- **Quantitative data**
- **Discrete data**
- **Ordinal categorical data**

2.4 Describing Quantitative Distributions

- **Shape**
- **Outliers**
- **Center**
- **Spread**
- **Modality**

2.5 Measures of Location and Outliers

2.6 Measures of Center

- **Mean (average)**
- **Median**
- **Sample mean**
- **Population mean**
- **Robust**
- **Mode**
- **Modality**

2.7 Measures of Spread

- Variation (variability, spread)
- Standard deviation
- Sample
- Population
- Variance
- Population
- Sample
- Z-score

Extra Practice

2.1 Introduction

1. The two types of descriptive statistical methods are:

Answer:

- Graphical
 - Numerical
-

2.2 Displaying and Describing Categorical Data

1. The two basic options for graphing categorical data are

Answer:

- Graphical
- Numerical

2. When describing categorical data we want to note:

Answer:

- Mode
- Level of variability

2. When describing the level of variability in categorical data we want to think about it as:

Answer:

- Diversity

2.3 Displaying Quantitative Data

1. Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1, 1, 1, 1, 1, 1, 1, 1, 1, 1

2, 2, 2, 2, 2, 2, 2, 2, 2, 2

3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3

4, 4, 4, 4, 4, 4

5, 5, 5, 5, 5

6, 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-73>

Calculate the number of bars as follows: $\frac{6.5-0.5}{\text{number of bars}} = 1$.

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.

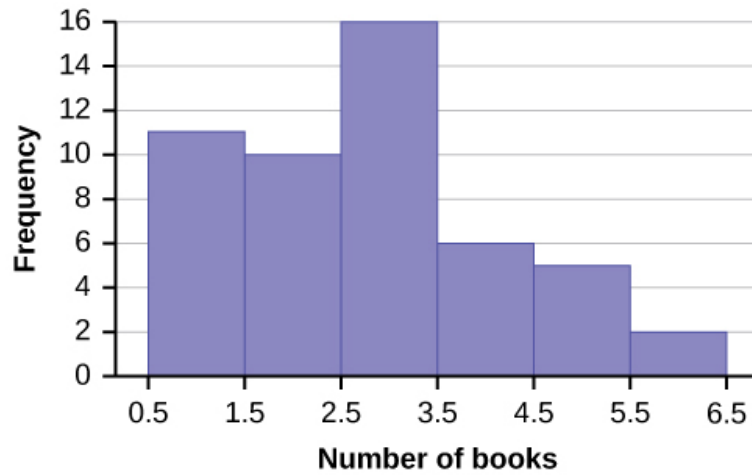


Figure 2.55

2. We will construct an overlay frequency polygon comparing the scores from the figure below with the students' final numeric grade.

Figure 2.56: Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Figure 2.57: Frequency Distribution for Calculus Final Grades

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

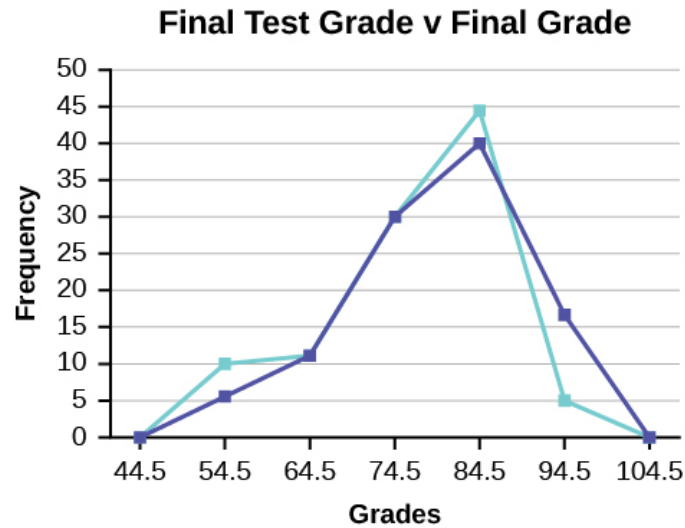


Figure 2.58

3. Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in the figure below.¹

Figure 2.59

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	3

4. Construct a frequency polygon for the following:

1. "Presidents." Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).

a. **Figure 2.60**

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b. **Figure 2.61**

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c. **Figure 2.62**

Tar (mg) in Non-filtered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

5. Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.²

2. “Food Security Statistics.” Food and Agriculture Organization of the United Nations. Available online at <http://www.fao.org/economic/ess/ess-fs/en/> (accessed April 3, 2013).

Figure 2.63

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

6. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlaid frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?³

Figure 2.64

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Figure 2.65

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

7. The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO₂ emissions for the United States.⁴

Figure 2.66: CO₂ Emissions

	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

8. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.⁵

4. "CO₂ emissions (kt)." The World Bank, 2013. Available online at <http://databank.worldbank.org/data/home.aspx> (accessed April 3, 2013).

5. "Births Time Series Data." General Register Office For Scotland, 2013. Available online at <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html> (accessed April 3, 2013).

Figure 2.67

	Female	Male	Total
1855	45,545	47,804	93,349
1856	49,582	52,239	101,821
1857	50,257	53,158	103,415
1858	50,324	53,694	104,018
1859	51,915	54,628	106,543
1860	51,220	54,409	105,629
1861	52,403	54,606	107,009
1862	51,812	55,257	107,069
1863	53,115	56,226	109,341
1864	54,959	57,374	112,333
1865	54,850	58,220	113,070
1866	55,307	58,360	113,667
1867	55,527	58,517	114,044
1868	56,292	59,222	115,514
1869	55,033	58,321	113,354
1870	56,431	58,959	115,390
1871	56,099	60,029	116,128
1872	57,472	61,293	118,765
1873	58,233	61,467	119,700
1874	60,109	63,602	123,711
1875	60,146	63,432	123,578

Solution:

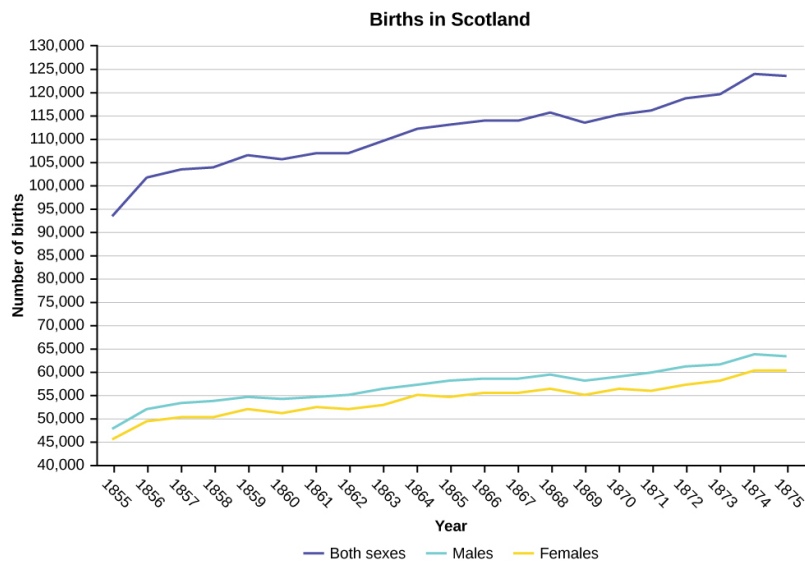


Figure 2.68

9. The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.⁶

6. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason’s book ‘Regression Analysis and its Application’, Marcel Dekker

Figure 2.69

	Police	Homicides
1961	260.35	8.6
1962	269.8	8.9
1963	272.04	8.52
1964	272.96	8.89
1965	272.51	13.07
1966	261.34	14.57
1967	268.89	21.36
1968	295.99	28.03
1969	319.87	31.49
1970	341.43	37.39
1971	356.59	46.26
1972	376.69	47.24
1973	390.19	52.33

- Construct a double time series graph using a common x-axis for both sets of data.
 - Which variable increased the fastest? Explain.
 - Did Detroit's increase in police officers have an impact on the murder rate? Explain.
-

2.4 Describing Quantitative Distributions

2.5 Measures of Location and Outliers

- Test scores for a college statistics class held during the day are: 99, 56, 78, 55.5, 32, 90, 80, 81, 56, 59, 45, 77, 84.5, 84, 70, 72, 68, 32, 79, 90. Test scores for a college statistics class held during the evening are: 98, 78, 68, 83, 81, 89, 88, 76, 65, 45, 98, 90, 80, 84.5, 85, 79, 78, 98, 90, 79, 81, 25.5.⁷
 - Find the smallest and largest values, the median, and the first and third quartile for the day class.
 - Find the smallest and largest values, the median, and the first and third quartile for the night class.
 - For each data set, what percentage of the data is between the smallest value and the first quartile? the

first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?

- d. Create a box plot for each set of data. Use one number line for both box plots.
- e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Solutions:

- Min = 32
- $Q_1 = 56$
- $M = 74.5$
- $Q_3 = 82.5$
- Max = 99

- Min = 25.5
- $Q_1 = 78$
- $M = 81$
- $Q_3 = 89$
- Max = 98

- b. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%.
Night class:

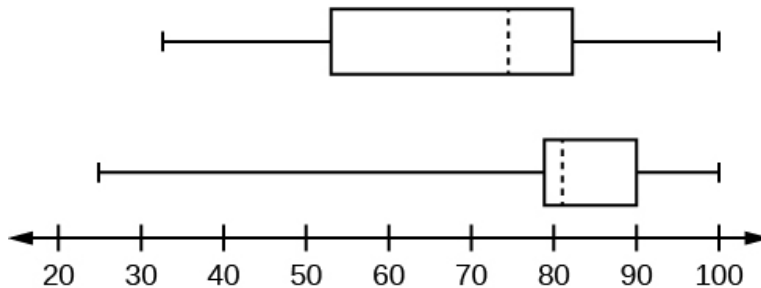


Figure 2.70

- c. 8
- d. The first data set has the wider spread for the middle 50% of the data. The IQR for the first data set is greater than the IQR for the second set. This means that there is more variability in the middle 50% of the first data set.

2. The following data set shows the heights in inches for the boys in a class of 40 students: 66, 66, 67, 67, 68, 68, 68, 68, 69, 69, 69, 70, 71, 72, 72, 72, 73, 73, 74. The following data set shows the heights in inches for the girls in a class of 40 students: 61, 61, 62, 62, 63, 63, 63, 65, 65, 65, 66, 66, 66, 67, 68, 68, 68, 69, 69, 69. Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

3. Graph a box-and-whisker plot for the data values shown.
10, 10, 10, 15, 35, 75, 90, 95, 100, 175, 420, 490, 515, 515, 790
The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q₁: 15
- Med: 95
- Q₃: 490
- Max: 790

Solution: The following graph shows the box-and-whisker plot.

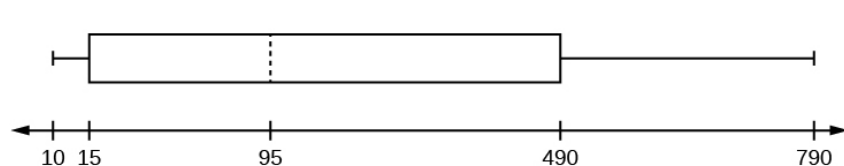


Figure 2.71

4. Graph a box-and-whisker plot for the data values shown.
0, 5, 5, 15, 30, 30, 45, 50, 50, 60, 75, 110, 140, 240, 330

5. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, nineteen generally sell four cars, twelve generally sell five cars, nine generally sell six cars, and eleven generally sell seven cars.

- Construct a box plot below. Use a ruler to measure and scale accurately.
- Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Solution: More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the

box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

6. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.

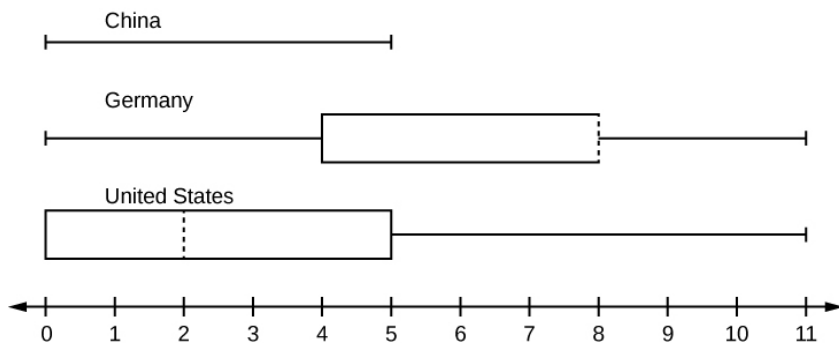


Figure 2.72

- In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
 - Have more Americans or more Germans surveyed been to over eight foreign countries?
 - Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?
-

7. Given the following box plot, answer the questions.

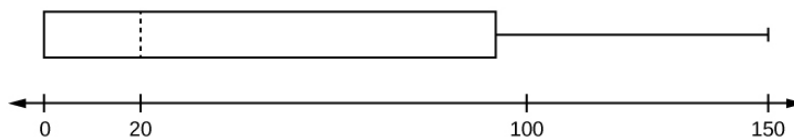


Figure 2.73

- Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
- What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

- a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
- b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.
-

8. Given the following box plots, answer the questions.

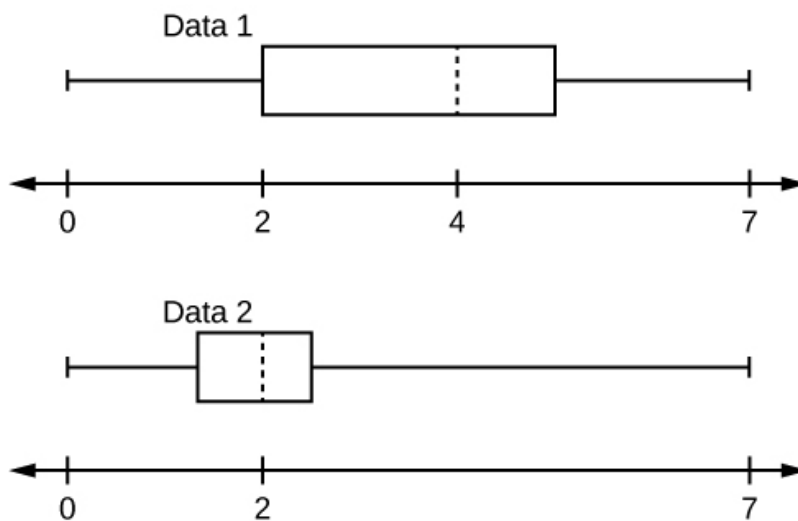


figure 2.74

- a. In complete sentences, explain why each statement is false.
- Data 1** has more data values above two than **Data 2** has above two.
 - The data sets cannot have the same mode.
 - For **Data 1**, there are more data values below four than there are above four.
- b. For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences.
-

9. A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.

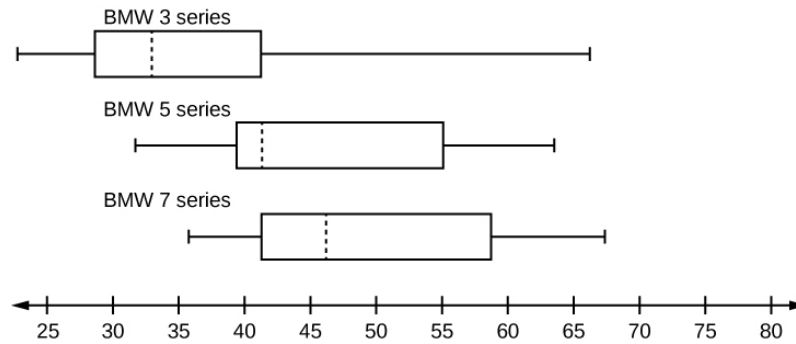


Figure 2.75

- In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
 - Which group is most likely to have an outlier? Explain how you determined that.
 - Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
 - Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
 - Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
 - Look at the BMW 5 series. Estimate the interquartile range (IQR).
 - Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
 - Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
 - 31–35
 - 38–41
 - 41–64
-
- Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.
 - The BMW 3 series is most likely to have an outlier. It has the longest whisker.
 - Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
 - The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
 - The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
 - IQR ~ 17 years
 - There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
 - The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that

fewer than 25% fall between 31 and 35.

10. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Figure 2.76

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Construct a box plot of the data.

11. Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Figure 2.77

Age Group	Percent of Community
0-17	18.9
18-24	8.0
25-34	22.8
35-44	15.0
45-54	13.1
55-64	11.9
65+	10.3

- a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
- b. What percentage of the community is under age 35?
- c. Which box plot most resembles the information above?

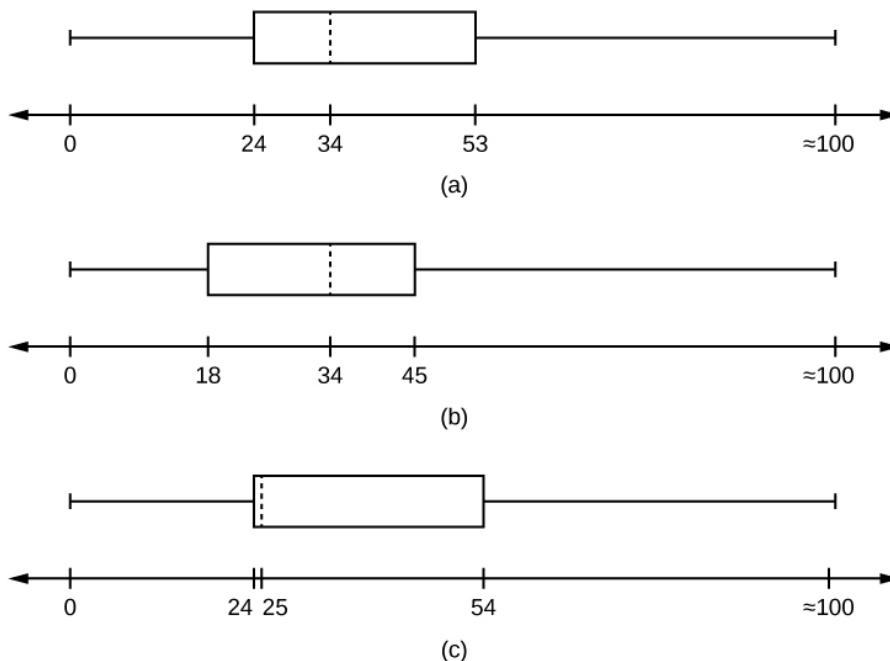


Figure 2.78

- For graph, check student's solution.
- 49.7% of the community is under the age of 35.
- Based on the information in the table, graph (a) most closely represents the data

12. For the following 13 real estate prices, calculate the IQR and determine if any prices are potential outliers. Prices are in dollars.

Data: 389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000.

Solution:

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$LF = Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$UF = Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than $-201,625$. However, $5,500,000$ is more than $1,159,375$. Therefore, $5,500,000$ is a potential outlier.

13. For the following 11 salaries, calculate the IQR and determine if any salaries are outliers. The salaries are in dollars.

\$33,000, \$64,500, \$28,000, \$54,000, \$72,000, \$68,500, \$69,000, \$42,000, \$54,000, \$120,000, \$40,500

14. For the two data sets in Example 1 (test scores), find the following:

- The interquartile range. Compare the two interquartile ranges.
- Any outliers in either set.

Solution:

The five number summary for the day and night classes is

Figure 2.79

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

- The IQR for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The IQR for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class IQR. This suggests more variation will be found in the day class's class test scores.

- Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier

15. Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A:

69, 96, 81, 79, 65, 76, 83, 99, 89, 67, 90, 77, 85, 98, 66, 91, 77, 69, 80, 94

Test Scores for Class B:

90, 72, 80, 92, 90, 97, 92, 75, 79, 68, 70, 80, 99, 95, 78, 73, 71, 68, 95, 100

16. Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

Figure 2.80

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

- a. Find the 28th percentile.
- b. Find the median.
- c. Find the third quartile.

Solution:

a. Notice the 0.28 in the “cumulative relative frequency” column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

b. Look again at the “cumulative relative frequency” column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

c. The third quartile is the same as the 75th percentile. You can “eyeball” this answer. If you look at the “cumulative relative frequency” column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and

sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight.** Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3 , is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

17. Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Figure 2.81

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

18. Using the table below:

Figure 2.82

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

- Find the 80th percentile.
- Find the 90th percentile.
- Find the first quartile. What is another name for the first quartile?

Solution: Using the data from the frequency table, we have:

- The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th percentile = $\frac{8+9}{2} = 8.5$

- b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
- c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six

19. Refer to the table below. Find the third quartile. What is another name for the third quartile?

Figure 2.83

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

20. Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

- a. Find the 40th percentile.
- b. Find the 78th percentile.

Solution:

- a. The 40th percentile is 37 years.
- b. The 78th percentile is 70 years.

21. Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest*.

18, 18, 21, 22, 25, 26, 27, 29, 30, 31, 31, 33, 36, 37, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

- a. Find the percentile of 37.
- b. Find the percentile of 72.

22. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Solution: Jesse graduated 37th out of a class of 180 students. There are $180 - 37 = 143$ students ranked below Jesse. There is one rank of 37.

$$x = 143 \text{ and } y = 1. \frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72. \text{ Jesse's rank of 37 puts him at the } 80^{\text{th}} \text{ percentile.}$$

23. For runners in a race, a *low time* means a faster run. The winners in a race have the shortest running times.

- Is it more desirable to have a finish time with a high or a low percentile when running a race?
 - The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
 - A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.
-

24. For runners in a race, a *higher speed* means a faster run.

- Is it more desirable to have a speed with a high or a low percentile when running a race?
- The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- For runners in a race it is more desirable to have a *high percentile for speed*. A high percentile means a higher speed which is faster.
 - 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).
-

25. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

26. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution: When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

27. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

28. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Solution: The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. Interpretation: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

29. The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
 - b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?
-

30. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is \$240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution: You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more.

31. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, nineteen generally sell four cars, twelve generally sell five cars, nine generally sell six cars, and eleven generally sell seven cars.

- a. First quartile = _____

b. Second quartile = median = 50th percentile = _____

c. Third quartile = _____

d. Interquartile range (IQR) = _____ - _____ = _____

e. 10th percentile = _____

f. 70th percentile = _____

Solution:

b. 4

d. $6 - 4 = 2$

f. 6

32. The median age for U.S. Black citizens currently is 30.9 years; for U.S. White citizens it is 42.3 years.

a. Based upon this information, give two reasons why the Black median age could be lower than the White median age.

b. Does the lower median age for Blacks necessarily mean that Blacks die younger than Whites? Why or why not?

c. How might it be possible for Blacks and Whites to die at approximately the same age, but for the median age for Whites to be higher?

33. Six hundred adult Americans were asked by telephone poll, “What do you think constitutes a middle-class income?” The results are in the figure below. Also, include left endpoint, but not the right endpoint.

Figure 2.84

Salary (\$)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered “not sure”?

b. What percentage think that middle-class is from \$25,000 to \$50,000?

- c. Construct a histogram of the data.
 - i. Should all bars have the same width, based on the data? Why or why not?
 - ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
- d. Find the 40th and 80th percentiles
- e. Construct a bar graph of the data

Solutions:

- a. $1 - (0.02 + 0.09 + 0.19 + 0.26 + 0.18 + 0.17 + 0.02 + 0.01) = 0.06$
- b. $0.19 + 0.26 + 0.18 = 0.63$
- c. Check student's solution.
- d. 40th percentile will fall between 30,000 and 40,000
80th percentile will fall between 50,000 and 75,000
- e. Check student's solution.

34. Given the following box plot:

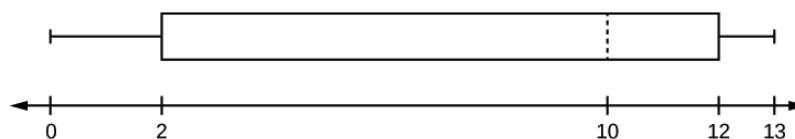


Figure 2.85

- a. which quarter has the smallest spread of data? What is that spread?
- b. which quarter has the largest spread of data? What is that spread?
- c. find the interquartile range (IQR).
- d. are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
- e. which interval has the fewest data in it? How do you know this?
 - i. 0–2
 - ii. 2–4
 - iii. 10–12
 - iv. 12–13
 - v. need more information

35. The following box plot shows the U.S. population for 1990, the latest available year.

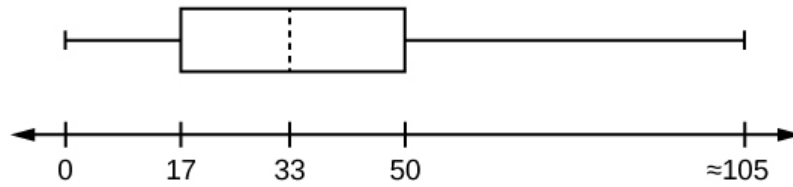


Figure 2.86

- a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
- b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

Solutions:

- a. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
- b. 62.4%

36. On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

37. On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

38. At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

39. During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

40. Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Figure 2.87

Weight Gain	Frequency
-2	3
-1	5
0	2
1	4
4	13
6	2
11	1

a. Calculate the following values:

- the average weight gain for the two weeks
- the standard deviation
- the first, second, and third quartiles

b. Construct a histogram and box plot of the data.

41. The figure below (Table 5) shows the amount, in inches, of annual rainfall in a sample of towns.

Figure 2.88

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95-4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97-6.99	7	$\frac{7}{50} = 0.14$	$0.12 + 0.14 = 0.26$
6.99-9.01	15	$\frac{15}{50} = 0.30$	$0.26 + 0.30 = 0.56$
9.01-11.03	8	$\frac{8}{50} = 0.16$	$0.56 + 0.16 = 0.72$
11.03-13.05	9	$\frac{9}{50} = 0.18$	$0.72 + 0.18 = 0.90$
13.05-15.07	5	$\frac{5}{50} = 0.10$	$0.90 + 0.10 = 1.00$
	Total = 50	Total = 1.00	

a. From the figure above find the percentage of rainfall that is less than 9.01 inches.

- b. Find the percentage of rainfall that is between 6.99 and 13.05 inches.
 - c. Find the number of towns that have rainfall between 2.95 and 9.01 inches.
 - d. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?
-

42. Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2, 5, 7, 3, 2, 10, 18, 15, 20, 7, 10, 18, 5, 12, 13, 12, 4, 5, 10. The following table was produced:

Figure 2.89: Frequency of commuting distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

- a. Is the table correct? If it is not correct, what is wrong?
 - b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
 - c. What fraction of the people surveyed commute five or seven miles?
 - d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution:





An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-74>

43. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in the figure below:

Figure 2.90: Flossing frequency for adults with gum disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

- Fill in the blanks in the figure above
- What percent of adults flossed six times per week?
- What percent flossed at most three times per week?

44. Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2, 5, 7, 2, 2, 10, 20, 15, 0, 7, 0, 20, 5, 12, 15, 12, 4, 5, 10.

Figure 2.91: Frequency of immigrant survey responses

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

- Fix the errors in the figure above. Also, explain how someone might have arrived at the incorrect number(s).
- Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
- Fix the statement in **b** to make it correct.
- What fraction of the people surveyed have lived in the U.S. five or seven years?
- What fraction of the people surveyed have lived in the U.S. at most 12 years?
- What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
- What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

45. The population in Park City is made up of children, working-age adults, and retirees. The figure below shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Figure 2.92

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

46. The data are the distances (in kilometers) from a home to local supermarkets.

1.1, 1.5, 2.3, 2.5, 2.7, 3.2, 3.3, 3.3, 3.5, 3.8, 4.0, 4.2, 4.5, 4.5, 4.7, 4.8, 5.5, 5.6, 6.5, 6.7, 12.3

a. Create a stemplot using the data.

b. Do the data seem to have any concentration of values?

Solution:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-75>

47. The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers: 0.5, 0.7, 1.1, 1.2, 1.2, 1.3, 1.3, 1.5, 1.5, 1.7, 1.7, 1.8, 1.9, 2.0, 2.2, 2.5, 2.6, 2.8, 2.8, 2.8, 3.5, 3.8, 4.4, 4.8, 4.9, 5.2, 5.5, 5.7, 5.8, 8.0

48. For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest): 32, 32, 33, 34, 38, 40, 42, 42, 43, 44, 46, 47, 47, 48, 48, 48, 49, 50, 50, 51, 52, 52, 52, 53, 54, 56, 57, 57, 60, 61. Construct a stem plot for the data.

49. The table below shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Figure 2.93: Atlanta hawks wins and losses

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

50. In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in the table below. Construct a line graph.

Figure 2.94

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

51. Using this data set, construct a histogram.

Figure 2.95: Number of hours my classmates spent playing video games on weekends

9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

52. The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22, 35, 15, 26, 40, 28, 18, 20, 25, 34, 39, 42, 24, 22, 19, 27, 22, 34, 40, 20, 38, and 28.

Use 10–19 as the first interval.

53. Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than \$19 each.

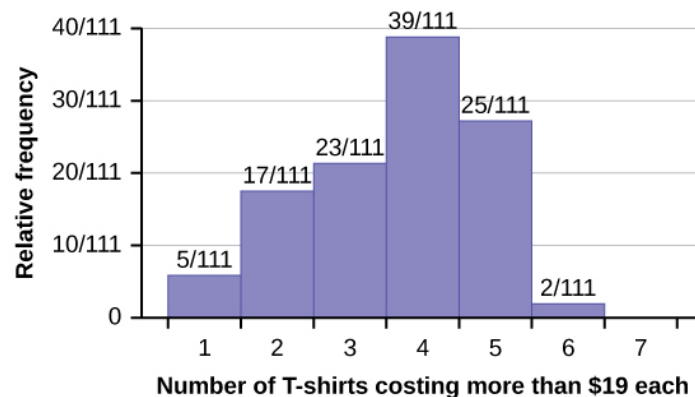


Figure 2.96

a. The percentage of people who own at most three t-shirts costing more than \$19 each is approximately:

- a. 21
- b. 59
- c. 41
- d. Cannot be determined

- b. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:
- cluster
 - simple random
 - stratified
 - convenience

54. Following are the 2010 obesity rates by U.S. states and Washington, DC.

Figure 2.97

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states. Answers will vary.

9. "Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

55. Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99.

- Construct a stem-and-leaf plot of the data.
- Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

56. The table below contains the 2010 obesity rates in U.S. states and Washington, DC.

10

Figure 2.98

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

- Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
- Construct a bar graph for all the states beginning with the letter “A.”
- Construct a bar graph for all the states beginning with the letter “M.”

10. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

Solution:

- a. Eight numbers are generated. The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}). If any numbers are repeated, generate a different number. Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}. Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

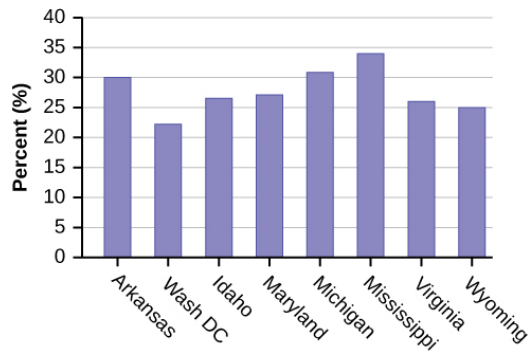


Figure 2.99

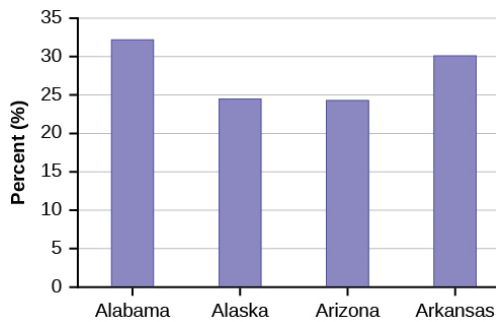


Figure 2.100

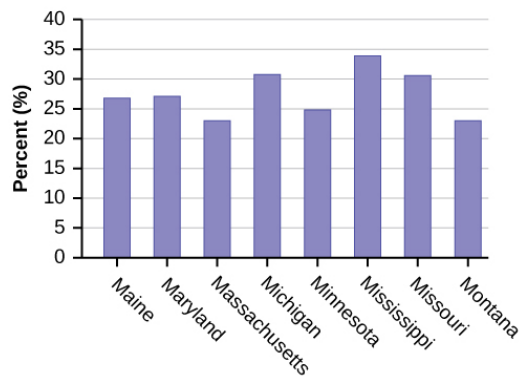


Figure 2.101

57. For each of the following data sets, create a stem plot and identify any outliers. The miles per gallon rating for 30 cars are shown below (lowest to highest).

19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Figure 2.102

Stem	Leaf
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

a. The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

b. The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Figure 2.103

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

c. The data are daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

58. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The figure below shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Figure 2.104

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Using the data from Mrs. Ramirez’s math class, construct a bar graph showing the percentages.

59. David County has six high schools. Each school sent students to participate in a county-wide science competition. The figure below shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

Figure 2.105

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Use the data from the David County science competition supplied above. Construct a bar graph that shows the county-wide population percentage of students at each school.

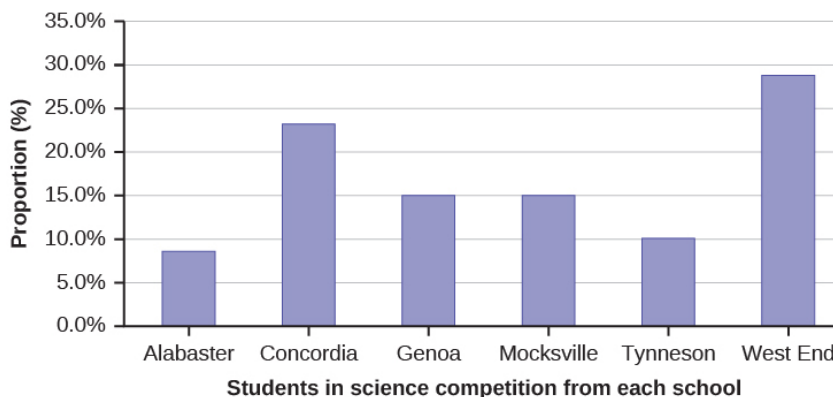


Figure 2.106

2.6 Measures of Center

1. The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 19, 19, 21, 21, 22, 22, 23, 24, 24, 24, 24

2. In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the “center”: the mean or the median?

3. The number of books checked out from the library from 25 students are as follows: 0, 0, 0, 1, 2, 3, 3, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 8, 9, 10, 10, 11, 11, 12, 12. Find the mode.

4. Find the mean for the following frequency tables.

a. **Figure 2.107**

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b. **Figure 2.108**

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c. **Figure 2.109**

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

5. The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16, 17, 19, 20, 20, 21, 23, 24, 25, 25, 25, 26, 26, 27, 27, 27, 28, 29, 30, 32, 33, 33, 34, 35, 37, 39, 40

a. Calculate the mean.

- Mean: $16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738$; $\frac{738}{27} = 27.33$

b. Identify the median.

c. Identify the mode.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-76>

6. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, nineteen generally sell four cars, twelve generally sell five cars, nine generally sell six cars, and eleven generally sell seven cars. Calculate the following:

1. sample mean = \bar{x} = _____

2. median = _____

3. mode = _____

7. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.¹¹

Figure 2.110

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

- What is the best estimate of the average obesity percentage for these countries?
 - The United States has an average obesity rate of 33.9%. Is this rate above average or below?
 - How does the United States compare to other countries?
-

8. The following figure gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?¹²

Figure 2.111

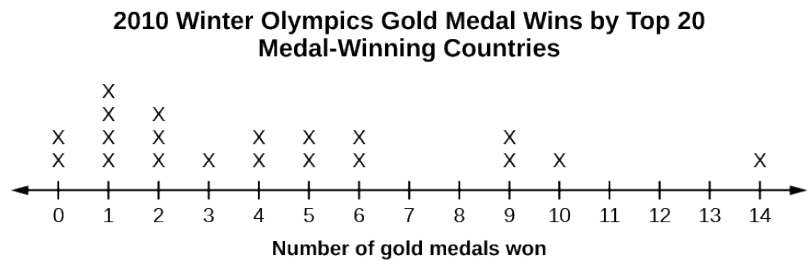
Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

The mean percentage, $\bar{x} = \frac{1328.65}{50} = 26.75$

- “Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).
- “Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).

9. Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.



b.

Figure 2.113: The Ages Former U.S Presidents Died

4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8 0 means 80.	

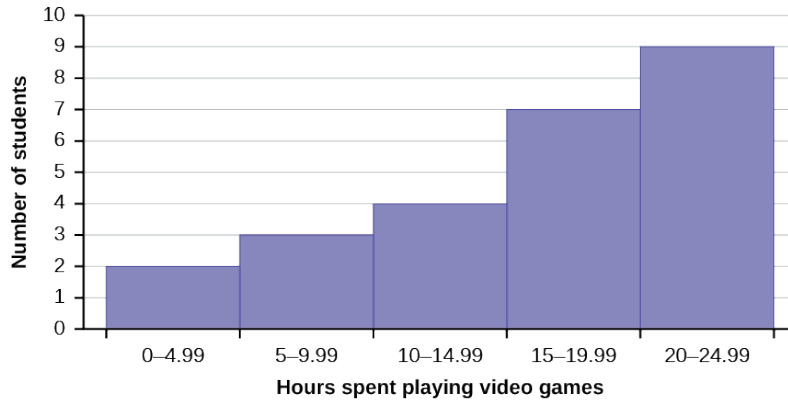
13

c.

13. “Presidents.” Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).

Hours Spent Playing Video Games on Weekends

Figure 2.114



10. State whether the data are symmetrical, skewed to the left, or skewed to the right.

a. 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-77>

b. 16, 17, 19, 22, 22, 22, 22, 22, 23

c. 87, 87, 87, 87, 87, 88, 89, 89, 90, 91



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-78>

11. When the data are skewed left, what is the typical relationship between the mean and median?

12. When the data are symmetrical, what is the typical relationship between the mean and median?

Solution: When the data are symmetrical, the mean and median are close or the same.

13. What word describes a distribution that has two modes?

14. Use the following graph to answer a-c.

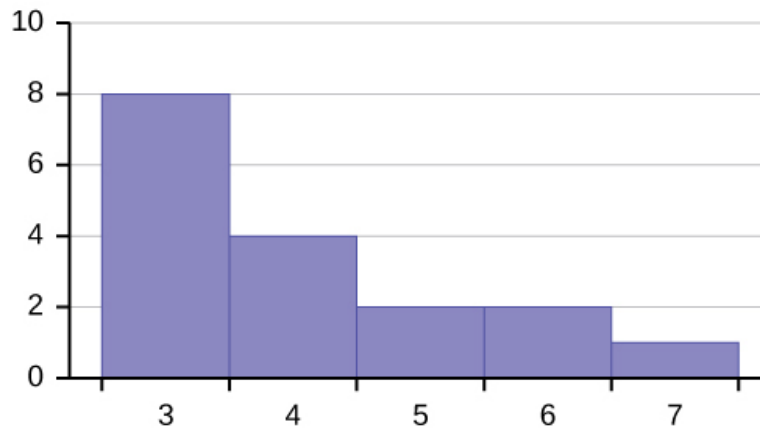


Figure 2.115

a. Describe the shape of this distribution.

- Solution: The distribution is skewed right because it looks pulled out to the right.

b. Describe the relationship between the mode and the median of this distribution.

c. Describe the relationship between the mean and the median of this distribution.

- Solution: The mean is 4.1 and is slightly greater than the median, which is four.
-

15. Data: 11, 11, 12, 12, 12, 12, 13, 15, 17, 22, 22, 22

a. Is the data perfectly symmetrical? Why or why not?

b. Which is the largest, the mean, the mode, or the median of the data set?

- Solution: The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.
-

16. Data: 56, 56, 56, 58, 59, 60, 62, 64, 64, 65, 67

- a. Is the data perfectly symmetrical? Why or why not?
 - b. Which is the largest, the mean, the mode, or the median of the data set?
-

17. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

- Solution: The mean tends to reflect skewing the most because it is affected the most by outliers.
-

18. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

19. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

- a. What does it mean for the median age to rise?
 - b. Give two reasons why the median age could rise.
 - c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?
-

20. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

Figure 2.116

	Javier	Ercilia
\bar{x}	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

- a. How can you determine which survey was correct ?
- b. Explain what the difference in the results of the surveys implies about the data.
- c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

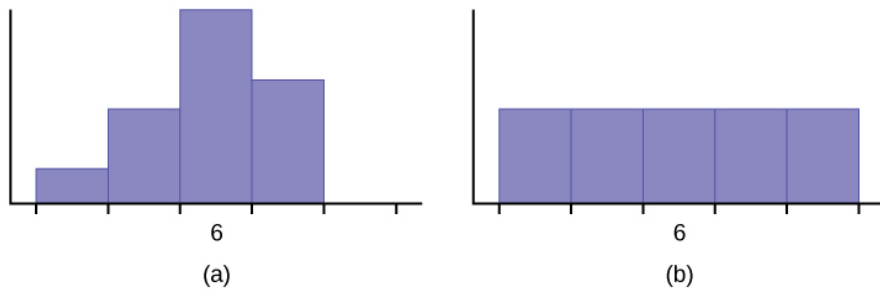


Figure 2.117

- d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

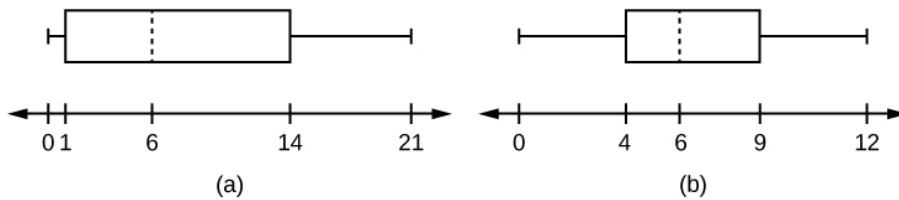


Figure 2.118

21. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Figure 2.119

Number of years	Frequency
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
Total = 20	



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-79>

What is the mode?

- a. 19
- b. 19.5
- c. 14 and 20
- d. 22.65

Is this a sample or the entire population?

- a. sample
- b. entire population
- c. neither

22. How much time does it take to travel to work? The figure below shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

Figure 2.120

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

23. Find the midpoint for each class. These will be graphed on the x -axis. The frequency values will be graphed on the y -axis values.

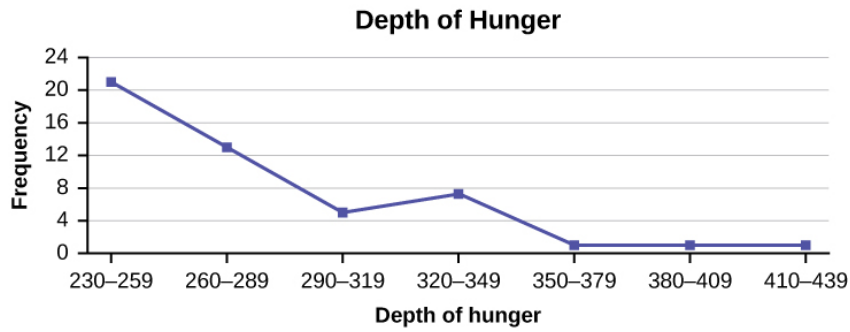


Figure 2.121

2.7 Measures of Spread

1. Use the following data (first exam scores) from Susan Dean's spring pre-calculus class: 33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100.

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.

b. Calculate the following to one decimal place:

- i. The sample mean
- ii. The sample standard deviation
- iii. The median
- iv. The first quartile
- v. The third quartile
- vi. IQR

c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solutions:

a.

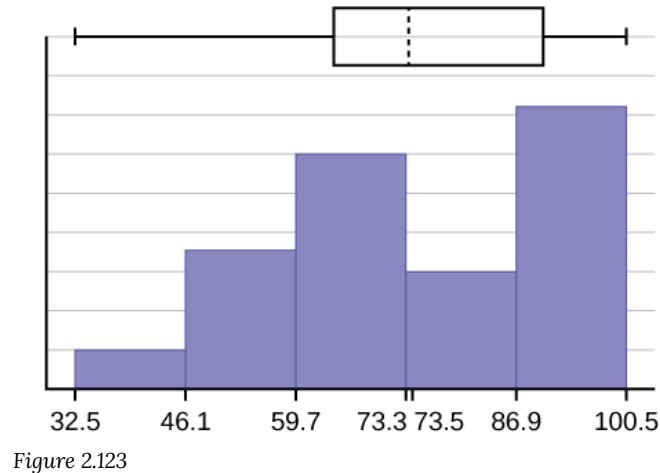
Figure 2.122

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

b.

- i. The sample mean = 73.5
- ii. The sample standard deviation = 17.9
- iii. The median = 73
- iv. The first quartile = 61
- v. The third quartile = 90
- vi. $IQR = 90 - 61 = 29$

c. The x -axis goes from 32.5 to 100.5; y -axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5$ = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ($IQR = 29$) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

2. The following data show the different types of pet food stores in the area carry: 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12. Calculate the sample mean and the sample standard deviation to one decimal place.

3. The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles: 29, 37, 38, 40, 58, 67, 68, 69, 76, 86, 87, 95, 96, 96, 99, 106, 112, 127, 145, 150.

a. Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

- *Solution:* $s = 34.5$

b. Find the value that is one standard deviation below the mean.

4. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Figure 2.124

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$

For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$

Fredo's z-score of -0.67 is higher than Karl's z-score of -0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Use the table above to find the value that is three standard deviations:

- above the mean
- below the mean

5. Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Figure 2.125

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Figure 2.126

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Figure 2.127

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solutions:

$$\begin{aligned} \text{a. } s_x &= \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88 \\ \text{b. } s_x &= \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62 \\ \text{c. } s_x &= \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14 \end{aligned}$$

6. The population parameters below describe the full-time equivalent number of students (FTES) each year at ABC University from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median = 1,014 FTES
- $\sigma = 474$ FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- $n = 29$ years

a. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

- The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

b. 75% of all years have an FTES:

- at or below: _____
- at or above: _____

c. The population standard deviation = _____

- 474 FTES

d. What percent of the FTES were from 528.5 to 1447.5? How do you know?

e. What is the IQR? What does the IQR represent?

- 919

f. How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Figure 2.128

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

g. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the IQR. Round to one decimal place.

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- IQR = 245

h. What additional information is needed to construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005?

i. Compare the IQR for the FTES for 1976–77 through 2004–2005 with the IQR for the FTES for 2005–2006 through 2010–2011. Why do you suppose the IQRs are so different? Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

7. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Figure 2.129

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

8. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing \$3,000, a guitar costing \$550, and a drum set costing \$600. The mean cost for a piano is \$4,000 with a standard deviation of \$2,500. The mean cost for a guitar is \$500 with a standard deviation of \$200. The mean cost for drums is \$700 with a standard deviation of \$100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

- Solution: For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.
-

9. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

- a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
 - b. Who is the fastest runner with respect to his or her class? Explain why.
-

10. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the table below:¹⁴

14. "Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).

Figure 2.130

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

Solutions:

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that $23.32 + 12.95 = 36.27$ is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

11. The figure below gives the percent of children under five considered to be underweight.¹⁵

15. “Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).

Figure 2.131

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

12. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Figure 2.132

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

- Find the sample mean \bar{x} .
- Find the approximate sample standard deviation, s .

Solutions:

- 1.48
 - 1.12
-

13. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

Figure 2.133

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

- Find the sample mean \bar{x}
 - Find the sample standard deviation, s
 - Construct a histogram of the data.
 - Complete the columns of the chart.
 - Find the first quartile.
 - Find the median.
 - Find the third quartile.
 - Construct a box plot of the data.
 - What percent of the students owned at least five pairs?
 - Find the 40th percentile.
 - Find the 90th percentile.
 - Construct a line graph of the data
 - Construct a stemplot of the data
-

14. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177, 205, 210, 210, 232, 205, 185, 185, 178, 210, 206, 212, 184, 174, 185, 242, 188, 212, 215, 247, 241, 223, 220, 260, 245, 259, 278, 270, 280, 295, 275, 285, 290, 272, 273, 280, 285, 286, 200, 215, 185, 230, 250, 241, 190, 260, 250, 302, 265, 290, 276, 228, 265

- Organize the data from smallest to largest value.
- Find the median.
- Find the first quartile.
- Find the third quartile.
- Construct a box plot of the data.
- The middle 50% of the weights are from _____ to _____.
- If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- If our population included every team member who ever played for the San Francisco 49ers, would the

above data be a sample of weights or the population of weights? Why?

- i. Assume the population was the San Francisco 49ers. Find:
 - i. the population mean, μ .
 - ii. the population standard deviation, σ .
 - iii. the weight that is two standard deviations below the mean.
 - iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
- j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmitt Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solutions:

- a. 174, 177, 178, 184, 185, 185, 185, 185, 188, 190, 200, 205, 205, 206, 210, 210, 210, 212, 212, 215, 215, 220, 223, 228, 230, 232, 241, 241, 242, 245, 247, 250, 250, 259, 260, 260, 265, 265, 270, 272, 273, 275, 276, 278, 280, 280, 285, 285, 286, 290, 290, 295, 302
- b. 241
- c. 205.5
- d. 272.5



Figure 2.134

- e.
- f. 205.5, 272.5
- g. sample
- h. population
 - i. 236.34
 - ii. 37.50
 - iii. 161.34
 - iv. 0.84 std. dev. below the mean
- i. Young

15. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

- What is the mean change score?
- What is the standard deviation for this population?
- What is the median change score?
- Find the change score that is 2.2 standard deviations below the mean.

16. Refer to the figures below and determine which of the following (a-d) are true and which are false. Explain your solution to each part in complete sentences.

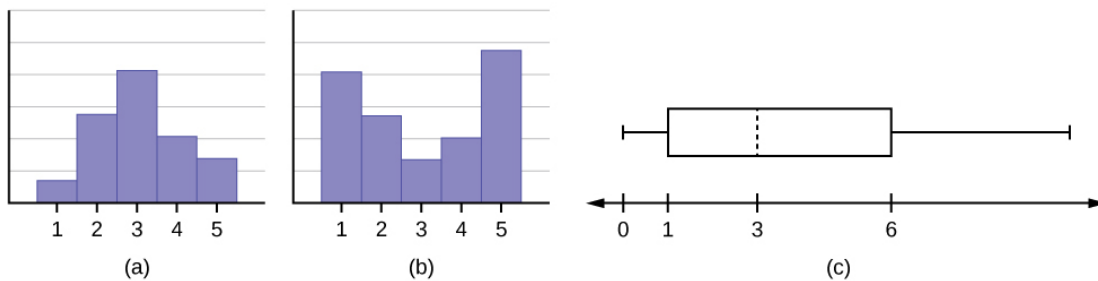


Figure 2.135

- The medians for all three graphs are the same.
- We cannot determine if any of the means for the three graphs is different.
- The standard deviation for graph b is larger than the standard deviation for graph a.
- We cannot determine if any of the third quartiles for the three graphs is different.

Solutions:

- True
- True
- True
- False

17. In a recent issue of the IEEE SPECTRUM, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- Organize the data in a chart.

- b. Find the median, the first quartile, and the third quartile.
 - c. Find the 65th percentile.
 - d. Find the 10th percentile.
 - e. Construct a box plot of the data.
 - f. The middle 50% of the conferences last from _____ days to _____ days.
 - g. Calculate the sample mean of days of engineering conferences.
 - h. Calculate the sample standard deviation of days of engineering conferences.
 - i. Find the mode.
 - j. If you were planning an engineering conference, which would you choose as the length of the conference: mean, median, or mode? Explain why you made that choice.
 - k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.
-

18. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414, 1550, 2109, 9350, 21828, 4300, 5944, 5722, 2825, 2044, 5481, 5200, 5853, 2750, 10012, 6357, 27000, 9414, 7681, 3200, 17500, 9200, 7380, 18314, 6557, 13713, 17768, 7493, 2771, 2861, 1263, 7285, 28165, 5080, 11622

- a. Organize the data into a chart with five intervals of equal width. Label the two columns “Enrollment” and “Frequency.”
- b. Construct a histogram of the data.
- c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- d. Calculate the sample mean.
- e. Calculate the sample standard deviation.
- f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Solutions:

- a. **Figure 2.136**

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

- b. Check student’s solution.
- c. mode
- d. 8628.74

- e. 6943.88
 - f. -0.09
-

19. X = the number of days per week that 100 clients use a particular exercise facility.

Figure 2.137

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

a. The 80th percentile is _____

- a. 5
- b. 80
- c. 3
- d. 4

Solution: a

b. The number that is 1.5 standard deviations BELOW the mean is approximately _____

- a. 0.7
 - b. 4.8
 - c. -2.8
 - d. Cannot be determined
-

20. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the figure below.

Figure 2.138

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

- Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
- If a data value is identified as an outlier, what should be done about it?
- Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
- Do parts a and c of this problem give the same answer?
- Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
- Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode

21. This figure contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Figure 2.139

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Answer each of the following questions and check your answers below.

- What is the frequency of deaths measured from 2006 through 2009?
- What percentage of deaths occurred after 2009?
- What is the relative frequency of deaths that occurred in 2003 or earlier?
- What is the percentage of deaths that occurred in 2004?
- What kind of data are the numbers of deaths?
- The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution:



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=119#h5p-80>

22. The following figure contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Figure 2.140

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

- What is the frequency of deaths measured from 2000 through 2004?
 - What percentage of deaths occurred after 2006?
 - What is the relative frequency of deaths that occurred in 2000 or before?
 - What is the percentage of deaths that occurred in 2011?
 - What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.
-

23. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Figure 2.141: Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Fill in the blanks in the figure above.

- What percent of students take exactly two courses?
 - What percent of students take one or two courses?
-

24. *Forbes* magazine published data on the best small firms in 2012. These were firms which had been publicly

traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The figure below shows the ages of the chief executive officers for the first 60 ranked firms.

Figure 2.142

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40-44	3		
45-49	11		
50-54	13		
55-59	16		
60-64	10		
65-69	6		
70-74	1		

- What is the frequency for CEO ages between 54 and 65?
- What percentage of CEOs are 65 years or older?
- What is the relative frequency of ages under 50?
- What is the cumulative relative frequency for CEOs younger than 55?
- Which graph shows the relative frequency and which shows the cumulative relative frequency?

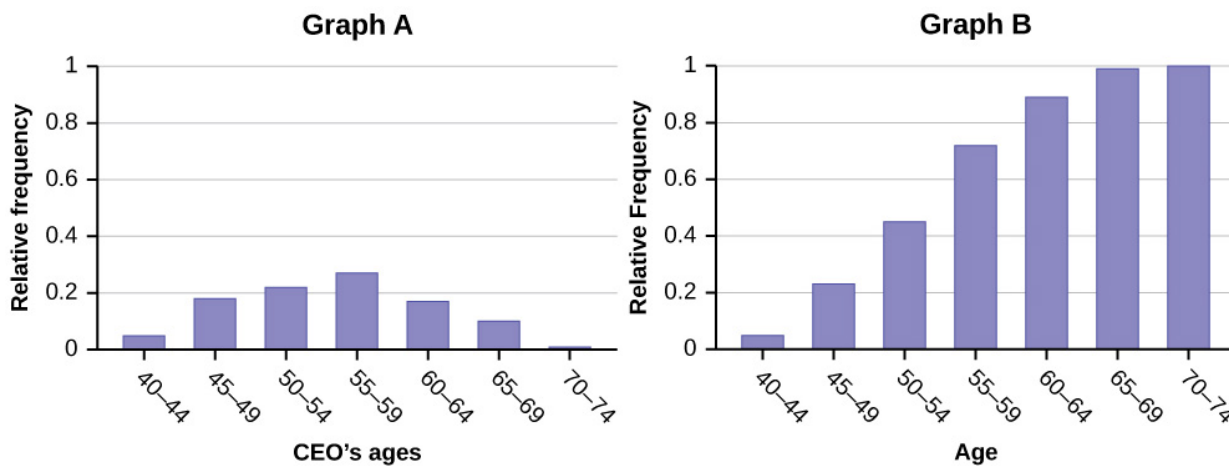


Figure 2.143

25. The figure below contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Figure 2.144: Frequency of Hurricane Direct Hits

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

- a. What is the relative frequency of direct hits that were category 4 hurricanes?
- 0.0768
 - 0.0659
 - 0.2601
 - Not enough information to calculate
- b. What is the relative frequency of direct hits that were AT MOST a category 3 storm?
- 0.3480
 - 0.9231
 - 0.2601
 - 0.3370

26. The following data are the shoe sizes of 50 male students. The sizes are discrete data since shoe size is measured in whole and half units only. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9, 9, 9.5, 9.5, 10, 10, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5

11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5, 11.5, 11.5, 11.5

12, 12, 12, 12, 12, 12, 12, 12.5, 12.5, 12.5, 12.5, 14

27. The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2

3, 3, 3, 3, 3, 3, 3

20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point

is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

28. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, nineteen generally sell four cars, twelve generally sell five cars, nine generally sell six cars, and eleven generally sell seven cars. Complete the table.

Figure 2.145

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

What does the frequency column sum to? Why?

What does the relative frequency column sum to? Why?

What is the difference between relative frequency and frequency for each data value?

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

What is the difference between cumulative relative frequency and relative frequency for each data value?

To construct the histogram for the data, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

29. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Figure 2.146: Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Figure 2.147: Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Figure 2.148: Publisher C

# of books	Freq.	Rel. Freq.
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

- Find the relative frequencies for each survey. Write them in the charts.
- Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
- In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
- Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
- Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
- Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more

similar or more different? Explain your answer.

30. Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Figure 2.149: Singles

Amount(\$)	Frequency	Rel. Frequency
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

Figure 2.150: Couples

Amount(\$)	Frequency	Rel. Frequency
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

- Fill in the relative frequency for each group.
- Construct a histogram for the singles group. Scale the x -axis by \$50 widths. Use relative frequency on the y -axis.
- Construct a histogram for the couples group. Scale the x -axis by \$50 widths. Use relative frequency on the y -axis.
- Compare the two graphs:
 - List two similarities between the graphs.
 - List two differences between the graphs.

- iii. Overall, are the graphs more similar or different?
- e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by \$50, scale it by \$100. Use relative frequency on the y-axis.
- f. Compare the graph for the singles with the new graph for the couples:
 - i. List two similarities between the graphs.
 - ii. Overall, are the graphs more similar or different?
- g. How did scaling the couples graph differently change the way you compared it to the singles graph?
- h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

Figure 2.151: Singles

Amount(\$)	Frequency	Relative Frequency
51-100	5	0.08
101-150	10	0.17
151-200	15	0.25
201-250	15	0.25
251-300	10	0.17
301-350	5	0.08

Figure 2.152: Couples

Amount(\$)	Frequency	Relative Frequency
100-150	5	0.07
201-250	5	0.07
251-300	5	0.07
301-350	5	0.07
351-400	10	0.14
401-450	10	0.14
451-500	10	0.14
501-550	10	0.14
551-600	5	0.07
601-650	5	0.07

- a. See the figures above.
- b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

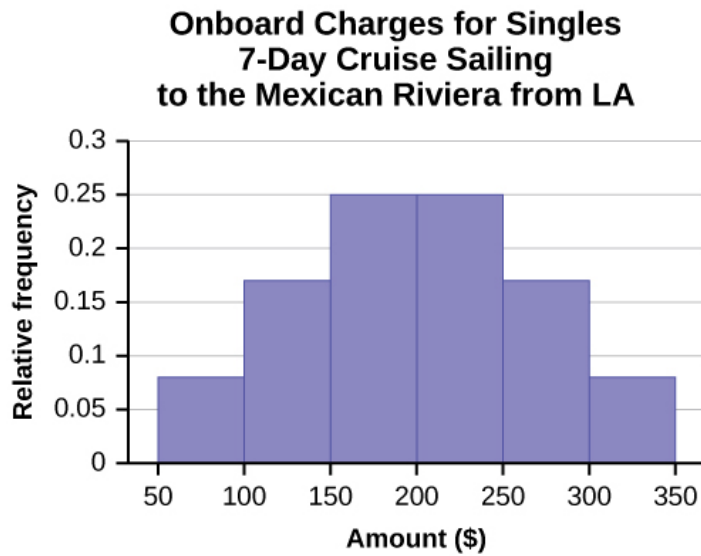


Figure 2.153

- c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

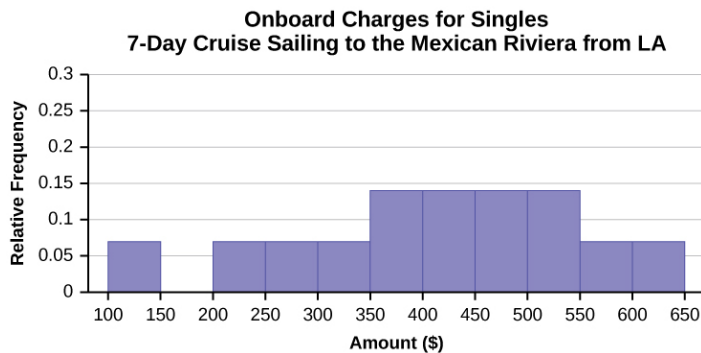


Figure 2.154

- d. Compare the two graphs:
- i. Answers may vary. Possible answers include:
 - Both graphs have a single peak.
 - Both graphs use class intervals with width equal to \$50.
 - ii. Answers may vary. Possible answers include:
 - The couples graph has a class interval with no values.
 - It takes almost twice as many class intervals to display the data for couples.

- iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.
- e. Check student's solution.
- f. Compare the graph for the Singles with the new graph for the Couples:
 - Both graphs have a single peak.
 - Both graphs display 6 class intervals.
 - Both graphs show the same general pattern.
- ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.
- g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
- h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

31. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

Figure 2.155

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

- a. Construct a histogram of the data.
- b. Complete the columns of the chart.

-
32. Use the data to construct a line graph.
- a. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown below.

Figure 2.156

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Solution:



b. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown below.

Figure 2.158

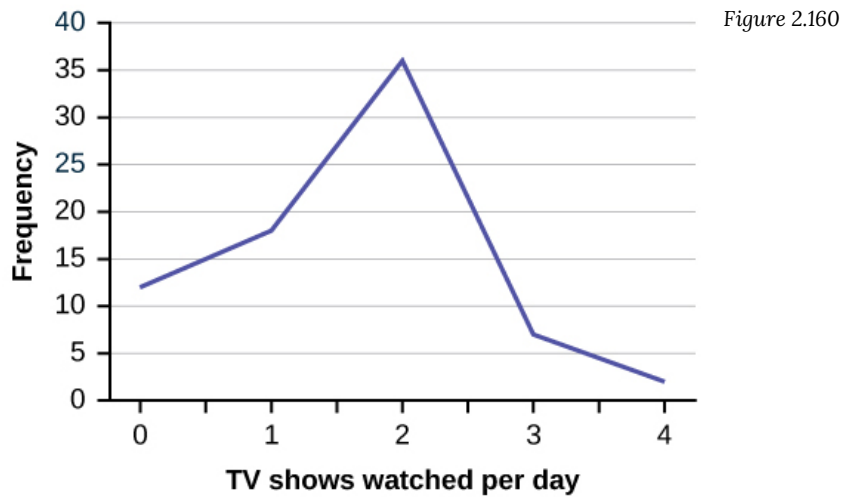
Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

c. Several children were asked how many TV shows they watch each day. The results of the survey are shown below.

Figure 2.159

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Solution:



References

Image References

Figure 2.55: Figure 2.6 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-2-histograms-frequency-polygons-and-time-series-graphs>

Figure 2.58: Figure 2.9 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-2-histograms-frequency-polygons-and-time-series-graphs>

Figure 2.68: Figure 2.8 from OpenIntro Introductory Statistics (2019) (CC BY-SA 3.0). Retrieved from <https://cnx.org/contents/pJuo4h-U@4.478:UMM7d-Hy/Display-Data>

Figure 2.70: Figure 2.14 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-4-box-plots>

Figure 2.71: Figure 2.17 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-4-box-plots>

Figure 2.72: Figure 2.45 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-homework>

Figure 2.73: Figure 2.46 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-homework>

Figure 2.74: Figure 2.47 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-homework>

Figure 2.75: Figure 2.46 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-homework>

Figure 2.78: Figure 2.47 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-bringing-it-together-homework>

Figure 2.85: Figure 2.43 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-homework>

Figure 2.86: Figure 2.44 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-homework>

Figure 2.96: Figure from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/2-homework>

Figure 2.99: Figure 2.58 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions>

Figure 2.100: Figure 2.59 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions>

Figure 2.101: Figure 2.60 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions>

Figure 2.106: Figure 2.54 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions>

Figure 2.112: Figure 2.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode>

Figure 2.114: Figure 2.25 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-6-skewness-and-the-mean-median-and-mode>

Figure 2.115: Figure 2.7.9 from LibreTexts Introductory Statistics (2020) (CC BY 4.0). Retrieved from [https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_\(OpenStax\)/02%3A_Descriptive_Statistics/2.07%3A_Skewness_and_the_Mean_Median_and_Mode](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(OpenStax)/02%3A_Descriptive_Statistics/2.07%3A_Skewness_and_the_Mean_Median_and_Mode)

Figure 2.117: Figure 2.9.1 from LibreTexts Introductory Business Statistics (2020) (CC BY 4.0). Retrieved from https://biz.libretexts.org/Courses/Gettysburg_College/MGT_235%3A_Introductory_Business_Statistics/02%3A_Descriptive_Statistics/2.09%3A_Homework

Figure 2.118: Figure 2.51 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-bringing-it-together-homework>

Figure 2.121: Figure 2.58 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-solutions>

Figure 2.123: Figure 2.8.2 from LibreTexts Introductory Statistics (2020) (CC BY 4.0). Retrieved from [https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_\(OpenStax\)/02%3A_Descriptive_Statistics/2.08%3A_Measures_of_the_Spread_of_the_Data](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(OpenStax)/02%3A_Descriptive_Statistics/2.08%3A_Measures_of_the_Spread_of_the_Data)

Figure 2.134: Figure from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-solutions#element-324s-solution>

Figure 2.135: Figure 2.52 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/2-bringing-it-together-homework>

Figure 2.143: Figure 1.11 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/1-homework>

Figure 2.153: Figure 2.36 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/2-solutions#eip-457-solution>

Figure 2.154: Figure 2.37 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/2-solutions#eip-457-solution>

Figure 2.157: Figure 2.51 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions#fs-idp113295424-solution>

Figure 2.160: Figure 2.52 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/2-solutions#fs-idp113295424-solution>

Text

“State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. <http://quickfacts.census.gov/qfd/index.html> (accessed May 1, 2013).

“Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, <http://www.nhc.noaa.gov/gifs/table5.gif> (accessed May 1, 2013).

“Levels of Measurement,” http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

David Lane. “Levels of Measurement,” Connexions, <http://cnx.org/content/m10809/latest> (accessed May 1, 2013).

Dekker, Marcel. Data on annual homicides in Detroit, 1961–73 in Gunst & Mason, *Regression Analysis and its Application*.

“Timeline: Guide to the U.S. Presidents: Information on every president’s birthplace, political party, term of office, and more.” Scholastic, 2013. Available online at <http://www.scholastic.com/teachers/article/timeline-guide-us-presidents> (accessed April 3, 2013).

“Presidents.” Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).

“Food Security Statistics.” Food and Agriculture Organization of the United Nations. Available online at <http://www.fao.org/economic/ess/ess-fs/en/> (accessed April 3, 2013).

“Consumer Price Index.” United States Department of Labor: Bureau of Labor Statistics. Available online at <http://data.bls.gov/pdq/SurveyOutputServlet> (accessed April 3, 2013).

“CO2 emissions (kt).” The World Bank, 2013. Available online at <http://databank.worldbank.org/data/home.aspx> (accessed April 3, 2013).

“Births Time Series Data.” General Register Office For Scotland, 2013. Available online at <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html> (accessed April 3, 2013).

“Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

“Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

Burbary, Ken. *Facebook Demographics Revisited – 2011 Statistics*, 2011. Available online at <http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/> (accessed August 21, 2013).

“9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at <http://apreport.collegeboard.org/goals-and-findings/promoting-equity> (accessed September 13, 2013).

Data from *West Magazine*.

Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” *USA Today*, 2012. Available online at <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1> (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/> (accessed April 3, 2013).

“1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/main/www/cen1990.html> (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Data from The World Bank, available online at <http://www.worldbank.org> (accessed April 3, 2013).

“Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).

Data from Microsoft Bookshelf.

King, Bill. “Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

CHAPTER 3: BASICS OF PROBABILITY

3.1 Introduction to Probability and Terminology

Learning Objectives

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability
- Determine whether two events are mutually exclusive
- Determine whether two events are independent
- Construct and interpret contingency tables
- Construct and interpret Venn diagrams
- Construct and interpret tree diagrams
- Calculate probabilities using the addition rules
- Calculate probabilities using the multiplication rules



Figure 3.1: Meteor Shower. Meteor showers are rare, but the probability of them occurring can be calculated.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Probability

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An experiment is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **probability experiment**. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

A **probability model** is a mathematical representation of a random process that lists all possible outcomes and assigns probabilities to each of them. This type of model is our ultimately our goal when moving forward in our study of statistics.

The Law of Large Numbers

An important characteristic of probability experiments, known as the **law of large numbers**, states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word empirical is often used instead of the word observed.)

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment, rather it gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Classical Approach to Probability

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a fair, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. This is often called the Classical Approach to probability.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event E = rolling a number that is at least five.

There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of “at least five”. You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

Example

If you toss a fair dime and a fair nickel, the sample space is {HH, TH, HT, TT} where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition {HT, TH}, so $P(A) = \frac{2}{4} = 0.5$.

The Axioms of Probability

Finding probabilities in more complicated situations starts with the three Axioms of Probability.

1. $P(S) = 1$
2. $0 \leq P(E) \leq 1$
3. For each two events E1 and E2 with $E1 \cap E2 = \emptyset$, $P(E1 \cup E2) = P(E1) + P(E2)$

The first two Axioms should be fairly intuitive. Axiom 1 says that the probabilities of all outcomes in a sample space will always add up to 1. Axiom 2 says the probability of any event must be between 0 and 1. The Third Axiom is called the disjoint addition rule which we will expand on in the future.

Relationships Between Events

Often we are not just interested in a single event, but multiple events happening at the same time. In order to find probabilities relating to multiple events, we first have to know about the relationship (or lack thereof) between them. The two main relationship terms we will look for are independence and mutually exclusive. Remember, these two terms certainly do not mean the same thing, neither are they opposites.

Consider two events, A&B. If it is not known whether they are mutually exclusive, assume they are not until you can show otherwise. Likewise, if it is not known whether A and B are independent, assume they are dependent until you can show otherwise. This “default” starting point is illustrated by the 4th position in the following table:

Figure 3.2: Relationships Between Events

		Independent?	
		Yes	No
Mutually Exclusive?	Yes	1*	2
	No	3	4

Depending on the information given in the problem and Assumptions you are able to make, as you move around the above grid, you will apply slightly different versions of each important probability rule.

***Note:** You will rarely, if ever, find yourself in this case

Mutually exclusive

A and B are **mutually exclusive** (or disjoint) events if they cannot occur at the same time. This means that A and B do not share any outcomes and $P(A \text{ AND } B) = 0$.

Example

Suppose the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $A = \{1, 2, 3, 4, 5\}$, $B = \{4, 5, 6, 7, 8\}$, and $C = \{7, 9\}$. $A \text{ AND } B = \{4, 5\}$. $P(A \text{ AND } B) = \frac{2}{10}$ and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(A \text{ AND } C) = 0$. Therefore, A and C are mutually exclusive.

Independent events

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. If two events are NOT independent, then we say that they are dependent. To show two events are independent, you only need to show one one of the equivalent conditions below:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

How you sample can have implications on independence. Sampling may be done with or without replacement

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

Example

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Consider the following scenarios

1. Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.
2. Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the

remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=123#h5p-81>

Compound Events

Often we are not just interested in a single event, but multiple events happening at the same time. There are several types of relationships and situations we may be interested in. The relationship between events can tell us a lot about the probability of compound events. Depending on the compound event we are looking for we will apply different rules.

Compliments

The **complement** of event A is denoted A' (read “A prime”). A' consists of all outcomes in the sample space, S, that are **NOT** in A.

There are several useful forms of the compliment rule:

- $P(A) + P(A') = 1$
- $1 - P(A) = P(A')$
- $1 - P(A') = P(A)$

Example

Let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = \{1, 2, 3, 4\}$. Then, $A' = \{5, 6\}$. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') =$

$$\frac{4}{6} + \frac{2}{6} = 1$$

Unions

The **union** of two events, denoted $A \cup B$, is the outcomes that are in either event A **OR** B (or both).

Example

Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. $A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

Intersections

The **intersection** of two events, denoted $A \cap B$, is the outcomes that are in both events A **AND** B.

Example

Let A and B be $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7, 8\}$, respectively. Then $A \text{ AND } B = \{4, 5\}$.

Conditional Probabilities

Sometimes knowing one event has already happened can change the probability of another event occurring. A **conditional probability** reduces the sample space by updating our probabilities based on what we already know. The conditional probability of A **GIVEN** B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. . We calculate the probability of A from the reduced sample space B. The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ where $P(B)$ is greater than zero.

Example

Suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let A = face is 2 or 3 and B = face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{the number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Your turn!

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

Are C and D independent?

Are C and D mutually exclusive?

What is $P(D|C)$?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=123#h5p-82>

Solving Probability Problems

The key to probability problems is sorting through and understanding important terminology and symbols. First read each problem carefully to think about and understand what you are looking for. Look for key words (and, or, not, w/ or w/o replacement etc..) to identify the event(s) of interest and the relationships between them. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional. Visualize them if possible using the tools we'll talk about in the future. We can then plug into our rules to get a numerical answer. Finally, based on your understanding of the situation, make sure the probability you came up with makes intuitive sense.

Example

Consider flipping two fair coins.

The sample space is {HH, HT, TH, TT} where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. Notice the outcomes HT and TH are different. HT means that the first coin showed heads and the second coin showed tails. TH means that the first coin showed tails and the second coin showed heads.

Find the probabilities of the events.

- a. Let A = the event of getting **at most one tail**.
- b. Let B = the event of getting all tails
- c. Let C = the event of getting all heads.
- d. Let D = event of getting **more than one** tail
- e. Let E = event of getting at least one tail in two flips
- f. Let F = the event of getting two faces that are the same.
- g. Let G = the event of getting a head on the first flip followed by a head or tail on the second flip.
- h. Let H = the event of getting all tails.
- i. Are A and F mutually exclusive?
- j. Are G and H mutually exclusive?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=123#h5p-83>

Your turn!

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

- a. Let F = the event of getting the white ball twice.
- b. Let G = the event of getting two balls of different colors.
- c. Let H = the event of getting white on the first pick.
- d. Are F and G mutually exclusive?
- e. Are G and H mutually exclusive?

Image Credits

Figure 3.1: Ed Sweeney (2009). “2009 Leonid Meteor.” CC BY 2.0. Retrieved from <https://flic.kr/p/7girE8>

3.2 Visualizing Probabilities

Sometimes, when the probability problems are complex, it can be helpful to visualize the situation. Contingency tables, tree diagrams and Venn diagrams are some tools that can help us visualize and solve conditional probabilities.

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

Figure 3.3: Driving Violations

	Speeding violation in the last year	No speeding violation in the last year	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that $305 + 450 = 755$ and $70 + 685 = 755$.

Calculate the following probabilities using the table.

- Find P (Driver is a cell phone user).
- Find P (driver had no violation in the last year).
- Find P (Driver had no violation in the last year AND was a cell phone user).
- Find P (Driver is a cell phone user OR driver had no violation in the last year).

- e. Find P (Driver is a cell phone user GIVEN driver had a violation in the last year).
 f. Find P (Driver had no violation last year GIVEN driver was not a cell phone user)



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-84>

Example

The figure below contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Figure 3.4: US Crime Index Rates

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	297.7	314.7	
2009	133.1	717.7	291.1	259.2	
2010	119.3	701	277.7	239.1	
2011	113.7	702.2	268.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

- Find P (2009 AND Robbery).
- Find P (2010 AND Burglary).
- Find P (2010 OR Burglary).
- Find P (2011|Rape).
- Find P (Vehicle|2008).



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-85>

Your turn!

The following figure shows the number of athletes who stretch before exercising and how many had injuries within the past year.

Figure 3.5: Injuries Among Athletes

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

- What is $P(\text{athlete stretches before exercising})$?
- What is $P(\text{athlete stretches before exercising}|\text{no injury in the last year})$?

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of “branches” that are labeled with either frequencies or probabilities. Tree diagrams can make some sample spaces easier to visualize. The following example illustrates how to use a tree diagram.

Example

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, with replacement. “With replacement” means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.

$$\text{Total} = 64 + 24 + 24 + 9 = 121$$

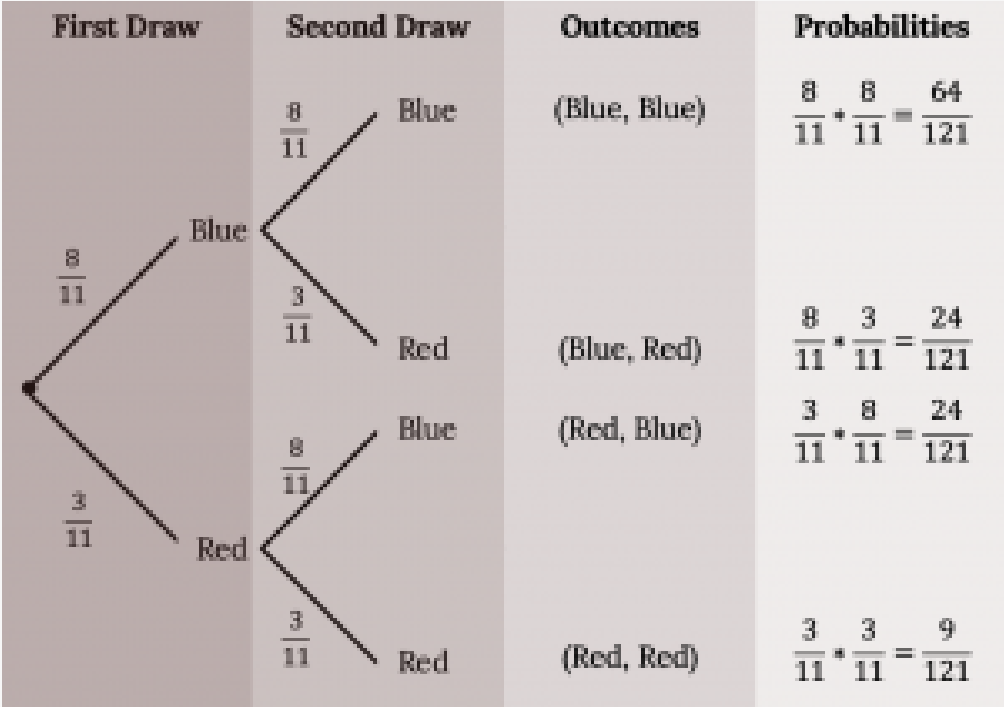


Figure 3.6: Tree Diagram

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the nine RR outcomes can be written as:

R1R1, R1R2, R1R3, R2R1, R2R2, R2R3, R3R1, R3R2, R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are $11(11) = 121$ outcomes, the size of the sample space.

- a. List the 24 BR outcomes: B1R1, B1R2, B1R3, ...



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-86>

- b. Using the tree diagram, calculate $P(RR)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-87>

- c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-88>

- d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-89>

- e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-90>

f. Using the tree diagram, calculate $P(BB)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-91>

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=129#h5p-92>

Your turn!

In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.

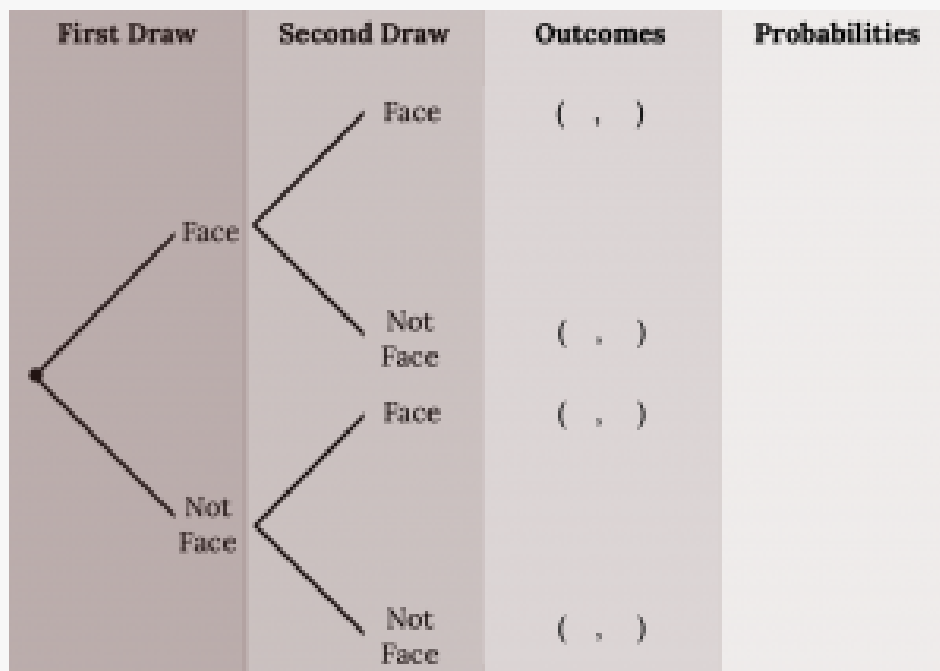


Figure 3.7: Tree Diagram

Venn Diagram

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space (S) together with circles or ovals. The circles or ovals represent events. It can often help us visualize relationships between events and compound events. We can visualize the entire rectangle as the Sample Space and areas in the circles as corresponding to probabilities.

Example

Suppose an experiment has the outcomes 1, 2, 3, ..., 12 where each outcome has an equal chance of

occurring. Let event $A = \{1, 2, 3, 4, 5, 6\}$ and event $B = \{6, 7, 8, 9\}$. Then $A \text{ AND } B = \{6\}$ and $A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The Venn diagram is as follows:

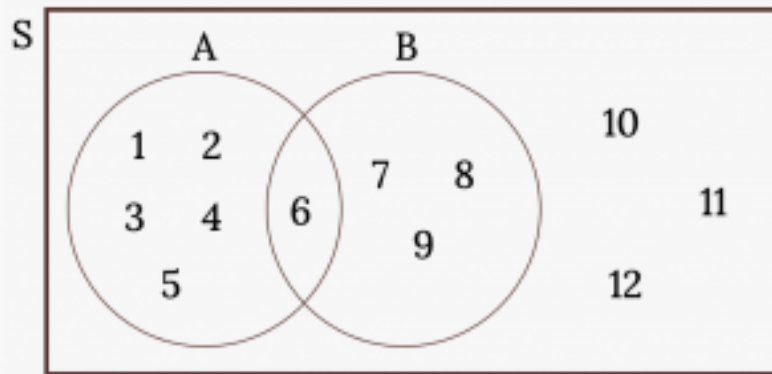


Figure 3.8: Venn Diagram

Example

Forty percent of the students at a local college belong to a club and 50% work part time. Five percent of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.

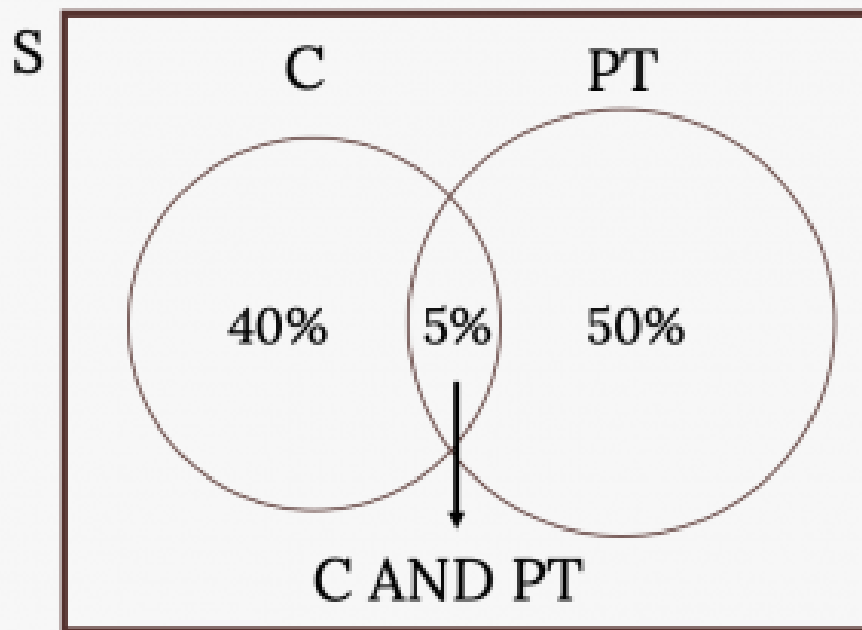


Figure 3.9: Student Activities Venn Diagram

If a student is selected at random, find

- the probability that the student belongs to a club. $P(C) = 0.40$
- the probability that the student works part time. $P(PT) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(C \text{ AND } PT) = 0.05$
- the probability that the student belongs to a club given that the student works part time.

$$P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{0.05}{0.50} = 0.1$$

- the probability that the student belongs to a club OR works part time. $P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = 0.40 + 0.50 - 0.05 = 0.85$

Your turn!

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event $C = \{\text{green, blue, purple}\}$ and event $P = \{\text{red, yellow, blue}\}$. Then $C \text{ AND } P = \{\text{blue}\}$ and $C \text{ OR } P = \{\text{green, blue, purple, red, yellow}\}$. Draw a Venn diagram representing this situation.

Image References

Figure 3.6: Kindred Grey (2020). "Figure 3.6 idea." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.6_idea.png

Figure 3.7: Kindred Grey (2020). "Figure 3.7 idea." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.7_idea.png

Figure 3.8: Kindred Grey (2020). "Figure 3.8." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.8.png

Figure 3.9: Kindred Grey (2020). "Figure 3.9." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.9.png

3.3 Compound Events

Recall the different combinations of relationships between two events:

Figure 3.10: Relationships Between Events

		Independent?	
		Yes	No
Disjoint?	Yes	1*	2
	No	3	4

We must always go into a problem assuming two events are not **mutually exclusive** or **independent**. This “default” starting point is illustrated by the 4th position in the table above. Depending on the information your are given and assumptions you are able to make, you may move potions on this grid. Where you fall on the grid will dictate how we apply the rules we will discuss in this section to find probabilities of compound events.

There are two types of compound events we may be interested in, Unions and Intersections, each with their own set of rules and assumptions.

***Note:** You will rarely, if ever, find yourself in this case

Finding Probabilities of Unions

To find a **union** we will typically use the **addition rule**. Intuitively the idea is that if we are looking for the outcomes in either event A or Event B, we should be able to simply add up the probabilities of each outcome. However, two events being mutually exclusive has big implications on how we apply the addition rule.

For Two Mutually Exclusive Events

The idea is simple when events are **mutually exclusive**. Picture a Venn diagram of two mutually exclusive events.

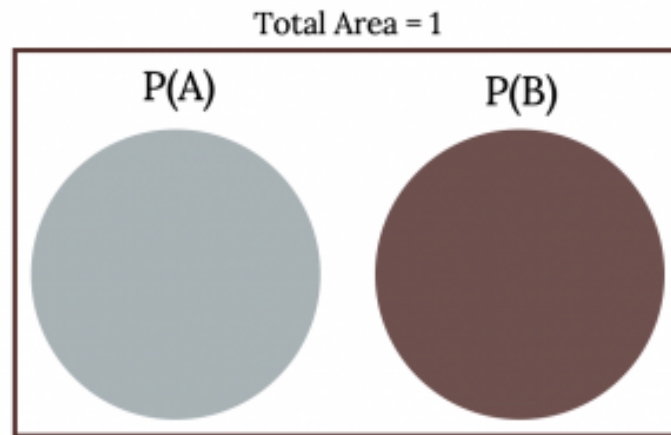


Figure 3.11: Two Mutually Exclusive Events

Not very exciting, but we need to note here that if A and B are mutually exclusive, then they have no shared outcomes. In other words no **intersection** exists between two disjoint events. In probability notation this means $A \cap B = \emptyset$ and $P(A \cap B) = 0$.

In this case, the Union of A OR B is simply:

$$P(A \cup B) = P(A) + P(B)$$

This is reflected in the previously mentioned *Third Axiom of Probability* (also called the disjoint addition rule).

Recall:

3. For each two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$ then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Example

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is $P(A) = 0.6$ and the probability that he chooses B is $P(B) = 0.35$.
- $P(A \text{ AND } B) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \text{ OR } B) = P(A) + P(B) = 0.6 + 0.35 = 0.95$.
- Note that the probability that he does not choose to go anywhere on vacation (the complement) must then be 0.05.

For Two Non-Mutually Exclusive Events

When two events are not mutually exclusive it gets a bit trickier. Consider a Venn diagram of two non-mutually exclusive events.

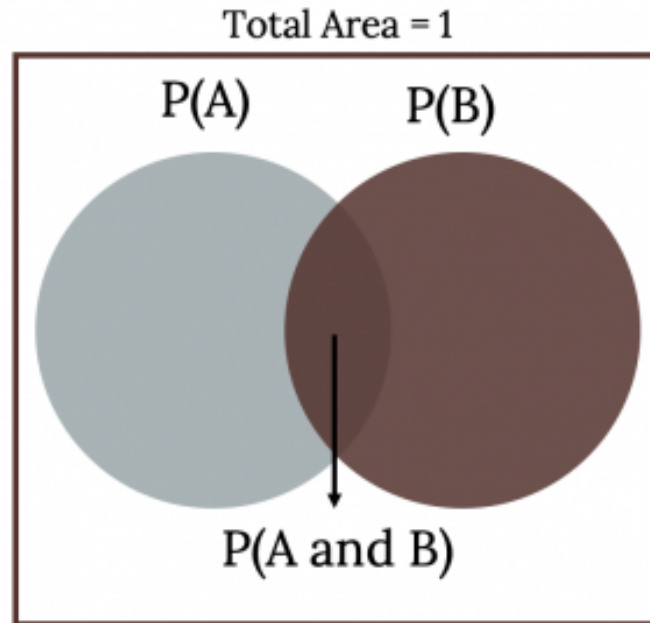


Figure 3.12: Two Non-Mutually Exclusive Events

Here we can see that when A and B are *not* mutually exclusive, then they *do* have shared outcomes, or an intersection. If we try to apply the addition rule, we need to be careful not to double count those shared outcomes

If A and B are defined on a sample space, then: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Example

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. Are A and B independent?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-93>

b. Are A and B mutually exclusive?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-94>

c. What is the probability that he makes both goals?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-95>

d. What is the probability that Carlos makes either the first goal or the second goal?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-96>

Your turn!

Felicity attends Reynolds CC in Richmond, VA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, M|S = math given speech

a. Are M and S independent? Is $P(M|S) = P(M)$?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-97>

b. Are M and S mutually exclusive? Is $P(M \text{ AND } S) = 0$?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-98>

c. What is the probability that Felicity enrolls in math and speech?

Find $P(M \text{ AND } S) = P(M|S)P(S)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-99>

d. What is the probability that Felicity enrolls in math or speech classes?

Find $P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S)$.





An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=134#h5p-100>

Applying the Addition Rule to Multiple Events

Let's try to extend the ideas of the addition rule to more than two events. Again it will depend on whether these events are mutually exclusive or not.

More Than Two Mutually Exclusive Events

Let's start by visualizing a Venn diagram of with 3 mutually exclusive events, A, B, and C. There would still be no intersections and we can simply apply the disjoint addition Rule resulting in:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

Extending this beyond just three events follows easily. If they are mutually exclusive none of them have intersections and we should be able to apply our disjoint addition rule infinitely.

$$P(A \cup B \cup \dots \cup N) = P(A) + P(B) + \dots + P(N)$$

As long as things are mutually exclusive we can just keep adding as many events we would like!

More Than Two Non-Mutually Exclusive Events

As with only two events things get a little bit trickier when we do have shared outcomes. Consider the Venn diagram below of three non-mutually exclusive events.

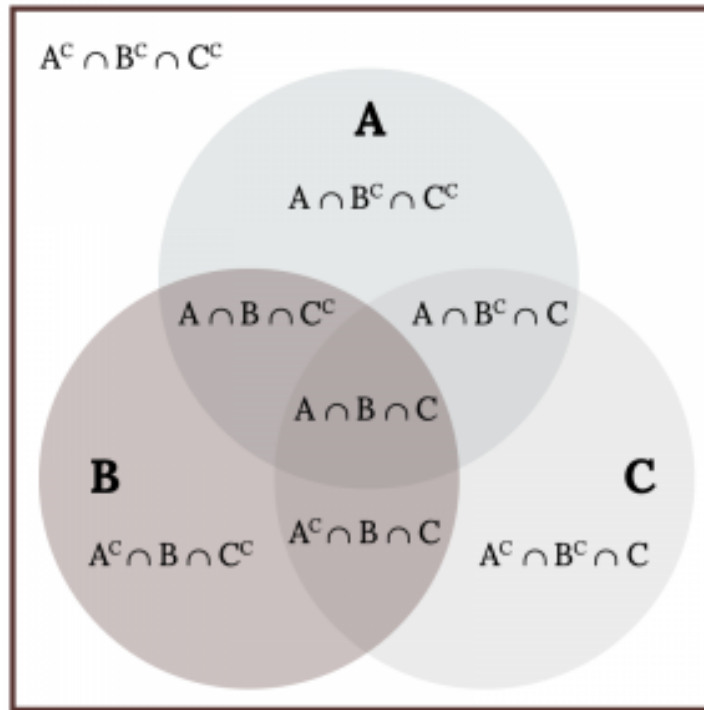


Figure 3.13: Three Non-Mutually Exclusive Events

We have multiple double counting of intersections issues here. We could try subtracting the intersections as we did with two events, but that causes a new problem. If you subtract all three intersections once, you have now subtracted the triple intersection ($A \cap B \cap C$) three times thus not counting those outcomes at all! So after subtracting the three intersections to eliminate double counting we need to add back the triple intersection finally resulting in:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Imagine extending this beyond 3 events! It's a bit messy and changes slightly based on whether we have an odd or even number of events but is certainly doable if we are meticulous.

Finding Probabilities of Intersections

Let's review a couple methods we have already seen that will help us find the **intersection** of two events:

1. If two events are mutually exclusive we have already established they have no intersection i.e. $P(A \cap B) = 0$.
2. If we are given the right pieces we could even rearrange the addition rule as: $P(A \cap B) = P(B) + P(A) - P(A \cup B)$

Beyond these to find the probability of an intersection we can use the **multiplication rule**:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

Notice we now have **conditional probabilities** involved and two different forms of the rule.

Finding a Conditional Probability

One way we can come up with conditional probabilities is to rewrite the multiplication rule as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(Notice what you are given goes in the denominator)

However, if we do not know the probability of the intersection this can leave us in a loop. In many cases we may just need to think carefully about a situation to come up with these conditional probabilities to apply our multiplication rule.

Conditional Probabilities for Two Independent Events

We now need to consider the implications of **independence** on conditional probabilities. Recall two independent events are events that have no effect on each other.

First, consider two dependent events: Let $P(O)$ be the probability you oversleep and $P(B)$ be the probability you eat breakfast. Oversleeping will likely have an effect on the probability you eat breakfast that morning so $P(B|O)$ is of interest.

Now consider two independent events: Let $P(G)$ be the probability you have green eyes and $P(B)$ the probability you eat breakfast. Your eye color has no effect on whether you will eat breakfast or not, therefore $P(B|G)$ is simply the same as $P(B)$.

The Multiplication Rule for Two Independent Events

Now let's think about the implications that independence then has on the multiplication rule. If A and B are independent then:

$P(A|B) = P(A)$. So the multiplication rule, $P(A \cap B) = P(A|B)P(B)$ becomes $P(A \cap B) = P(A)P(B)$.

If two events are independent conditional probabilities are eliminated from the multiplication rule and things become much simpler

Showing Independence of Two Events

Knowing what we now know about the implications of Independence on conditional probabilities and the multiplication rule, we can establish the following conditions. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ AND } B) = P(A)P(B)$

Note that we only have to show one, all of these conditions are equivalent and imply each other.

Applying the Multiplication Rule to Multiple Events

Now we'll extend the ideas of the multiplication rule to more than two events. How we apply it will again depend on whether these events are independent.

More than two independent events

We saw that if events were independent conditional probabilities were eliminated from the multiplication rule and things got pretty easy. Extending the independent multiplication rule to three events would look like:

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

Extending this beyond three events:

$$P(A \cap B \cap \dots \cap N) = P(A)P(B) \dots * P(N)$$

This makes it very easy to find the probability of a sequence of independent events.

More than two dependent events

If we have a sequence of three dependent events we will have to sequentially update conditional probabilities. For example:

$$P(A \cap B \cap C) = P(A)*P(B|A)*P(C|A \cap B)$$

This is doable for just a handful of events but could get quite messy for a lot of dependent events.

Image References

Figure 3.11: Kindred Grey (2020). “Figure 3.11.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.11.png

Figure 3.12: Kindred Grey (2020). “Figure 3.12.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.12.png

Figure 3.13: Kindred Grey (2020). “Figure 3.13.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_3.13.png

Chapter 3 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=152#h5p-101>

Section Reviews

3.1 Introduction

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$

3.2 Visualizing Probabilities

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

3.3 Compound Events

The multiplication rule and the addition rule are used for computing the probability of A and B , as well as the probability of A or B for two given events A , B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

The multiplication rule: $P(A \text{ AND } B) = P(A|B)P(B)$

The addition rule: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

3.1 Introduction

- Probability
- Probability experiment
- Outcome
- Sample space
- Event
- Probability model
- Law of large numbers
- Mutually exclusive
- Independent
- Complement
- Intersection
- Conditional probability

3.2 Visualizing Probabilities

- Contingency table
- Tree diagram
- Venn diagram

3.3 Compound Events

- Mutually exclusive
- Independent
- Union
- Intersection
- Conditional probabilities
- Independence

Extra Practice

3.1 Introduction

1. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

- Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?
 - Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?
-

2. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs.

- Suppose you pick four cards, but do not put any cards back into the deck. Your cards are QS, 1D, 1C, QD.
- Suppose you pick four cards and put each card back before you pick the next card. Your cards are KH, 7D, 6D, KH.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

3. You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=152#h5p-102>

4. Flip two fair coins. The sample space is {HH, HT, TH, TT} where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as {HH, HT, TH}. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as {TT}. B is the **complement** of A, so $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for A and for B are $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{4}$.
- Let C = the event of getting all heads. $C = \{HH\}$. Since $B = \{TT\}$, $P(B \text{ AND } C) = 0$. B and C are mutually exclusive. (B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $D = \{TT\}$. $P(D) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. $F = \{HT, TH, TT\}$. $P(F) = \frac{3}{4}$

5. Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

6. Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then $A = \{1, 3, 5\}$. Let event B = a face is even. Then $B = \{2, 4, 6\}$.

- Find the complement of A, A' . The complement of A, A' , is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then $C = \{3, 5\}$. Let event D = all even faces smaller than five. Then $D = \{2, 4\}$. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.

- Let event E = all faces less than five. $E = \{1, 2, 3, 4\}$.

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

No. $C = \{3, 5\}$ and $E = \{1, 2, 3, 4\}$. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is $\{3, 5\}$ and event A is $\{1, 3, 5\}$. To find $P(C|A)$, find the probability of C using the sample space A . You have reduced the sample space from the original sample space $\{1, 2, 3, 4, 5, 6\}$ to $\{1, 3, 5\}$. So, $P(C|A) = \frac{2}{3}$

7. Let event A = learning Spanish. Let event B = learning German. Then $A \text{ AND } B$ = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show **ONE** of the following:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A)$$

The events are independent because $P(A|B) = P(A)$.

8. Let event G = taking a math class. Let event H = taking a science class. Then, $G \text{ AND } H$ = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

NOTE: The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. Show that $P(G|H) = P(G)$.

$$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$$

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

$$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

9. In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

10. A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

- a. Find $P(B|D)$.
 - b. Find $P(D|B)$.
 - c. Are B and D independent?
 - d. Are B and D mutually exclusive?
-

11. In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$. S has eight outcomes.

- $P(R) = \frac{3}{8}$, $P(B) = \frac{5}{8}$, $P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)
- $P(E) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(E|B) = \frac{2}{5}$. (There are five blue cards: $B1, B2, B3, B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E) = \frac{2}{3}$. (There are three even-numbered cards: $R2, B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.

- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = \frac{1}{4}$. (The only card in H that has a number greater than three is $B4$.) Since $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, which means that G and H are independent.
-

12. In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

13. In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

NOTE: **The choice you make depends on the information you have.** You could use the first or last condition on the list for this example. You do not know $P(F|L)$ yet, so you cannot use the second condition.

Solution 1 Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Solution 2 Check whether $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

14. Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.56$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

15. Fill in the blanks to the following questions.

- Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are _____ outcomes.
 - Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are _____ outcomes.
 - Multiply the two numbers of outcomes. The answer is _____.
 - If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in part c. is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are H1 and T6.)
 - Event A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.
A = {_____}. Find $P(A)$.
 - Event B = heads on the coin followed by a three on the die. B = {_____}. Find $P(B)$.
 - Are A and B mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.)
 - Are A and B independent? (Hint: Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then A and B are independent. If not, then they are dependent).
- H and T; 2
 - 1, 2, 3, 4, 5, 6; 6
 - $2(6) = 12$
 - T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6
 - A = {H2, H4, H6}; $P(A) = \frac{3}{12}$
 - B = {H3}; $P(B) = \frac{1}{12}$
 - Yes, because $P(A \text{ AND } B) = 0$
 - $P(A \text{ AND } B) = 0.P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so A and B are dependent.
-

16. A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

- Compute $P(T)$.
- Compute $P(T|F)$.

- c. Are T and F independent?
 - d. Are F and S mutually exclusive?
 - e. Are F and S independent?
-

17. E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E \mid F)$.

18. J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

- $P(J) = 0.3$
-

19. U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

- a. $P(U \text{ AND } V) =$
 - b. $P(U|V) =$
 - c. $P(U \text{ OR } V) =$
-

20. Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

- $P(Q \text{ AND } R) = P(Q)P(R)$
 - $0.1 = (0.4)P(R)$
 - $P(R) = 0.25$
-

21. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

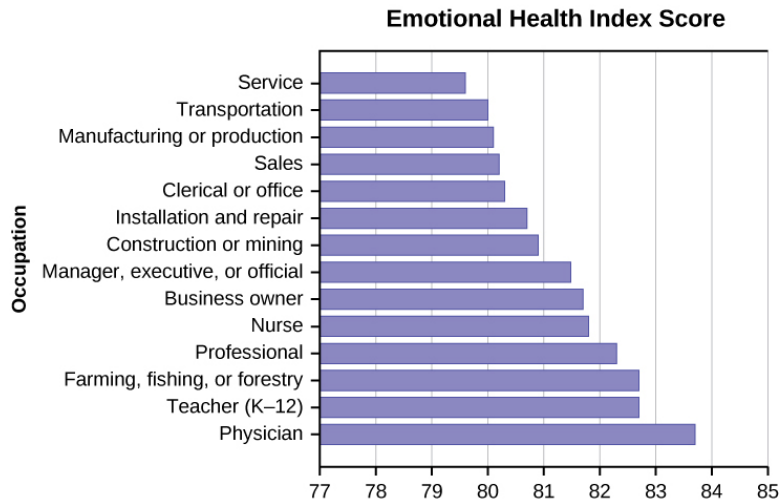


Figure 3.14

- Find the probability that an Emotional Health Index Score is 82.7.
- Find the probability that an Emotional Health Index Score is 81.0.
 - 0
- Find the probability that an Emotional Health Index Score is more than 81?
- Find the probability that an Emotional Health Index Score is between 80.5 and 82?
 - 0.3571
- If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?
- What is the probability that an Emotional Health Index Score is 80.7 or 82.7?
 - 0.2142
- What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81?
- What occupation has the highest emotional index score?
 - Physician (83.7)
- What occupation has the lowest emotional index score?

j. What is the range of the data?

- $83.7 - 79.6 = 4.1$

k. Compute the average EHIS.

l. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

- $P(\text{Occupation} < 81.3) = 0.5$
-

22. A previous year, the weights of the members of the San Francisco 49ers and the Dallas Cowboys were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into the figure below.¹

Figure 3.15

Shirt#	≤ 210	211–250	251–290	$290 \leq$
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\# } 1\text{--}33 | \leq 210 \text{ pounds})$?

23. The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write “not enough information” for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

- $P(C) = \underline{\hspace{2cm}}$
- $P(P|C) = \underline{\hspace{2cm}}$
- $P(P|C^c) = \underline{\hspace{2cm}}$
- If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

- a. $P(C) = 0.4567$
 - b. not enough information
 - c. not enough information
 - d. No, because over half (0.51) of men have at least one false positive test
-

24. Given events G and H : $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) = 0.14$

- a. Find $P(H \text{ OR } G)$.
 - b. Find the probability of the complement of event $(H \text{ AND } G)$.
 - c. Find the probability of the complement of event $(H \text{ OR } G)$.
-

25. Given events J and K : $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

- a. Find $P(J \text{ AND } K)$.
 - b. Find the probability of the complement of event $(J \text{ AND } K)$.
 - c. Find the probability of the complement of event $(J \text{ OR } K)$.
- a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37 - P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$
 - b. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 0.10 = 0.90$
 - c. $P(\text{NOT } (J \text{ OR } K)) = 1 - P(J \text{ OR } K) = 1 - 0.45 = 0.55$
-

26. The sample space S is the whole numbers starting at one and less than 20.

- a. $S = \underline{\hspace{4cm}}$

Let event A = the even numbers and event B = numbers greater than 13.

- b. $A = \underline{\hspace{4cm}}$, $B = \underline{\hspace{4cm}}$
 - c. $P(A) = \underline{\hspace{2cm}}$, $P(B) = \underline{\hspace{2cm}}$
 - d. $A \text{ AND } B = \underline{\hspace{4cm}}$, $A \text{ OR } B = \underline{\hspace{4cm}}$
 - e. $P(A \text{ AND } B) = \underline{\hspace{2cm}}$, $P(A \text{ OR } B) = \underline{\hspace{2cm}}$
 - f. $A' = \underline{\hspace{4cm}}$, $P(A') = \underline{\hspace{2cm}}$
 - g. $P(A) + P(A') = \underline{\hspace{2cm}}$
 - h. $P(A|B) = \underline{\hspace{2cm}}$, $P(B|A) = \underline{\hspace{2cm}}$; are the probabilities equal?
- a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
 - b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
 - c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$

- d. $A \text{ AND } B = \{14, 16, 18\}$, $A \text{ OR } B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
 e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$
 f. $A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(A') = \frac{10}{19}$
 g. $P(A) + P(A') = 1$ ($\frac{9}{19} + \frac{10}{19} = 1$)
 h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{3}{9}$ No
-

27. The sample space S is all the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

- a. $S =$ _____ Let event A = the sum is even and event B = the first number is prime.
 b. $A =$ _____, $B =$ _____
 c. $P(A) =$ _____, $P(B) =$ _____
 d. $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____
 e. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____
 f. $B' =$ _____, $P(B') =$ _____
 g. $P(A) + P(A') =$ _____
 h. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?
-

28. A fair, six-sided die is rolled. Describe the sample space S , identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

- a. Event T = the outcome is two.
 b. Event A = the outcome is an even number.
 c. Event B = the outcome is less than four.
 d. The complement of A .
 e. $A \text{ GIVEN } B$
 f. $B \text{ GIVEN } A$
 g. $A \text{ AND } B$
 h. $A \text{ OR } B$
 i. $A \text{ OR } B'$
 j. Event N = the outcome is a prime number.
 k. Event I = the outcome is seven.

- a. $T = \{2\}$, $P(T) = \frac{1}{6}$
 b. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
 c. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$

- d. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
 - e. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
 - f. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
 - g. $A \text{ AND } B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$
 - h. $A \text{ OR } B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
 - i. $A \text{ OR } B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$
 - j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
 - k. A six-sided die does not have seven dots. $P(7) = 0$.
-

29. The figure below describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

Figure 3.16

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

- a. $P(M)$
 - b. $P(F)$
 - c. $P(R)$
 - d. $P(L)$
 - e. $P(M \text{ AND } R)$
 - f. $P(F \text{ AND } L)$
 - g. $P(M \text{ OR } F)$
 - h. $P(M \text{ OR } R)$
 - i. $P(F \text{ OR } L)$
 - j. $P(M')$
 - k. $P(R|M)$
 - l. $P(F|L)$
 - m. $P(L|F)$
-
- a. $P(M) = 0.52$
 - b. $P(F) = 0.48$
 - c. $P(R) = 0.87$
 - d. $P(L) = 0.13$
 - e. $P(M \text{ AND } R) = 0.43$

- f. $P(F \text{ AND } L) = 0.04$
 - g. $P(M \text{ OR } F) = 1$
 - h. $P(M \text{ OR } R) = 0.96$
 - i. $P(F \text{ OR } L) = 0.57$
 - j. $P(M') = 0.48$
 - k. $P(R|M) = 0.8269$ (rounded to four decimal places)
 - l. $P(F|L) = 0.3077$ (rounded to four decimal places)
 - m. $P(L|F) = 0.0833$
-

30. In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
 - Let M be the event that a student is male.
 - Let S be the event that a student has short hair.
 - Let L be the event that a student has long hair.
- a. The probability that a student does not have long hair.
 - b. The probability that a student is male or has short hair.
 - c. The probability that a student is a female and has long hair.
 - d. The probability that a student is male, given that the student has long hair.
 - e. The probability that a student has long hair, given that the student is male.
 - f. Of all the female students, the probability that a student has short hair.
 - g. Of all students with long hair, the probability that a student is female.
 - h. The probability that a student is female or has long hair.
 - i. The probability that a randomly selected student is a male student with short hair.
 - j. The probability that a student is female.
- a. $P(L') = P(S)$
 - b. $P(M \text{ OR } S)$
 - c. $P(F \text{ AND } L)$
 - d. $P(M|L)$
 - e. $P(L|M)$
 - f. $P(S|F)$
 - g. $P(F|L)$
 - h. $P(F \text{ OR } L)$
 - i. $P(M \text{ AND } S)$
 - j. $P(F)$
-

31. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let H = the event of getting a hat.

Let N = the event of getting a noisemaker.

Let F = the event of getting a finger trap.

Let C = the event of getting a bag of confetti.

a. Find $P(H)$.

b. Find $P(N)$.

- $P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

c. Find $P(F)$.

d. Find $P(C)$.

- $P(C) = \frac{5}{42} = 0.12$

32. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let B = the event of getting a blue jelly bean

Let G = the event of getting a green jelly bean.

Let O = the event of getting an orange jelly bean.

Let P = the event of getting a purple jelly bean.

Let R = the event of getting a red jelly bean.

Let Y = the event of getting a yellow jelly bean.

a. Find $P(B)$.

b. Find $P(G)$.

- $P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

c. Find $P(P)$.

d. Find $P(R)$.

- $P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

e. Find $P(Y)$.

f. Find $P(O)$.

- $P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

33. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let A = the event that a country is in Asia.

Let E = the event that a country is in Europe.

Let F = the event that a country is in Africa.

Let N = the event that a country is in North America.

Let O = the event that a country is in Oceania.

Let S = the event that a country is in South America.

a. Find $P(A)$.

b. Find $P(E)$.

- $P(E) = \frac{47}{194} = 0.24$

c. Find $P(F)$.

d. Find $P(N)$.

- $P(N) = \frac{23}{194} = 0.12$

e. Find $P(O)$.

f. Find $P(S)$.

- $P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

g. What is the probability of drawing a red card in a standard deck of 52 cards?

h. What is the probability of drawing a club in a standard deck of 52 cards?

- $\frac{13}{52} = \frac{1}{4} = 0.25$

i. What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

j. What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

- $\frac{3}{6} = \frac{1}{2} = 0.5$

34. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

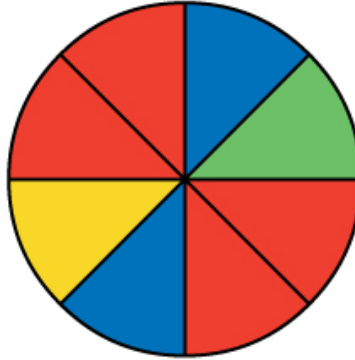


Figure 3.17

Let B = the event of landing on blue.
Let R = the event of landing on red.
Let G = the event of landing on green.
Let Y = the event of landing on yellow.

- If you land on Y, you get the biggest prize. Find $P(Y)$.
- If you land on red, you don't get a prize. What is $P(R)$?

- $P(R) = \frac{4}{8} = 0.5$

35. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let I = the event that a player is an infielder.
Let O = the event that a player is an outfielder.
Let H = the event that a player is a great hitter.
Let N = the event that a player is not a great hitter.

- Write the symbols for the probability that a player is not an outfielder.
- Write the symbols for the probability that a player is an outfielder or is a great hitter.

- $P(O \text{ OR } H)$

- Write the symbols for the probability that a player is an infielder and is not a great hitter.

d. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

- $P(H|I)$

e. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

f. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

- $P(N|O)$

g. Write the symbols for the probability that of all the great hitters, a player is an outfielder.

h. Write the symbols for the probability that a player is an infielder or is not a great hitter.

- $P(I \text{ OR } N)$

i. Write the symbols for the probability that a player is an outfielder and is a great hitter.

j. Write the symbols for the probability that a player is an infielder.

- $P(I)$

k. What is the word for the set of all possible outcomes?

l. What is conditional probability?

m. The likelihood that an event will occur given that another event has already occurred.

36. A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let F = event that book is fiction

Let N = event that book is nonfiction

What is the sample space?

What is the sum of the probabilities of an event and its complement?

- 1

37. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

- What does $P(E|M)$ mean in words?
- What does $P(E \text{ OR } M)$ mean in words?
- the probability of landing on an even number or a multiple of three

38. The graph below displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

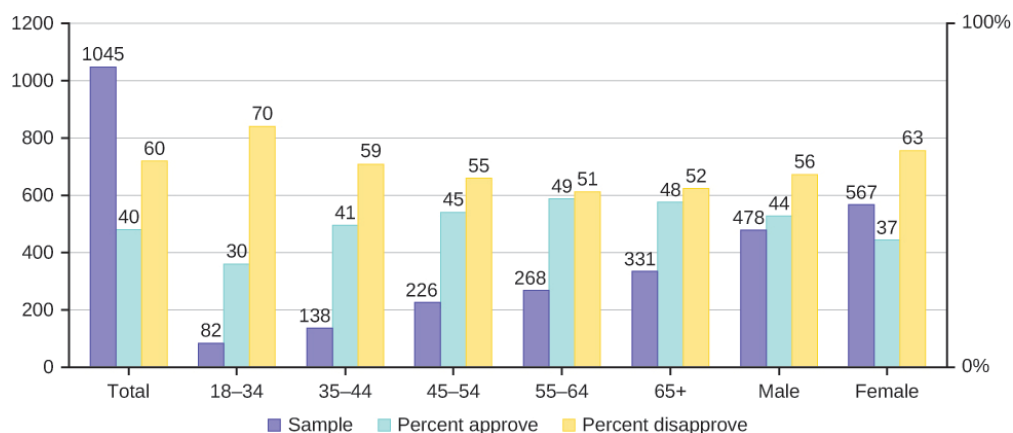


Figure 3.18

- Define three events in the graph.
- Describe in words what the entry 40 means.
- Describe in words the complement of the entry in question 2.
- Describe in words what the entry 30 means.
- Out of the males and females, what percent are males?
- Out of the females, what percent disapprove of Mayor Ford?
- Out of all the age groups, what percent approve of Mayor Ford?
- Find $P(\text{Approve}|\text{Male})$.
- Out of the age groups, what percent are more than 44 years old?
- Find $P(\text{Approve}|\text{Age} < 35)$.

39. Explain what is wrong with the following statements. Use complete sentences.

- If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.

- b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.
- a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
- b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

3.2 Visualizing Probabilities

1. The following figure shows a random sample of 100 hikers and the areas of hiking they prefer.

Figure 3.19: Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	---	45
Male	---	---	14	55
Total	---	41	---	---

- a. Complete the table.

a.

Figure 3.20: Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

- b. Are the events “being female” and “preferring the coastline” independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$

$$2. P(F)P(C) = \left(\frac{45}{100}\right) \left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(\text{---}|\text{---}) = \text{---}$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

c. The word 'given' tells you that this is a conditional.

1. $P(M|L) = \frac{25}{41}$
2. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

d.

1. $P(F) = \frac{45}{100}$
2. $P(P) = \frac{25}{100}$
3. $P(F \text{ AND } P) = \frac{11}{100}$
4. $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

2. The figure below shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Figure 3.21

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

- a. Out of the males, what is the probability that the cyclist prefers a hilly path?
 - b. Are the events “being male” and “preferring the hilly path” independent events?
-

3. Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Figure 3.22: Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	----
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	----
Total	----	----	----	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right) \left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right) \left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

a.

Figure 3.23: Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. What is the probability that Alissa does not catch Muddy?

b. $\frac{41}{60}$

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

c. $\frac{9}{19}$

4. The figure below relates the weights and heights of a group of individuals participating in an observational study.

Figure 3.24

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

- Find the total for each row and column
- Find the probability that a randomly chosen individual from this group is Tall.
- Find the probability that a randomly chosen individual from this group is Obese and Tall.
- Find the probability that a randomly chosen individual from this group is Tall given that the individual is Obese.
- Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
- Find the probability a randomly chosen individual from this group is Tall and Underweight.
- Are the events Obese and Tall independent?

5. There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. The figure below shows a random sample of musicians and how they learned to play their instruments.

Figure 3.25

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

- Find $P(\text{musician is a female})$.
- Find $P(\text{musician is a male AND had private instruction})$.

- $P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$

c. Find $P(\text{musician is a female OR is self taught})$.

d. Are the events “being a female musician” and “learning music in school” mutually exclusive events?

- $P(\text{being a female musician AND learning music in school}) = \frac{38}{130} = \frac{19}{65} = 0.29$
- No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

6. An article in the NEW ENGLAND JOURNAL OF MEDICINE, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.²

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Figure 3.26: Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

a. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

2. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at <http://www.nejm.org/doi/full/10.1056/NEJMoa033250> (accessed May 2, 2013).

- $\frac{35,065}{100,450}$

b. Find the probability that the person was Latino.

c. In words, explain what it means to pick one person from the study who is “Japanese American **AND** smokes 21 to 30 cigarettes per day.” Also, find the probability.

- To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

d. In words, explain what it means to pick one person from the study who is “Japanese American **OR** smokes 21 to 30 cigarettes per day.” Also, find the probability.

e. In words, explain what it means to pick one person from the study who is “Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day.” Also, find the probability.

- To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

f. Prove that smoking level/day and ethnicity are dependent events.

7. The figure below contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S. ³

Figure 3.27: United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

3. United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at <http://www.disastercenter.com/crime/> (accessed May 2, 2013).

- Find $P(2009 \text{ AND Robbery})$.
- Find $P(2010 \text{ AND Burglary})$.
- Find $P(2010 \text{ OR Burglary})$.
- Find $P(2011|Rape)$.
- Find $P(\text{Vehicle}|2008)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=152#h5p-85>

8. The following figure shows a random sample of 100 hikers and the areas of hiking they prefer.

Figure 3.28: Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	---	45
Male	---	---	14	55
Total	---	41	---	---

- Complete the table.
- Are the events “being female” and “preferring the coastline” independent events? Let F = being female and let C = preferring the coastline.
 - Find $P(F \text{ AND } C)$.
 - Find $P(F)P(C)$
 - Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.
- Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.
 - What word tells you this is a conditional?
 - Fill in the blanks and calculate the probability: $P(____| ____) = ____$.
 - Is the sample space for this problem all 100 hikers? If not, what is it?
- Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.
 - Find $P(F)$.

2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solutions:

a.

Figure 3.29: Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

b.

- $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
- $P(F)P(C) = \left(\frac{45}{100}\right) \left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$
- $P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

c.

- The word 'given' tells you that this is a conditional.
- $P(M|L) = \frac{25}{41}$
- No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

- $P(F) = \frac{45}{100}$
- $P(P) = \frac{25}{100}$
- $P(F \text{ AND } P) = \frac{11}{100}$
- $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

9. The figure below shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Figure 3.30

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

- Out of the males, what is the probability that the cyclist prefers a hilly path?
- Are the events “being male” and “preferring the hilly path” independent events?

10. Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Figure 3.31: Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	----
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	----
Total	----	----	----	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right) \left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right) \left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

- Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

a.

Figure 3.32: Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. What is the probability that Alissa does not catch Muddy?

b. $\frac{41}{60}$

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

c. $\frac{9}{19}$

11. The figure below relates the weights and heights of a group of individuals participating in an observational study.

Figure 3.33

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

- Find the total for each row and column
- Find the probability that a randomly chosen individual from this group is Tall.
- Find the probability that a randomly chosen individual from this group is Obese and Tall.
- Find the probability that a randomly chosen individual from this group is Tall given that the individual is Obese.
- Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
- Find the probability a randomly chosen individual from this group is Tall and Underweight.
- Are the events Obese and Tall independent?

12. Use the following information to answer the next four exercises. The figure below shows a random sample of musicians and how they learned to play their instruments.

Figure 3.34

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Find $P(\text{musician is a female})$.

Find $P(\text{musician is a male AND had private instruction})$.

$$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$$

Find $P(\text{musician is a female OR is self taught})$.

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

$$P(\text{being a female musician AND learning music in school}) = \frac{38}{130} = \frac{19}{65} = 0.29$$

No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

13. Use the following information to answer the next seven exercises. An article in the NEW ENGLAND JOURNAL OF MEDICINE, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.⁴

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

4. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at <http://www.nejm.org/doi/full/10.1056/NEJMoa033250> (accessed May 2, 2013).

Figure 3.35: Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

$$\frac{35,065}{100,450}$$

Find the probability that the person was Latino.

In words, explain what it means to pick one person from the study who is “Japanese American **AND** smokes 21 to 30 cigarettes per day.” Also, find the probability.

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

In words, explain what it means to pick one person from the study who is “Japanese American **OR** smokes 21 to 30 cigarettes per day.” Also, find the probability.

In words, explain what it means to pick one person from the study who is “Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day.” Also, find the probability.

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

Prove that smoking level/day and ethnicity are dependent events.

14. In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, **with replacement**. “With replacement” means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.

$$\text{Total} = 64 + 24 + 24 + 9 = 121$$

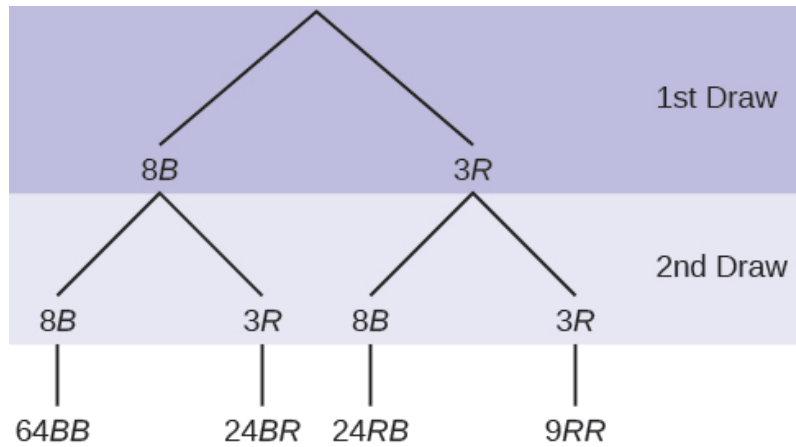


Figure 3.36

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the nine RR outcomes can be written as:

R1R1R1R2R1R3R2R1R2R2R2R3R3R1R3R2R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are $11(11) = 121$ outcomes, the size of the sample space.

a. List the 24 BR outcomes: B1R1, B1R2, B1R3, ...

- B1R1B1R2B1R3B2R1B2R2B2R3B3R1B3R2B3R3B4R1B4R2B4R3B5R1B5R2B5R3B6R1B6R2B6R3B7R1B7R2B7R3B8R1B8R2B8R3

b. Using the tree diagram, calculate $P(RR)$.

- $P(RR) = \left(\frac{3}{11}\right) \left(\frac{3}{11}\right) = \frac{9}{121}$

c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

- $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right) \left(\frac{8}{11}\right) + \left(\frac{8}{11}\right) \left(\frac{3}{11}\right) = \frac{48}{121}$

d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

- $P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = P(RB) = \left(\frac{3}{11}\right) \left(\frac{8}{11}\right) = \frac{24}{121}$

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

- $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw}) = P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are $24 + 64 = 88$ possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

f. Using the tree diagram, calculate $P(BB)$.

- $P(BB) = \frac{64}{121}$

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw} | R \text{ on the first draw})$.

- $P(B \text{ on 2nd draw} | R \text{ on 1st draw}) = \frac{8}{11}$

There are $9 + 24$ outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then $9 + 24 = 33$. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

15. An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **“Without replacement”** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example:

$$\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}.$$

$$\text{Total} = \frac{56+24+24+6}{110} = \frac{110}{110} = 1$$

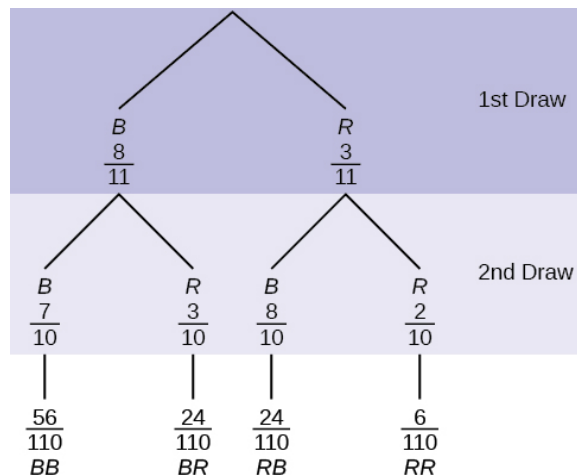


Figure 3.37

NOTE: If you draw a red on the first draw from the three red possibilities, there are two red marbles left to

draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

a. $P(RR) = \underline{\hspace{2cm}}$

- $P(RR) = \left(\frac{3}{11}\right) \left(\frac{2}{10}\right) = \frac{6}{110}$

b. Fill in the blanks: $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right) \left(\frac{8}{10}\right) + (\underline{\hspace{1cm}})(\underline{\hspace{1cm}}) = \frac{48}{110}$

- $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right) \left(\frac{8}{10}\right) + \left(\frac{8}{11}\right) \left(\frac{3}{10}\right) = \frac{48}{110}$

c. $P(R \text{ on 2nd} | B \text{ on 1st}) =$

- $\frac{3}{10}$

d. Fill in the blanks. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\underline{\hspace{1cm}})(\underline{\hspace{1cm}}) = \frac{24}{100}$

- $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right) \left(\frac{8}{10}\right) = \frac{24}{100}$

e. Find $P(BB)$.

- $P(BB) = \left(\frac{8}{11}\right) \left(\frac{7}{10}\right)$

f. Find $P(B \text{ on 2nd} | R \text{ on 1st})$.

- Using the tree diagram, $P(B \text{ on 2nd} | R \text{ on 1st}) = P(R|B) = \frac{8}{10}$

If we are using probabilities, we can label the tree in the following general way.

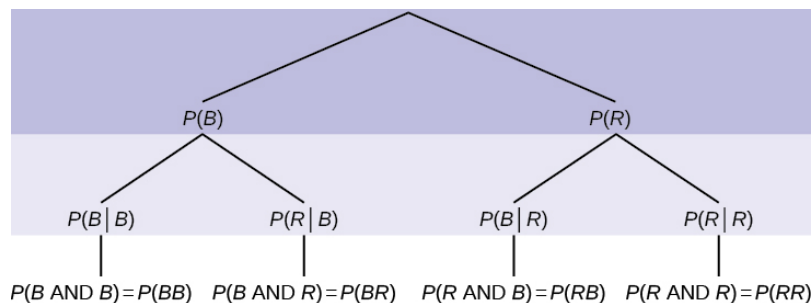


Figure 3.38

- $P(R|R)$ here means $P(\text{R on 2nd}|\text{R on 1st})$
- $P(B|R)$ here means $P(\text{B on 2nd}|\text{R on 1st})$
- $P(R|B)$ here means $P(\text{R on 2nd}|\text{B on 1st})$
- $P(B|B)$ here means $P(\text{B on 2nd}|\text{B on 1st})$

16. In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.

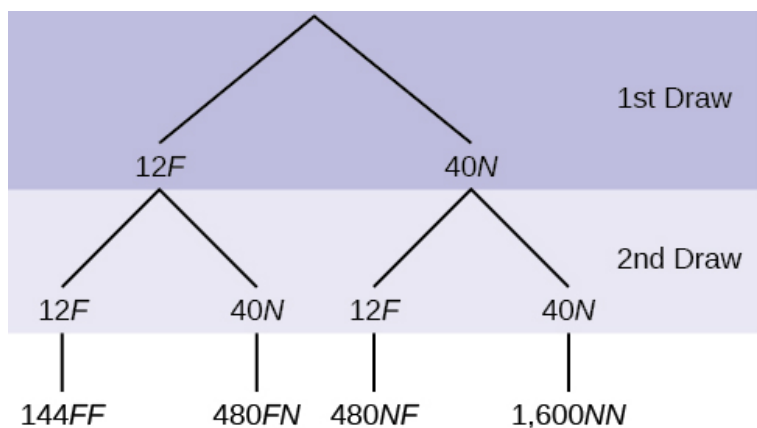


Figure 3.39

17. In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

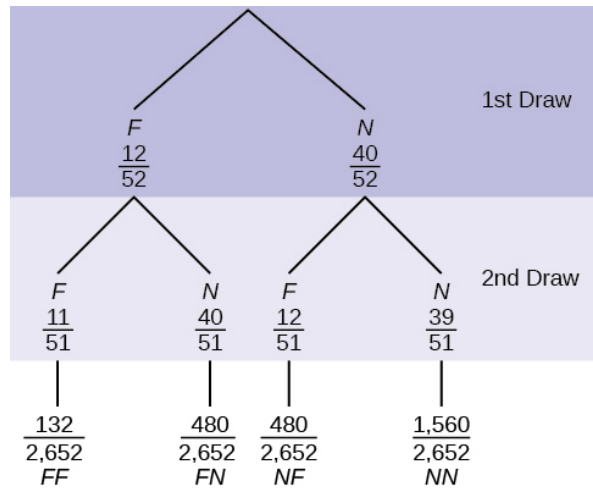


Figure 3.40

- Find $P(FN \text{ OR } NF)$.
- Find $P(N|F)$.
- Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
- Find $P(\text{at least one face card})$.
Hint: "At least one face card" means one or two face cards.

18. A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.

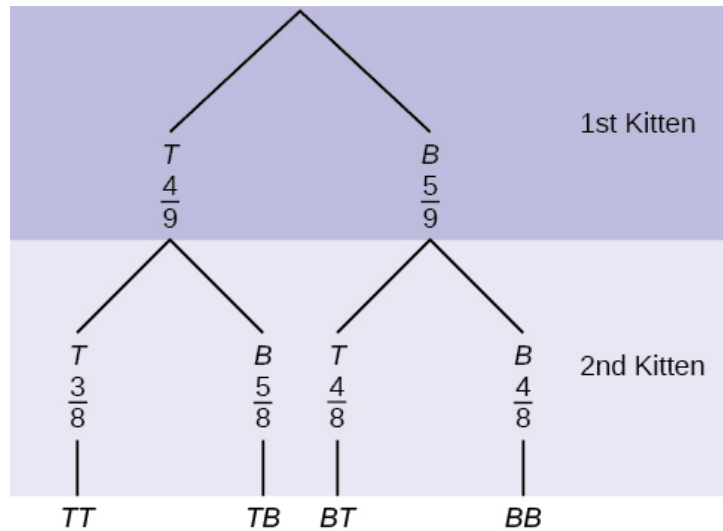


Figure 3.41

1. What is the probability that both kittens are tabby?

- a. $\left(\frac{1}{2}\right) \left(\frac{1}{2}\right)$
- b. $\left(\frac{4}{9}\right) \left(\frac{4}{9}\right)$
- c. $\left(\frac{4}{9}\right) \left(\frac{3}{8}\right)$
- d. $\left(\frac{4}{9}\right) \left(\frac{5}{8}\right)$

2. What is the probability that one kitten of each coloring is selected?

- a. $\left(\frac{4}{9}\right) \left(\frac{5}{9}\right)$
- b. $\left(\frac{4}{9}\right) \left(\frac{5}{8}\right)$
- c. $\left(\frac{4}{9}\right) \left(\frac{3}{8}\right) + \left(\frac{5}{9}\right) \left(\frac{4}{9}\right)$
- d. $\left(\frac{4}{9}\right) \left(\frac{5}{8}\right) + \left(\frac{5}{9}\right) \left(\frac{4}{8}\right)$

3. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?

4. What is the probability of choosing two kittens of the same color?

Solutions: 1. c, 2. d, 3. $\frac{4}{8}$, 4. $\frac{32}{72}$

19. Suppose there are four red balls and three yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

20. Flip two fair coins. Let A = tails on the first coin. Let B = tails on the second coin. Then $A = \{TT, TH\}$ and $B = \{TT, HT\}$. Therefore, $A \text{ AND } B = \{TT\}$. $A \text{ OR } B = \{TH, TT, HT\}$.

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$. The outcome HH is in NEITHER A NOR B . Draw a Venn Diagram.

Solution:

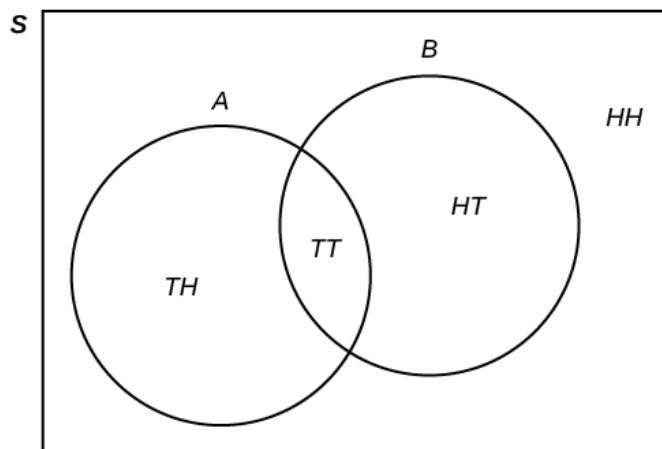


Figure 3.42

21. Roll a fair, six-sided die. Let A = a prime number of dots is rolled. Let B = an odd number of dots is rolled. Then $A = \{2, 3, 5\}$ and $B = \{1, 3, 5\}$. Therefore, $A \text{ AND } B = \{3, 5\}$. $A \text{ OR } B = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

22. Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event $C = \{\text{green, blue, purple}\}$ and event $P = \{\text{red, yellow, blue}\}$. Then $C \text{ AND } P = \{\text{blue}\}$ and $C \text{ OR } P = \{\text{green, blue, purple, red, yellow}\}$. Draw a Venn diagram representing this situation.

23. Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let W = works a second job and S = spouse also works.

24. A person with type O blood and a negative Rh factor (Rh^-) can donate blood to any person with any blood

type. Four percent of African Americans have type O blood and a negative RH factor, 5–10% of African Americans have the Rh- factor, and 51% have type O blood.^{5 6}

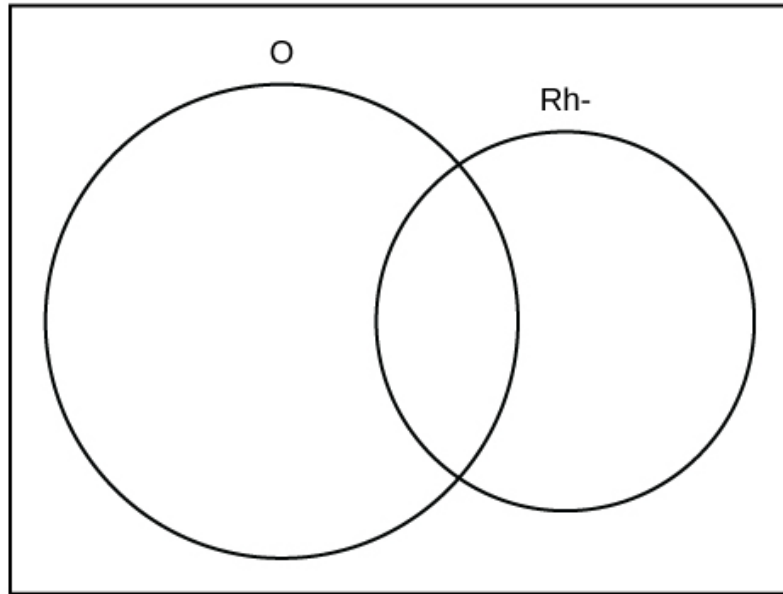


Figure 3.43

The “O” circle represents the African Americans with type O blood. The “Rh-” oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let O = African American with Type O blood and R = African American with Rh- factor.

- $P(O) =$ _____
- $P(R) =$ _____
- $P(O \text{ AND } R) =$ _____
- $P(O \text{ OR } R) =$ _____
- In the Venn Diagram, describe the overlapping area using a complete sentence.
- In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

5. “Blood Types.” American Red Cross, 2013. Available online at <http://www.redcrossblood.org/learn-about-blood/bloodtypes> (accessed May 3, 2013).

6. Samuel, T. M. “Strange Facts about RH Negative Blood.” Healthfully, 2017. Available online at <https://healthfully.com/strange-rh-negative-blood-5552003.html> (accessed January 26, 2021).

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

25. In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

- Draw a Venn diagram representing the situation.
 - Find the probability that the customer buys either a novel or anon-fiction book.
 - In the Venn diagram, describe the overlapping area using a complete sentence.
 - Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.
-

26. The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: C = a man develops cancer in his lifetime; P = man has at least one false positive. Construct a tree diagram of the situation.

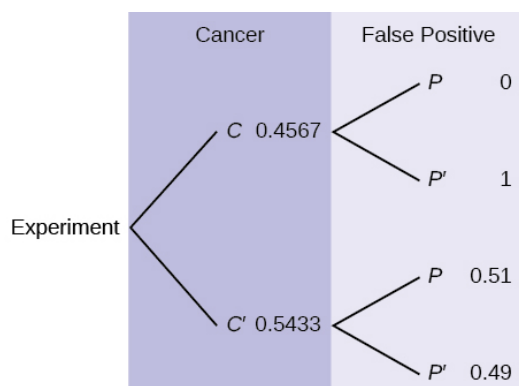


Figure 3.44

27. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (R), four yellow (Y) and five blue (B) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where H is heads and T is tails.

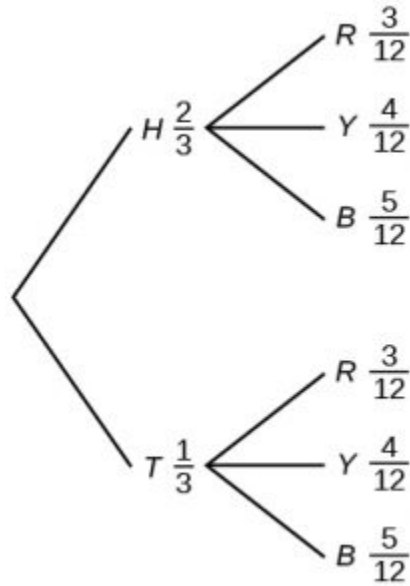


Figure 3.45

Find $P(\text{tossing a Head on the coin AND a Red bead})$

- a. $\frac{2}{3}$
- b. $\frac{5}{15}$
- c. $\frac{6}{36}$
- d. $\frac{9}{36}$

Find $P(\text{Blue bead})$.

- a. $\frac{15}{36}$
- b. $\frac{10}{36}$
- c. $\frac{12}{36}$
- d. $\frac{6}{36}$

28. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

- a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
- b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first

selection? Explain.

- For each complete path through the tree, write the event it represents and find the probabilities.
- Let S be the event that both cookies selected were the same flavor. Find $P(S)$.
- Let T be the event that the cookies selected were different flavors. Find $P(T)$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
- Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.

29. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

Suppose that you randomly draw two cards, one at a time, **with replacement**.

Let G_1 = first card is green

Let G_2 = second card is green

- Draw a tree diagram of the situation.
- Find $P(G_1 \text{ AND } G_2)$.
- Find $P(\text{at least one green})$.
- Find $P(G_2|G_1)$.
- Are G_2 and G_1 independent events? Explain why or why not.

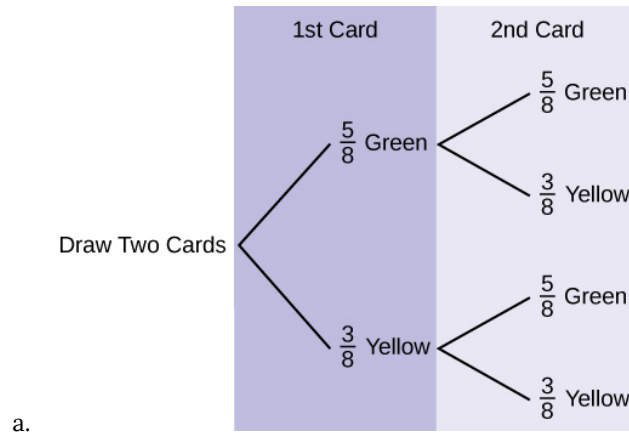


Figure 3.46

- $P(GG) = \left(\frac{5}{8}\right) \left(\frac{5}{8}\right) = \frac{25}{64}$
- $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
- $P(G|G) = \frac{5}{8}$
- Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Suppose that you randomly draw two cards, one at a time, **without replacement**.

G_1 = first card is green
 G_2 = second card is green

- Draw a tree diagram of the situation.
- Find $P(G_1 \text{ AND } G_2)$.
- Find $P(\text{at least one green})$.
- Find $P(G_2|G_1)$.
- Are G_2 and G_1 independent events? Explain why or why not.

30. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.⁸

Complete the following.

- Construct a table or a tree diagram of the situation.
- Find $P(\text{driver is female})$.
- Find $P(\text{driver is age 65 or over}|\text{driver is female})$.
- Find $P(\text{driver is age 65 or over AND female})$.
- In words, explain the difference between the probabilities in part c and part d.
- Find $P(\text{driver is age 65 or over})$.
- Are being age 65 or over and being female mutually exclusive events? How do you know?

a. **Figure 3.47**

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

- $P(F) = 0.486$
- $P(>64|F) = 0.1361$
- $P(>64 \text{ and } F) = P(F) P(>64|F) = (0.486)(0.1361) = 0.0661$
- $P(>64|F)$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } F)$ is the percentage of drivers who are female and 65 or older.
- $P(>64) = P(>64 \text{ and } F) + P(>64 \text{ and } M) = 0.1356$
- No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } F) = 0.0661$.

8. Data from the Federal Highway Administration, part of the United States Department of Transportation.

Suppose that 10,000 U.S. licensed drivers are randomly selected.

- How many would you expect to be male?
- Using the table or tree diagram, construct a contingency table of gender versus age group.
- Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

31. Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.⁹

- Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
- Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
- Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
- Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

a. **Figure 3.48**

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

- If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$.
- Make the same assumptions as in (b) we have: $(0.7708)(1,000) = 771$
- $(0.1332)(1,000) = 133$

32. When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

- Based on the given data, find $P(H)$ and $P(T)$.
- Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.

9. Data from the Federal Highway Administration, part of the United States Department of Transportation.

- c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
- d. Use the tree to find the probability of obtaining at least one head.

33. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

Figure 3.49: AIDS statistics * includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	-----
Male	2,146	463	60	135	-----
Totals	-----	-----	-----	-----	-----

Suppose a person with AIDS in Santa Clara County is randomly selected.

- a. Find P(Person is female).
- b. Find P(Person has a risk factor heterosexual contact).
- c. Find P(Person is female OR has a risk factor of IV drug user).
- d. Find P(Person is female AND has a risk factor of homosexual/bisexual).
- e. Find P(Person is male AND has a risk factor of IV drug user).
- f. Find P(Person is female GIVEN person got the disease from heterosexual contact).
- g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

The completed contingency table is as follows:

Figure 3.50: AIDS statistics solution * includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	255
Male	2,146	463	60	135	2,804
Totals	2,146	533	196	184	3,059

- a. $\frac{255}{3059}$
- b. $\frac{196}{3059}$
- c. $\frac{718}{3059}$
- d. 0
- e. $\frac{463}{3059}$
- f. $\frac{136}{196}$

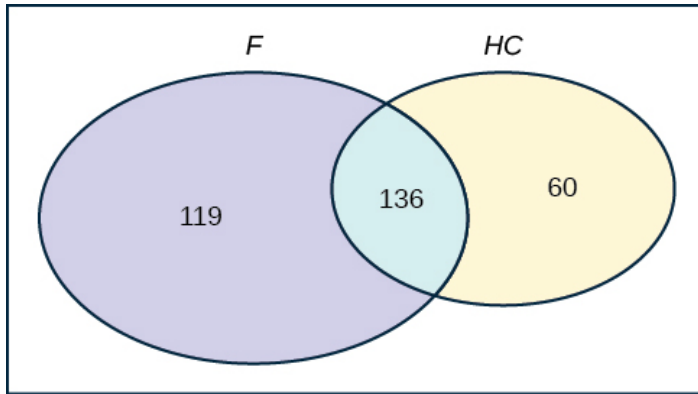


Figure 3.51

g.

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

- Find $P(\text{Person is female})$.
- Find $P(\text{Person obtained the disease through heterosexual contact})$.
- Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$
- Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

34. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.¹¹

Figure 3.52

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

- What is the probability that a randomly selected senator has an “Other” affiliation?

- 0

11. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

- b. What is the probability that a randomly selected senator is up for reelection in November 2016?
- c. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?
- $\frac{10}{67}$
- d. What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?
- e. Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?
- $\frac{10}{34}$
- f. Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?
- g. The events “Republican” and “Up for reelection in 2016” are _____
- a. mutually exclusive.
 - b. independent.
 - c. both mutually exclusive and independent.
 - d. neither mutually exclusive nor independent.
- h. The events “Other” and “Up for reelection in November 2016” are _____
- a. mutually exclusive.
 - b. independent.
 - c. both mutually exclusive and independent.
 - d. neither mutually exclusive nor independent.
-

35. The figure below gives the number of suicides estimated in the U.S. for a recent year by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Figure 3.53

Race and Sex	1-14	15-24	25-64	over 64	TOTALS
white, male	210	3,360	13,610		22,050
white, female	80	580	3,380		4,930
black, male	10	460	1,060		1,670
black, female	0	40	270		330
all others					
TOTALS	310	4,650	18,780		29,760

Do not include “all others” for parts f and g.

- Fill in the column for the suicides for individuals over age 64.
- Fill in the row for all other races.
- Find the probability that a randomly selected individual was a white male.
- Find the probability that a randomly selected individual was a black female.
- Find the probability that a randomly selected individual was black
- Find the probability that a randomly selected individual was a black or white male.
- Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.

a. Figure 3.54

Race and Sex	1-14	15-24	25-64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b. Figure 3.55

Race and Sex	1-14	15-24	25-64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c.
$$\frac{22,050}{29,760}$$

- d. $\frac{330}{29,760}$
- e. $\frac{2,000}{29,760}$
- f. $\frac{23720}{(29760-780)} = \frac{23720}{28980}$
- g. $\frac{5010}{(6020-100)} = \frac{5010}{5920}$

36. The table of data obtained from WWW.BASEBALL-ALMANAC.COM shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.¹²

Figure 3.56

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Find $P(\text{hit was made by Babe Ruth})$.

- a. $\frac{1518}{2873}$
- b. $\frac{12351}{583}$
- c. $\frac{12351}{4189}$
- d. $\frac{4189}{12351}$

Find $P(\text{hit was made by Ty Cobb} | \text{The hit was a Home Run})$.

- a. $\frac{4189}{12351}$
- b. $\frac{114}{1720}$
- c. $\frac{4189}{114}$
- d. $\frac{114}{12351}$

37. The figure below identifies a group of children by one of four hair colors, and by type of hair.

12. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Figure 3.57

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

- Complete the table.
- What is the probability that a randomly selected child will have wavy hair?
- What is the probability that a randomly selected child will have either brown or blond hair?
- What is the probability that a randomly selected child will have wavy brown hair?
- What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
- If B is the event of a child having brown hair, find the probability of the complement of B.
- In words, what does the complement of B represent?

38. In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data were compiled into the following table.¹³

Figure 3.58

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

- Find the probability that his shirt number is from 1 to 33.
- Find the probability that he weighs at most 210 pounds.
- Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
- Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
- Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

a. $\frac{26}{106}$

b. $\frac{33}{106}$

c. $\frac{21}{106}$

d. $\left(\frac{26}{106}\right) + \left(\frac{33}{106}\right) - \left(\frac{21}{106}\right) = \left(\frac{38}{106}\right)$

e. $\frac{21}{33}$

3.3 Compound Events

1. Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(C) = 0.75$. D = the event Helen makes the second shot. $P(D) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

2. A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?

- $\frac{28}{150}$

b. What is the probability that the member practices four times a week?

- $\frac{80}{150}$

c. What is the probability that the member is an advanced swimmer and practices four times a week?

- $\frac{40}{150}$

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

- $P(\text{advanced AND intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

- No, these are not independent events.
- $P(\text{novice AND practices four times per week}) = 0.0667$
- $P(\text{novice})P(\text{practices four times per week}) = 0.0996$

- $0.0667 \neq 0.0996$

3. A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

4. A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

- Find $P(B \text{ AND } D)$.
- Find $P(B \text{ OR } D)$.

5. Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

- $P(B) = 0.143$; $P(N) = 0.85$

b. Given that the woman has breast cancer, what is the probability that she tests negative?

- $P(N|B) = 0.02$

c. What is the probability that the woman has breast cancer AND tests negative?

- $P(B \text{ AND } N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$

d. What is the probability that the woman has breast cancer or tests negative?

- $P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = 0.143 + 0.85 - 0.0029 = 0.9901$

e. Are having breast cancer and testing negative independent events?

- No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

f. Are having breast cancer and testing negative mutually exclusive?

- No. $P(B \text{ AND } N) = 0.0029$. For B and N to be mutually exclusive, $P(B \text{ AND } N)$ must be zero.
-

6. A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

7. Refer to the information in Question 6. P = tests positive.

- Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
- What is the probability that a woman develops breast cancer and tests positive. Find $P(B \text{ AND } P) = P(P|B)P(B)$.
- What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.
- What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

8. A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

- Find $P(B')$.
 - Find $P(D \text{ AND } B)$.
 - Find $P(B|D)$.
 - Find $P(D \text{ AND } B')$.
 - Find $P(D|B')$.
-

9. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

a. Find $P(C)$.

b. Find $P(L)$.

- 0.376

c. Find $P(C|L)$.

d. In words, what is $C|L$?

- $C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

e. Find $P(L \text{ AND } C)$.

f. In words, what is $L \text{ AND } C$?

- $L \text{ AND } C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

g. Are L and C independent events? Show why or why not.

h. Find $P(L \text{ OR } C)$.

- 0.6492

i. In words, what is $L \text{ OR } C$?

j. Are L and C mutually exclusive events? Show why or why not.

- No, because $P(L \text{ AND } C)$ does not equal 0.

10. On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either

very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.¹⁴

In this problem, let:

- C = California registered voters who support same-sex marriage.
 - B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
 - A = California registered voters who are 18 to 39 years old.
- a. Find $P(C)$.
 - b. Find $P(B)$.
 - c. Find $P(C|A)$.
 - d. Find $P(B|C)$.
 - e. In words, what is $C|A$?
 - f. In words, what is $B|C$?
 - g. Find $P(C \text{ AND } B)$.
 - h. In words, what is $C \text{ AND } B$?
 - i. Find $P(C \text{ OR } B)$.
 - j. Are C and B mutually exclusive events? Show why or why not.

11. After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval.¹⁵

These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
 - In mid-2011, 57 percent of the population approved of his actions.
 - In late 2011, the percentage of popular approval was measured at 42 percent.
- a. What is the sample size for this study?
 - b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
 - c. How many people polled responded that they approved of Mayor Ford in late 2011?

14. DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at <https://web.archive.org/web/20130512064934/http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf> (accessed May 12, 2013).

15. Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

- d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
- e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

- a. The Forum Research surveyed 1,046 Torontonians.
- b. 58%
- c. 42% of 1,046 = 439 (rounding to the nearest integer)
- d. 0.57
- e. 0.60.

12. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.¹⁶

00	3	6	9	12	15	18	21	24	27	30	33	36	12 to 1
0	2	5	8	11	14	17	20	23	26	29	32	35	12 to 1
	1	4	7	10	13	16	19	22	25	28	31	34	2 to 1
1st Dozen				2nd Dozen				3rd Dozen					
1 to 18		EVEN						ODD		19 to 36			

Figure 3.59

- a. List the sample space of the 38 possible outcomes in roulette.
- b. You bet on red. Find $P(\text{red})$.
- c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
- d. You bet on an even number. Find $P(\text{even number})$.
- e. Is getting an odd number the complement of getting an even number? Why?
- f. Find two mutually exclusive events.
- g. Are the events Even and 1st Dozen independent?

Compute the probability of winning the following types of bets:

16. "Roulette." Wikipedia. Available online at <http://en.wikipedia.org/wiki/Roulette> (accessed May 2, 2013).

- Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
- Betting on three numbers in a line, as in 1-2-3
- Betting on one number
- Betting on four numbers that touch each other to form a square, as in 10-11-13-14
- Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
- Betting on 0-00-1-2-3
- Betting on 0-1-2; or 0-00-2; or 00-2-3

- $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$
- $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
- $P(\text{Betttng on one number}) = \frac{1}{38}$
- $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$
- $P(\text{Betting on two number that touch each other on the table}) = \frac{2}{38}$
- $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$
- $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

Compute the probability of winning the following types of bets:

- Betting on a color
- Betting on one of the dozen groups
- Betting on the range of numbers from 1 to 18
- Betting on the range of numbers 19-36
- Betting on one of the columns
- Betting on an even or odd number (excluding zero)

13. Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

- List the sample space.
- $P(G) = \underline{\hspace{2cm}}$
- $P(G|E) = \underline{\hspace{2cm}}$
- $P(G \text{ AND } E) = \underline{\hspace{2cm}}$
- $P(G \text{ OR } E) = \underline{\hspace{2cm}}$
- Are G and E mutually exclusive? Justify your answer numerically.

- $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$

- b. $\frac{1}{6}$
- c. $\frac{1}{3}$
- d. $\frac{1}{2}$
- e. $\frac{2}{3}$
- f. No, because $P(G \text{ AND } E)$ does not equal 0.

14. Roll two fair dice separately. Each die has six faces.

- a. List the sample space.
- b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
- c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
- d. In words, explain what " $P(A|B)$ " represents. Find $P(A|B)$.
- e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
- f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

15. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

- a. List the sample space.
- b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
- c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
- d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

NOTE: The coin toss is independent of the card picked first.

- a. $\{(G,H) (G,T) (B,H) (B,T) (R,H) (R,T)\}$
- b. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right) \left(\frac{1}{2}\right) = \frac{3}{20}$
- c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \text{ AND } B) = 0$
- d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \text{ AND } C) = P(A) = \frac{3}{20}$

16. An experiment consists of first rolling a die and then tossing a coin.

- a. List the sample space.
 - b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
 - c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
-

17. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

- a. List the sample space.
 - b. Let A be the event that there are at least two tails. Find $P(A)$.
 - c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.
- a. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
 - b. $\frac{4}{8}$
 - c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \text{ AND } B) = 0$.
-

18. Consider the following scenario:

Let $P(C) = 0.4$.

Let $P(D) = 0.5$.

Let $P(C|D) = 0.6$.

- a. Find $P(C \text{ AND } D)$.
 - b. Are C and D mutually exclusive? Why or why not?
 - c. Are C and D independent events? Why or why not?
 - d. Find $P(C \text{ OR } D)$.
 - e. Find $P(D|C)$.
-

19. Y and Z are independent events.

- a. Rewrite the basic Addition Rule $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y \text{ AND } Z)$ using the information that Y and Z are independent events.
 - b. Use the rewritten rule to find $P(Z)$ if $P(Y \text{ OR } Z) = 0.71$ and $P(Y) = 0.42$.
- a. If Y and Z are independent, then $P(Y \text{ AND } Z) = P(Y)P(Z)$, so $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y)P(Z)$.
 - b. 0.5
-

20. G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

- Explain why the following statement MUST be false: $P(H|G) = 0.4$.
 - Find $P(H \text{ OR } G)$.
 - Are G and H independent or dependent events? Explain in a complete sentence.
-

21. Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.¹⁷

Let: E = speaks English at home

E' = speaks another language at home

S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Figure 3.60

Probability Statements	Answers
a. $P(E')$ =	i. 0.8043
b. $P(E)$ =	ii. 0.623
c. $P(S \text{ and } E')$ =	iii. 0.1957
d. $P(S E')$ =	iv. 0.1219

22. 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

- What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
 - In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
 - Are G and F independent or dependent events? Justify your answer numerically and also explain why.
 - Are G and F mutually exclusive events? Justify your answer numerically and explain why.
-

17. Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at <http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf> (accessed May 2, 2013).

23. Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with \$10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

- Write a probability statement for the overall percent of money returned.
 - Write a probability statement for the percent of money returned out of the economics classes.
 - Write a probability statement for the percent of money returned out of the other classes.
 - Is money being returned independent of the class? Justify your answer numerically and explain it.
 - Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.
- $P(R) = 0.44$
 - $P(R|E) = 0.56$
 - $P(R|O) = 0.31$
 - No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
 - No, this study definitely does not support that notion; in fact, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

24. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.¹⁸

Figure 3.61

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are “the hit being made by Hank Aaron” and “the hit being a double” independent events?

- Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$

18. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

- b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$
- c. No, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
- d. Yes, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) = P(\text{hit is a double})$

25. United Blood Services¹⁹ is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

- a. Find the probability that a person has both type O blood and the Rh- factor.
 - b. Find the probability that a person does NOT have both type O blood and the Rh- factor.
- a. $P(\text{type O OR Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O AND Rh-})$
 $0.52 = 0.43 + 0.15 - P(\text{type O AND Rh-});$ solve to find $P(\text{type O AND Rh-}) = 0.06$
 6% of people have type O, Rh- blood
- b. $P(\text{NOT}(\text{type O AND Rh-})) = 1 - P(\text{type O AND Rh-}) = 1 - 0.06 = 0.94$
 94% of people do not have type O, Rh- blood

26. At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

- a. Find the probability that a course has a final exam or a research project.
- b. Find the probability that a course has NEITHER of these two requirements.

27. In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

- a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
 - b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).
- a. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
- b. $P(C \text{ OR } N) = P(C) + P(N) - P(C \text{ AND } N) = 0.36 + 0.12 - 0.08 = 0.40$

19. "Human Blood Types." United Blood Services, 2011. Available online at <https://web.archive.org/web/20130807103902/http://www.unitedbloodservices.org/learnMore.aspx> (accessed August 22, 2013).

c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \text{ OR } N) = 1 - 0.40 = 0.60$

28. A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student

- Find $P(D \text{ AND } E)$.
- Find $P(E|D)$.
- Find $P(D \text{ OR } E)$.
- Using an appropriate test, show whether D and E are independent.
- Using an appropriate test, show whether D and E are mutually exclusive.

References

Image References

Figure 3.14: Figure 3.12 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/3-homework>

Figure 3.17: Figure 3.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/3-practice>

Figure 3.18: Figure 3.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/3-homework>

Figure 3.36: Figure 3.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.37: Figure 3.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.38: Figure from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.39: Figure 3.12 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.40: Figure 3.14 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.41: Figure from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.42: Figure 3.16 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.43: Figure 3.18 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-5-tree-and-venn-diagrams>

Figure 3.44: Figure 3.15 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/3-solutions#element-750-solution>

Figure 3.45: Figure 3.6.1 from LibreTexts Math 40: Statistics and Probability (2020) (CC BY 4.0). Retrieved from [https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_and_Probability/04%3A_Probability_and_Counting/4.E%3A_Probability_Topics_\(Optional_Exercises\)](https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_and_Probability/04%3A_Probability_and_Counting/4.E%3A_Probability_Topics_(Optional_Exercises))

Figure 3.46: Figure 3.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/3-solutions>

Figure 3.51: Figure 3.17 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/3-solutions#element-456-solution>

Figure 3.59: Figure 3.4.1 from LibreTexts Introductory Statistics (2020) (CC BY 4.0). Retrieved from [https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Exercises_\(Introductory_Statistics\)/Exercises%3A_OpenStax/03.E%3A_Probability_Topics_\(Exercises\)](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Exercises_(Introductory_Statistics)/Exercises%3A_OpenStax/03.E%3A_Probability_Topics_(Exercises))

Text

“Countries List by Continent.” Worldatlas, 2013. Available online at <http://www.worldatlas.com/cntycont.htm> (accessed May 2, 2013).

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. <http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx> (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

“Blood Types.” American Red Cross, 2013. Available online at <http://www.redcrossblood.org/learn-about-blood/blood-types> (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at <http://www.nejm.org/doi/full/10.1056/NEJMoa033250> (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at <http://www.disastercenter.com/crime/> (accessed May 2, 2013).

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at <http://lib.stat.cmu.edu/DASL> (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at <http://data.worldbank.org/topic/environment> (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 12, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News_Archives/News_Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at <http://en.wikipedia.org/wiki/Roulette> (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at <http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf> (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at <http://www.ropercenter.uconn.edu/> (accessed May 2, 2013).

Data from Field Research Corporation. Available online at <https://web.archive.org/web/20130512064934/http://www.field.com/fieldpollonline> (accessed May 12, 2013).

Roulette Image: film8ker/wikibooks

CHAPTER 4: DISCRETE RANDOM VARIABLES

4.1 Introduction to Discrete Random Variables and Notation

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions
- Calculate and interpret probabilities, expected values, and standard deviations of general random variables
- Recognize the binomial probability distribution and apply it appropriately



Figure 4.1: Lightning Strike. You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A random variable describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variables

Random variables are **probability models** quantifying situations. Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If X is a random variable, then X is written in words, and x is given as a number.

There are both continuous and discrete random variables. We will begin with **discrete RVs** and revisit **continuous RVs** in the future.

Discrete Random Variables

We have seen the word discrete before associated with types of data. Discrete means we have a countable number of outcomes. So a **discrete random variable** is a RV that models a process or experiment that produces discrete data. Consider the following example of a discrete random variable:

Let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is $TTT, THH, HTH, HHT, HTT, THT, TTH, HHH$. Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Example

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X =

the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$.

$P(x)$ = probability that X takes on a value x .

Figure 4.2: Newborn Baby Crying

x	$P(x)$
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

Is this a valid discrete probability distribution?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=156#h5p-103>

Your turn!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$. $P(x)$ = the probability that X takes on value x . Is this a discrete probability distribution function (two reasons)?

Figure 4.3: Post-Op Patients

X	P(x)
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Characteristics and Notation

The distribution of a discrete random variable is often pictured in a table, but may also be represented by a graph or formula. Two main characteristics it should exhibit are:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

The **probability mass function (PMF)** of a DRV tells you the probability of a single value. Notation-wise this means $P(X = x)$. This is also sometimes (erroneously) called probability distribution function (PDF).

The **cumulative distribution function (CDF)** of a DRV tells you the probability of being less than or equal to a value. Notation-wise this means $P(X \leq x)$.

A probability distribution function is a pattern. You try to fit a probability problem into a pattern or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Example

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

a. Let X = the number of days Nancy _____ .



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=156#h5p-104>

b. X takes on what values?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=156#h5p-105>

c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one below. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Figure 4.4:
Blank PDF

x	$P(x)$
0	
1	
2	
3	



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=156#h5p-106>

d. Construct the cumulative probability distribution function



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=156#h5p-107>

Your turn!

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Image Credits

Figure 4.1: Michael D (2018). “Storm at dawn.” Public domain. Retrieved from <https://unsplash.com/photos/2cDIzRnVq0Q>

4.2 Measures of General DRVs

Once we know how to work with Discrete Random Variables we may be interested in some other measures such as the mean, variance, and standard deviation. The ideas here are slightly different than we have seen before within our new context of Random Variables.

The Expected Value (Mean) of a Discrete Random Variable

Recall the Law of Large Numbers which states as the number of trials in a probability experiment increases our results become closer to what we expect. When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This “long-term average” is known as the mean or **expected value** of the random variable and is denoted by the Greek letter μ or $E[X]$ in the context of random variables. In other words, after conducting many trials of an experiment, you would expect this average value.

To find the expected value or long term average we simply multiply each value of the random variable by its probability and add the products.

$$\text{Mean or Expected Value: } \mu = \sum_{x \in X} xP(x)$$

Example

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ , of the number of days per week the men's soccer team plays soccer.

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x \cdot P(x)$. In this column, you will multiply each x value by its probability.

Figure 4.5: Expected Value Table. This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	$P(x)$	$x \cdot P(x)$
0	0.2	$(0)(0.2) = 0$
1	0.5	$(1)(0.5) = 0.5$
2	0.3	$(2)(0.3) = 0.6$

What is the expected value?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-108>

Your turn!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

**Figure 4.3 (repeat):
Post-Op Patients**

x	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

The Variance and Standard Deviation of a Discrete Random Variable

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root.

Finding the variance, σ^2 or $V[X]$, and standard deviation, σ or $SD[X]$ of a random variable starts similar to what we have seen before but differs at step 4:

1. Find the mean
2. Subtract the mean from each value of x to get your deviations
3. Square each deviation
4. Multiply each squared deviation by its probability, $P(x)$
5. Sum each of the products

At this point you now have the variance then can of course take the square root of the variance to get your standard deviation. The formula looks like this:

$$\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$$

Example

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

Figure 4.6: Newborn Baby Crying. You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0) \left(\frac{2}{50} \right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x = 1) = \left(\frac{11}{50} \right)$	$(1) \left(\frac{11}{50} \right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x = 2) = \frac{23}{50}$	$(2) \left(\frac{23}{50} \right) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x = 3) = \frac{9}{50}$	$(3) \left(\frac{9}{50} \right) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x = 4) = \frac{4}{50}$	$(4) \left(\frac{4}{50} \right) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x = 5) = \frac{1}{50}$	$(5) \left(\frac{1}{50} \right) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

a. Add the values in the third column of the table to find the expected value of X .



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-109>

b. Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x , multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$).

c. Add the values in the fourth column of the table:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-110>

d. The standard deviation of X is the square root of this sum.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-111>

e. The mean, μ , of a discrete probability function is the expected value.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-112>

f. The standard deviation, Σ , of the PDF is the square root of the variance.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=158#h5p-113>

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

Your turn!

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X .

Note on Calculations

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For many special cases of probability distributions, there are short-cut formulas for calculating μ , σ , and associated probabilities. We will see some of these in the future.

4.3 The Binomial Distribution

We have seen how to deal with general **discrete random variables**, but there are also special cases of DRVs. If we can identify them, they can provide us some insight and shortcuts. The first of these is the Binomial Distribution.

The Binomial Setting

There are three characteristics of a **binomial experiment**.

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called “success” and “failure,” for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are **independent** and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p , of a success and probability, q , of a failure remain the same.

For example: At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A “success” could be defined as an individual who withdrew. The random variable X = the number of students who withdraw from the randomly selected elementary physics class.

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli trials.

For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same. This situation meets the Binomial requirements.

The following example illustrates a problem that is not binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn without replacement. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the

second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Example

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

a. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-114>

b. If we are interested in the number of students who do their homework on time, then how do we define X?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-115>

c. What values does x take on?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-116>

d. What is a “failure,” in words?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-117>

e. If $p + q = 1$, then what is q ?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-118>

f. The words “at least” translate as what kind of inequality for the probability question $P(x \text{ ____ } 40)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-119>

Your turn!

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Can we use the binomial here?

Notation for the Binomial

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X counts the number of successes obtained in the n independent trials.

$$X \sim B(n, p)$$

Read this as “ X is a random variable with a binomial distribution.” The parameters are n and p : n = number of trials, p = probability of a success on each trial.

Since the Binomial counts the number of successes, x , in n trials, the range of values for a binomial random variable could be anything from 0 to n ($x=0,1,2,\dots, n$).

Binomial Probability Function

Once we have decided we can use the binomial for a given situation, we can use the binomial probability function to find the probability of a specific number of successes, $P(X=x)$. The binomial **PMF** is made up of two parts:

First, we need to find out how many different ways we can get x successes in n trials. To do this we can use the “Choose” function, also called the binomial coefficient, written as:

$$nC_x = C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Note: The the ! mark is the factorial operator.

The next part gives us the probability of a single one of those ways to get x successes in n trials. We can do this by using our independent multiplication rule. We multiply the probability of success (p) raised to the number of successes (x) by the probability of failure ($q=1-p$) raised to the number of failures ($n-x$).

$$p^x q^{(n-x)}$$

Since we know each of these ways are equally likely and how many ways are possible we can now put the two pieces together. We multiply the probability of one way by how many we have to give us our overall probability of x successes in n trials.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

Unfortunately the binomial does not have a nice form of **CDF**, but it is simply the sum of PDFs up until that point. Consider the following example to demonstrate this point.

Example

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. 20 adult workers are randomly selected.

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$, $p = 0.41$, and $q = 1 - 0.41 = 0.59$. $X \sim B(20, 0.41)$

The y-axis contains the probability of x , where X = the number of workers who have only a high school diploma.

The graph of $X \sim B(20, 0.41)$ is as follows:

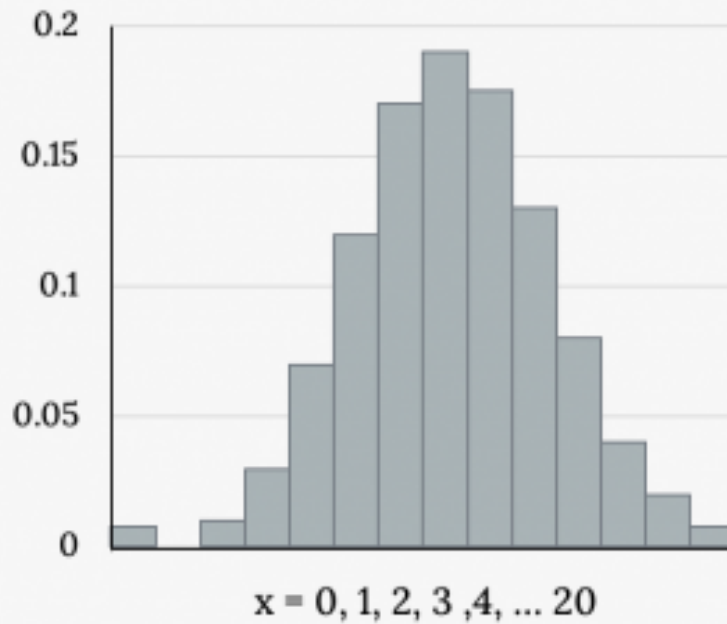


Figure 4.7: Workers With Diplomas

Find the probability that:

(a) Exactly 12 of them have a high school diploma



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-120>

(b) At most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-121>

Your turn!

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find:

- (a) The probability that exactly 14 of them participate in a community volunteer program outside of school. First try plugging in to the binomial formula by hand, then check yourself with technology.
- (b) The probability that exactly 14 of them participate in a community volunteer program outside of school. Rely on technology for this cumulative probability.

Measures of the Binomial Distribution

The mean, μ , and variance, σ^2 , for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ , is then $\sigma = \sqrt{npq}$.

Example

In the 2013 Jerry's Artarama art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

1. What values does x take on?
2. What is the probability distribution? Find the following probabilities
 - 2a. the probability that two pages feature signature artists.
 - 2b. the probability that at most six pages feature signature artists
 - 2c. the probability that more than three pages feature signature artists.
3. Using the formulas, calculate the (3a) mean and (3b) standard deviation.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=161#h5p-122>

Your turn!

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

- a. What is the probability distribution for X ?
- b. Use your calculator to find the following probabilities:
 - i. the probability that 25 adults in the sample prefer saving over spending
 - ii. the probability that at most 20 adults prefer saving
 - iii. the probability that more than 30 adults prefer saving

- c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X .

Image References

Figure 4.7: Kindred Grey (2020). “Figure 4.7” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_4.7.png

Chapter 4 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=164#h5p-123>

Section Reviews

4.1 Introduction

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

4.2 Measures of General DRVs

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$

Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$

4.3 The Binomial Distribution

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n .
2. There are only two possible outcomes, called “success” and, “failure” for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

$X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p .

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x = 0, 1, 2, 3, \dots, n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1$

$q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

4.1 Introduction

- Random variable
- Probability model
- Discrete random variable
- Continuous random variable
- Probability mass function (PMF)
- Cumulative distribution function (CDF)

4.2 Measures of General DRVs

- Expected value

4.3 The Binomial Distribution

- Discrete random variable
- Binomial experiment
- Independent
- Bernoulli trial
- Probability mass function
- Cumulative distribution function

Extra Practice

4.1 Introduction

1. A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution. Let X = the number of years a new hire will stay with the company. Let $P(x)$ = the probability that a new hire will stay with the company x years. Complete the figure below using the data provided.

Figure 4.8

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

a. $P(x = 4) =$ _____

- 0.10

b. $P(x \geq 5) =$ _____

- $0.10 + 0.05 = 0.15$

c. On average, how long would you expect a new hire to stay with the company?

- $0 + 0.18 + 0.60 + 0.45 + 0.40 + 0.50 + 0.30 = 2.43$ years

d. What does the column " $P(x)$ " sum to?

- 1
-

2. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

**Figure
4.9**

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

a. Define the random variable X .

- Let X = the number of batches that the baker will sell.

b. What is the probability the baker will sell more than one batch? $P(x > 1) = \underline{\hspace{2cm}}$

- $0.35 + 0.40 + 0.10 = 0.85$

c. What is the probability the baker will sell exactly one batch? $P(x = 1) = \underline{\hspace{2cm}}$

- 0.15

d. On average, how many batches should the baker make?

- $1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45$
-

3. Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

a. Define the random variable X.

- Let X = the number of days Ellen attends practice per week.

b. Construct a probability distribution table for the data.

**Figure
4.10**

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

c. We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

- Each probability is between zero and one, inclusive.
-

4. Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

a. Define the random variable X.

- Let X = the number of events Javier volunteers for each month.

b. What values does x take on?

- 0, 1, 2, 3, 4, 5

c. Construct a PDF table.

**Figure
4.11**

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

d. Find the probability that Javier volunteers for less than three events each month. $P(x < 3) =$ _____

- $0.05 + 0.05 + 0.10 = 0.20$

e. Find the probability that Javier volunteers for at least one event each month. $P(x > 0) =$ _____

- $1 - 0.05 = 0.95$

5. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in below.

**Figure
4.12**

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. In words, define the random variable X .

b. What does it mean that the values zero, one, and two are not included for x in the PDF?

4.2 Measures of General DRV's

1. Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

- To do this problem, set up an expected value table for the amount of money you can profit. Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and -2 dollars.
- To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is
- $\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)(10^{-5}) = 0.00001$.
- Therefore, the probability of winning is 0.00001 and the probability of losing is $1 - 0.00001 = 0.99999$.
- The expected value table is as follows:

Figure 4.13: Add the last column. $-1.99998 + 1 = -0.99998$

	x	$P(x)$	$x \cdot P(x)$
Loss	-2	0.99999	$(-2)(0.99999) = -1.99998$
Profit	100,000	0.00001	$(100000)(0.00001) = 1$

Since -0.99998 is about -1, you would, on average, expect to lose approximately \$1 for each game you play. However, each time you play, you either lose \$2 or profit \$100,000. The \$1 is the average or expected LOSS per game after playing this game over and over.

2. You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay \$1 to play. If you guess the right suit every time, you get your money back and \$256. What is your expected profit of playing the game over the long term?

3. Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead?

a. Define a random variable X .

- X = amount of profit

b. Complete the following expected value table.

Figure 4.14

	x	-----	-----
WIN	10	$\frac{1}{3}$	-----
LOSE	-----	-----	$\frac{-12}{3}$

Figure 4.15

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	-6	$\frac{2}{3}$	$\frac{-12}{3}$

c. What is the expected value, μ ? Do you come out ahead?

- Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

4. Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay \$10. If you land on blue, you don't pay or win anything. If you land on green, you win \$10. Complete the following expected value table.

Figure 4.16

	x	$P(x)$	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

5. Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like the one in Number 4 and calculate the mean μ and standard deviation σ of X .

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

Figure 4.17

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

Figure 4.18: Calculating μ and σ .

x	$P(x)$	$xP(x)$	$(x - \mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum: $\sigma = \sqrt{\frac{18}{36}} \approx 0.701$

6. On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20. Let X = the amount of profit from a bet.¹

$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 21.42\%$

$P(\text{loss}) = P(\text{one moderate earthquake will not occur}) = 100\% - 21.42\%$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X ? Construct a table similar to the ones in 4 and 5 to help you answer these questions.

1. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012.
http://www.worldearthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

Figure 4.19

	x	$P(x)$	$x(Px)$	$(x - \mu)^2 P(x)$
win	50	0.2142	10.71	$[50 - (-5.006)]^2 (0.2142) = 648.0964$
loss	-20	0.7858	-15.716	$[-20 - (-5.006)]^2 (0.7858) = 176.6636$

Mean = Expected Value = $10.71 + (-15.716) = -5.006$.

If you make this bet many times under the same conditions, your long term outcome will be an average loss of \$5.01 per bet.

$$\text{Standard Deviation} = \sqrt{648.0964 + 176.6636} \approx 28.7186$$

7. Complete the expected value table.

Figure 4.20

x	$P(x)$	$x \cdot P(x)$
0	0.2	
1	0.2	
2	0.4	
3	0.2	

8. Find the expected value from the expected value table.

Figure 4.21

x	$P(x)$	$x \cdot P(x)$
2	0.1	$2(0.1) = 0.2$
4	0.3	$4(0.3) = 1.2$
6	0.4	$6(0.4) = 2.4$
8	0.2	$8(0.2) = 1.6$

- $0.2 + 1.2 + 2.4 + 1.6 = 5.4$

Find the standard deviation.

Figure 4.22

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 P(x)$
2	0.1	$2(0.1) = 0.2$	$(2-5.4)^2(0.1) = 1.156$
4	0.3	$4(0.3) = 1.2$	$(4-5.4)^2(0.3) = 0.588$
6	0.4	$6(0.4) = 2.4$	$(6-5.4)^2(0.4) = 0.144$
8	0.2	$8(0.2) = 1.6$	$(8-5.4)^2(0.2) = 1.352$

- $\sigma = 1.156 + 0.588 + 0.144 + 1.352 = 3.24 = 1.8$
-

9. Identify the mistake in the probability distribution table.

Figure 4.23

x	$P(x)$	$x \cdot P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

- The values of $P(x)$ do not sum to one.
-

10. Identify the mistake in the probability distribution table.

Figure 4.24

x	$P(x)$	$x \cdot P(x)$
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

- The values of $xP(x)$ are not correct.
-

11. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

Figure 4.25

x	$P(x)$	$x \cdot P(x)$
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

a. Define the random variable X .

- Let X = the number of years a physics major will spend doing post-graduate research.

b. Define $P(x)$, or the probability of x .

- Let $P(x)$ = the probability that a physics major will do post-graduate research for x years.

c. Find the probability that a physics major will do post-graduate research for four years. $P(x = 4) = \underline{\hspace{2cm}}$

- $1 - 0.35 - 0.20 - 0.15 - 0.10 - 0.05 = 0.15$

d. Find the probability that a physics major will do post-graduate research for at most three years. $P(x \leq 3) = \underline{\hspace{2cm}}$

- $0.35 + 0.20 + 0.15 = 0.70$

e. On average, how many years would you expect a physics major to spend doing post-graduate research?

- $1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 0.35 + 0.40 + 0.45 + 0.60 + 0.50 + 0.30 = 2.6$ years

12. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let $P(x)$ = the probability that a student will study ballet x years.

Complete the figure below using the data provided.

Figure 4.26

x	$P(x)$	$x \cdot P(x)$
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

a. In words, define the random variable X .

- X is the number of years a student studies ballet with the teacher.

b. $P(x = 4) = \underline{\hspace{2cm}}$

- $1 - 0.10 - 0.05 - 0.10 - 0.30 - 0.20 - 0.10 = 0.15$

c. $P(x < 4) = \underline{\hspace{2cm}}$

- $0.10 + 0.05 + 0.10 = 0.25$

d. On average, how many years would you expect a child to study ballet with this teacher?

- $1(0.10) + 2(0.05) + 3(0.10) + 4(0.15) + 5(0.30) + 6(0.20) + 7(0.10) = 4.5$ years

e. What does the column " $P(x)$ " sum to and why?

- The sum of the probabilities sum to one because it is a probability distribution.

f. What does the column " $x \cdot P(x)$ " sum to and why?

- The sum of $xP(x) = 4.5$; it is the mean of the distribution.

13. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win \$30. If it is not a face card, you pay \$2. There are 12 face cards in a deck of 52 cards.

a. What is the expected value of playing the game?

$$-2 \left(\frac{40}{52} \right) + 30 \left(\frac{12}{52} \right) = -1.54 + 6.92 = 5.38$$

b. Should you play the game?

Yes, because there is a positive expected value, and the more you play, the more likely you are to get closer to the expected value.

14. A theater group holds a fund-raiser. It sells 100 raffle tickets for \$5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of \$150.

- What are you interested in here?
- In words, define the random variable X .
- List the values that X may take on.
- Construct a PDF.
- If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

- I am interested in the average profit or loss. Let X = the return from the raffle Win(\$150) or Lose (\$0)
 - $150\left(\frac{1}{100}\right) + 0\left(\frac{99}{100}\right) - 20 = -\18.50
-

15. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win \$6
- If the card is a face card, and the coin lands on Tails, you win \$2
- If the card is not a face card, you lose \$2, no matter what the coin shows.

- Find the expected value for this game (expected net gain or loss).
- Explain what your calculations indicate about your long-term average profits and losses on this game.
- Should you play this game to win money?

The variable of interest is X , or the gain or loss, in dollars.

The face cards jack, queen, and king. There are $(3)(4) = 12$ face cards and $52 - 12 = 40$ cards that are not face cards.

We first need to construct the probability distribution for X . We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Figure 4.27

Card Event	X net gain/loss	P(X)
Face Card and Heads	6	$\left(\frac{12}{52}\right) \left(\frac{1}{2}\right) = \left(\frac{6}{52}\right)$
Face Card and Tails	2	$\left(\frac{12}{52}\right) \left(\frac{1}{2}\right) = \left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	-2	$\left(\frac{40}{52}\right) (1) = \left(\frac{40}{52}\right)$

- Expected value = $\left(6\right) \left(\frac{6}{52}\right) + \left(2\right) \left(\frac{6}{52}\right) + \left(-2\right) \left(\frac{40}{52}\right) = -\frac{32}{52} = -\0.62 (rounded to the nearest cent)
- If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

16. You buy a lottery ticket to a lottery that costs \$10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one \$500 prize, two \$100 prizes, and four \$25 prizes. Find your expected gain or loss.

- Start by writing the probability distribution. X is net gain or loss = prize (if any) less \$10 cost of ticket.
- Expected Value = $(490)\left(\frac{1}{100}\right) + (90)\left(\frac{2}{100}\right) + (15)\left(\frac{4}{100}\right) + (-10)\left(\frac{93}{100}\right) = -\2 . There is an expected loss of \$2 per ticket, on average.

Complete the PDF and answer the questions.

Figure 4.28

x	P(x)	xP(x)
0	0.3	
1	0.2	
2		
3	0.4	

- Find the probability that $x = 2$.
- Find the expected value.

- 0.1
- 1.6

17. Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win \$10. If you roll a four or five, you win \$5. If you roll a one, two, or three, you pay \$6.

- What are you ultimately interested in here (the value of the roll or the money you win)?
- In words, define the Random Variable X .
- List the values that X may take on.
- Construct a PDF.
- Over the long run of playing this game, what are your expected average winnings per game?
- Based on numerical values, should you take the deal? Explain your decision in complete sentences.

- the money won X = the amount of money won or lost \$5, -\$6, \$10
- Expected Value = $(10) \frac{1}{6} + (5) \frac{2}{6} - (6) \frac{3}{6} = 0.33$ Yes, the expected value is 33 cents

18. A venture capitalist, willing to invest \$1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning \$5,000,000 profit, a 30% chance of returning \$1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning \$3,000,000 profit, a 40% chance of returning \$1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning \$6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

- Construct a PDF for each investment.
- Find the expected value for each investment.
- Which is the safest investment? Why do you think so?
- Which is the riskiest investment? Why do you think so?
- Which investment has the highest expected return, on average?

a.

Figure 4.29: Software Company	
x	$P(x)$
5,000,000	0.10
1,000,000	0.30
-1,000,000	0.60

Figure 4.30: Hardware Company	
x	$P(x)$
3,000,000	0.20
1,000,000	0.40
-1,000,00	0.40

Figure 4.31: Biotech Firm	
x	$P(x)$
6,00,000	0.10
0	0.70
-1,000,000	0.20

- b. \$200,000; \$600,000; \$400,000
- c. third investment because it has the lowest probability of loss
- d. first investment because it has the highest probability of loss
- e. second investment

19. Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

Figure 4.32

x	$P(x)$	$xP(x)$
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

- a. Find the probability that a married adult has three children.
 - b. In words, what does the expected value in this example represent?
 - c. Find the expected value.
 - d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?
- 0.2
 - The average number of children married adults have 2.35
 - two to three children

20. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given below.

**Figure
4.33**

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. On average, how many years do you expect it to take for an individual to earn a B.S.?

- 4.85 years

21. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

**Figure
4.34**

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

- Describe the random variable X in words.
- Find the probability that a customer rents three DVDs.
- Find the probability that a customer rents at least four DVDs.
- Find the probability that a customer rents at most two DVDs.

Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

Figure 4.35

x	$P(x)$
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

- At which store is the expected number of DVDs rented per customer higher?
- If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
- If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
- Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Solutions:

X = the number of video rentals per customer 0.12 0.11 0.77 Video To Go (1.82 expected value vs. 1.4 for Entertainment Headquarters) The expected number of videos rented to 300 Video To Go customers is 546. The expected number of videos rented to 420 Entertainment Headquarters customers is 588. Entertainment Headquarters will rent more videos. The standard deviation for the number of videos rented at Video To Go is 1.1609. The standard deviation for the number of videos rented at Entertainment Headquarters is 1.4293. Entertainment Headquarters has more variation.

22. A “friend” offers you the following “deal.” For a \$10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth \$6.
- Eighty of the coupons are for a free gift worth \$8.
- Six of the coupons are for a free gift worth \$12.
- Four of the coupons are for a free gift worth \$40.

Based upon the financial gain or loss over the long run, should you play the game?

- Yes, I expect to come out ahead in money.
- No, I expect to come out behind in money.
- It doesn’t matter. I expect to break even.

- Answer: b

23. Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.²

- What is the average class size assuming each class is filled to capacity?
- Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Define the PDF for X .
- Find the mean of X .
- Find the standard deviation of X .

- Solutions: The average class size is: $30 + 8(60) + 70 + 4(100) = 70$ $P(x=30) = \frac{1}{14}$ $P(x=60) = \frac{8}{14}$ $P(x=70) = \frac{1}{14}$ $P(x=100) = \frac{4}{14}$ Complete the following table to find the mean and standard deviation of X . c Mean of $X = 30 \cdot \frac{1}{14} + 60 \cdot \frac{8}{14} + 70 \cdot \frac{1}{14} + 100 \cdot \frac{4}{14} = 70$ d Standard Deviation of $X = \sqrt{114.2857 + 57.1429 + 0 + 257.1429} = 20.702$

24. In a lottery, there are 250 prizes of \$5, 50 prizes of \$25, and ten prizes of \$100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

Let X = the amount of money to be won on a ticket. The following table shows the PDF for X .

Figure 4.36

x	$P(x)$
0	0.969
5	$\frac{250}{10,000} = 0.025$
25	$\frac{50}{10,000} = 0.005$
100	$\frac{10}{10,000} = 0.001$

- Calculate the expected value of X .

- $0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35$
- A fair price for a ticket is \$0.35. Any price over \$0.35 will enable the lottery to raise money.

2. Class Catalogue at the Florida State University. Available online at <https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy> (accessed May 15, 2013).

4.3 The Binomial Distribution

1. The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a “success” be in this case?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=164#h5p-124>

2. A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=164#h5p-125>

3. A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

4. The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%).³ Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

- What is the probability distribution for X ?
- Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- Find the probability that at most eight people develop pancreatic cancer

3. “What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics> (accessed May 15, 2013).

d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

5. During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots.⁴ Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

- What is the probability distribution for X ?
 - Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
 - Find the probability that DeAndre scored with 60 of these shots.
 - Find the probability that DeAndre scored with more than 50 of these shots.
-

6. A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

7. The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status.⁵ Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

a. In words, define the random variable X .

- X = the number that reply “yes”

b. $X \sim \text{_____}(\text{_____,} \text{_____})$

- $B(8, 0.713)$

4. “NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

5. Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. The American Freshman: National Norms Fall 2011. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

c. What values does the random variable X take on?

- 0, 1, 2, 3, 4, 5, 6, 7, 8

d. Construct the probability distribution function (PDF).

Figure 4.37

x	$P(x)$

e. On average (μ), how many would you expect to answer yes?

- 5.7

f. What is the standard deviation (σ)?

- 1.2795

g. What is the probability that at most five of the freshmen reply “yes”?

- 0.4151

h. What is the probability that at least two of the freshmen reply “yes”?

- 0.9990

8. According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

- 0.2709

9. Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not

just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

a. Define the random variable and list its possible values.

- X = the number of patients calling in claiming to have the flu, who actually have the flu.
- $X = 0, 1, 2, \dots, 25$

b. State the distribution of X .

- $B(25, 0.04)$

c. Find the probability that at least four of the 25 patients actually have the flu.

- 0.0165

d. On average, for every 25 patients calling in, how many do you expect to have the flu?

- one

10. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given below. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

**Figure
4.38**

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

- Describe the random variable X in words.
- Find the probability that a customer rents three DVDs.
- Find the probability that a customer rents at least four DVDs.
- Find the probability that a customer rents at most two DVDs.

- X = the number of DVDs a Video to Go customer rents
- 0.12

- c. 0.11
- d. 0.77

11. A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

- a. In words, define the random variable X.
- b. List the values that X may take on.
- c. Give the distribution of X. $X \sim \text{_____}(\text{_____,_____})$
- d. How many of the 12 students do we expect to attend the festivities?
- e. Find the probability that at most four students will attend.
- f. Find the probability that more than two students will attend.

Solutions: X = the number of students who will attend Tet. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 $X \sim B(12, 0.18)$ 2.16
0.9511 0.3702

12. The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date).⁶ An upcoming monthly schedule contains 12 games.

a. The expected number of wins for that upcoming month is:

- a. 1.67
- b. 12
- c. $\frac{382}{1043}$
- d. 4.43

• d

b. Let X = the number of games won in that upcoming month. What is the probability that the San Jose Sharks win six games in that upcoming month?

- a. 0.1476
- b. 0.2336
- c. 0.7664

6. Hockey Reference - San Jose Sharks <https://www.hockey-reference.com/teams/SJS/history.html> (accessed January 26, 2021).

d. 0.8903

- a

c. What is the probability that the San Jose Sharks win at least five games in that upcoming month

a. 0.3694

b. 0.5266

c. 0.4734

d. 0.2305

- c
-

13. A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

Solution: X = number of questions answered correctly $X \sim B(10, 0.5)$ We are interested in AT LEAST 70% of ten questions correct. 70% of ten is seven. We want to find the probability that X is greater than or equal to seven. The event “at least seven” is the complement of “less than or equal to six”. Use your software or calculator to get 0.171875. The probability of getting at least 70% of the ten questions correct when randomly guessing is approximately 0.172.

14. A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

- X = number of questions answered correctly
 - $X \sim B\left(32, \frac{1}{3}\right)$
 - We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event “more than 24” is the complement of “less than or equal to 24.”
 - Using your technology of choice: $1 - \text{binomcdf}(32, 1/3, 24)$
 - $P(x > 24) = 0$
 - The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.
-

15. Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable X .

- b. List the values that X may take on.
- c. Give the distribution of X. $X \sim \text{_____}(\text{_____,} \text{_____})$
- d. On average, how many dice would you expect to show a one?
- e. Find the probability that all six dice show a one.
- f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

Solution: X = the number of dice that show a one 0, 1, 2, 3, 4, 5, 6 $X \sim B(6, 1/6)$ 10.00002 three dice

16. More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

- a. In words, define the random variable X.
 - b. List the values that X may take on.
 - c. Give the distribution of X. $X \sim \text{_____}(\text{_____,} \text{_____})$
 - d. On average, how many schools would you expect to offer such courses?
 - e. Find the probability that at most ten offer such courses.
 - f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.
-
- a. X = the number of college and universities that offer online offerings.
 - b. 0, 1, 2, ..., 13
 - c. $X \sim B(13, 0.96)$
 - d. 12.48
 - e. 0.0135
 - f. $P(x = 12) = 0.3186$ $P(x = 13) = 0.5882$ More likely to get 13.
-

17. Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

- a. In words, define the random variable X.
- b. List the values that X may take on.
- c. Give the distribution of X. $X \sim \text{_____}(\text{_____,} \text{_____})$
- d. How many are expected to attend their graduation?
- e. Find the probability that 17 or 18 attend.
- f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Solution: X = the number of students who attend their graduation 0, 1, 2, ..., 22 $X \sim B(22, 0.85)$ 18.7 0.3249 $P(x = 22) = 0.0280$ (less than 3%) which is unusual

18. At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

- In words, define the random variable X .
- List the values that X may take on.
- Give the distribution of X . $X \sim \text{_____}(\text{_____,} \text{_____})$
- How many are expected to **not** to use the foil as their main weapon?
- Find the probability that six do **not** use the foil as their main weapon.
- Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

- X = the number of fencers who do **not** use the foil as their main weapon
 - 0, 1, 2, 3,... 25
 - $X \sim B(25, 0.40)$
 - 10
 - 0.0442
 - The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.
-

19. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

- In words, define the random variable X .
- List the values that X may take on.
- Give the distribution of X . $X \sim \text{_____}(\text{_____,} \text{_____})$
- How many seniors are expected to have participated in after-school sports all four years of high school?
- Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
- Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Solution: X = the number of high school students who participate in after school sports all four years of high school. 0, 1, 2, ..., 60 $X \sim B(60, 0.08)$ 4.8 Yes, $P(x = 0) = 0.0067$, which is a small probability $P(x = 4) = 0.1873$, $P(x = 5) = 0.1824$. More likely to get four.

20. The chance of an IRS audit for a tax return with over \$25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

- In words, define the random variable X.
- List the values that X may take on.
- Give the distribution of X. $X \sim \text{_____}(\text{_____,} \text{_____})$
- How many audits are expected in a 20-year period?
- Find the probability that a person is not audited at all.
- Find the probability that a person is audited more than twice.

- X = the number of audits in a 20-year period
- 0, 1, 2, ..., 20
- $X \sim B(20, 0.02)$
- 0.4
- 0.6676
- 0.0071

21. It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.⁷

- In words, define the random variable X.
- List the values that X may take on.
- Give the distribution of X. $X \sim \text{_____}(\text{_____,} \text{_____})$
- What is the probability that at least eight have adequate earthquake supplies?
- Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
- How many residents do you expect will have adequate earthquake supplies?

Solution: X = the number of California residents who do have adequate earthquake supplies. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 $B(11, 0.30)$ 0.0043 $P(x = 0) = 0.0198$. $P(x = 11) = 0$ or none 3.3

22. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese

7. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012.
http://www.worldearthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being \$1. The player places a bet on a number or object. The “house” rolls three dice. If none of the dice show the number or object that was bet, the house keeps the \$1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her \$1 bet, plus \$1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her \$1 bet, plus \$2 profit. If all three dice show the number or object bet, the player gets back his or her \$1 bet, plus \$3 profit. Let X = number of matches and Y = profit per game.

- In words, define the random variable X .
- List the values that X may take on.
- Give the distribution of X . $X \sim \text{_____}(\text{_____,} \text{_____})$
- List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
- Calculate the average expected matches over the long run of playing this game for the player.
- Calculate the average expected earnings over the long run of playing this game for the player.
- Determine who has the advantage, the player or the house.

- X = the number of matches
- 0, 1, 2, 3
- $X \sim B(3, \frac{1}{6})$
- In dollars: -1, 1, 2, 3
- $\frac{1}{2}$
- Multiply each Y value by the corresponding X probability from the PDF table. The answer is -0.0787. You lose about eight cents, on average, per game.
- The house has the advantage.

23. According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

- What is the probability distribution for X ?
- Using the formulas, calculate the mean and standard deviation of X .
- Find the probability that 15 people in the sample have access to electricity.
- Find the probability that at most ten people in the sample have access to electricity.
- Find the probability that more than 25 people in the sample have access to electricity.

Solution: $X \sim B(150, 0.09)$ Mean = $np = 150(0.09) = 13.5$ Standard Deviation = $npq = 150(0.09)(0.91) \approx 3.5050$ $P(x = 15) = \text{binompdf}(150, 0.09, 15) = 0.0988$ $P(x \leq 10) = \text{binomcdf}(150, 0.09, 10) = 0.1987$ $P(x > 25) = 1 - P(x \leq 25) = 1 - \text{binomcdf}(150, 0.09, 25) = 1 - 0.9991 = 0.0009$

24. The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.⁸

- Sketch a graph of the probability distribution of X .
- Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- Find the probability that more than five people in the sample are literate. Is it more likely that three people or four people are literate.

a. $X \sim B(15, 0.281)$

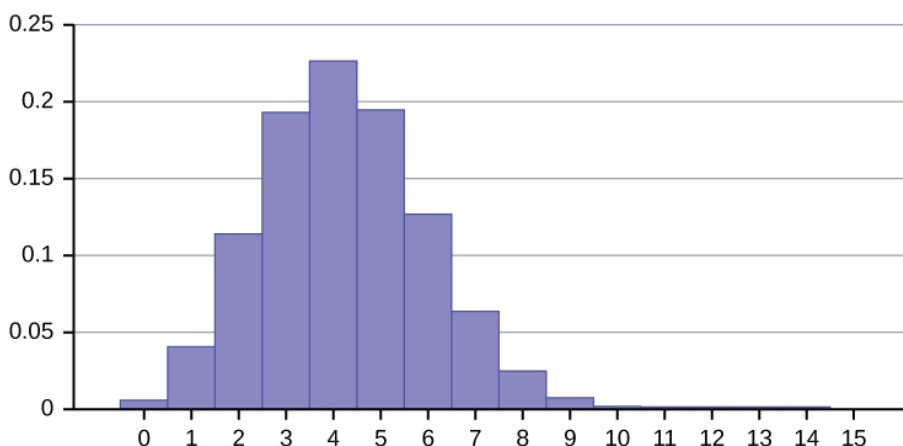


Figure 4.39

- Mean = $\mu = np = 15(0.281) = 4.215$
 - Standard Deviation = $\sigma = \sqrt{npq} = \sqrt{15(0.281)(0.719)} = 1.7409$
- b. $P(x > 5) = 1 - P(x \leq 5) = 1 - \text{binomcdf}(15, 0.281, 5) = 1 - 0.7754 = 0.2246$
 $P(x = 3) = \text{binompdf}(15, 0.281, 3) = 0.1927$
 $P(x = 4) = \text{binompdf}(15, 0.281, 4) = 0.2259$
 It is more likely that four people are literate than three people are.

8. “UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] 300 Chapter 4 | Discrete Random Variables This OpenStax book is available for free at <http://cnx.org/content/col11562/1.18> and writing skills,” UNICEF Television. Video available online at <http://www.unicefusa.org/assets/video/afghan-femaleliteracy-centers.html> (accessed May 15, 2013).

References

Image References

Figure 4.39: Figure 4.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/4-solutions#fs-idm104093808-solution>

Text

Class Catalogue at the Florida State University. Available online at <https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy> (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

“Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

“Distance Education.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

“NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. “Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income,” GALLUP® Economy, 2013. Available online at <http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx> (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics> (accessed May 15, 2013).

CHAPTER 5: CONTINUOUS RANDOM VARIABLES

5.1 Introduction to Continuous Random Variables and The Uniform Distribution

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions
- Recognize the uniform probability distribution and apply it appropriately



Figure 5.1: Plant Heights. The heights of these plants are continuous random variables.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

Note: The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (PDF). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (CDF). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x -axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d . $P(c < x < d)$ is the area under the curve, above the x -axis, to the right of c and the left of d .
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x -axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions however much of the work has already been done for us. The formulas to find the area in this textbook have already been found by using the techniques of integral calculus.

Some Continuous Distributions

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way. We do not often handle general CRVs, but more often study special known cases. The following graphs illustrate some of these these distributions.

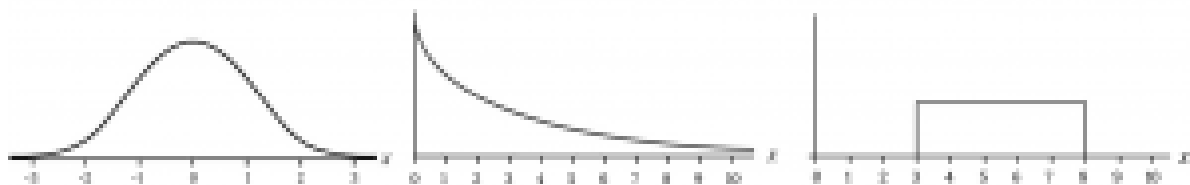


Figure 5.2: Continuous Distributions

Probability Density Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. For continuous probability distributions, you can think about it as: **PROBABILITY = AREA**.

The Uniform Distribution

The (continuous) **uniform distribution** is fairly simple and is a great place to start to demonstrate the ideas of continuous distributions. It is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive of endpoints.

The notation for the uniform distribution is $X \sim U(a, b)$ where a = the lowest value of x and b = the highest value of x .

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Consider the following example.

Example

The following data are 55 smiling times, in seconds, of an eight-week-old baby.

Figure 5.3: Smiling Times

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23

seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is equally likely. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week-old baby's smile.

For this example, $X \sim U(0, 23)$ and $f(x) = \frac{1}{23-0}$ for $0 \leq X \leq 23$.

For this problem, the theoretical mean and standard deviation are

$$\mu = \frac{0 + 23}{2} = 11.50 \text{ seconds and } \sigma = \sqrt{\frac{(23 - 0)^2}{12}} = 6.64 \text{ seconds.}$$

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

Example

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$.

- x = a real number
- The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

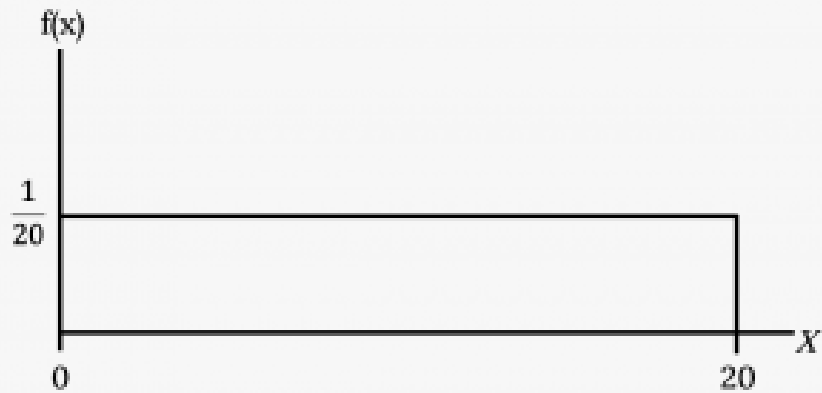


Figure 5.4: Example of a Function

- $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$.
- The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.
- The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x -axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.
- $\text{AREA} = 20 \left(\frac{1}{20} \right) = 1$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x -axis where $0 < x < 2$.

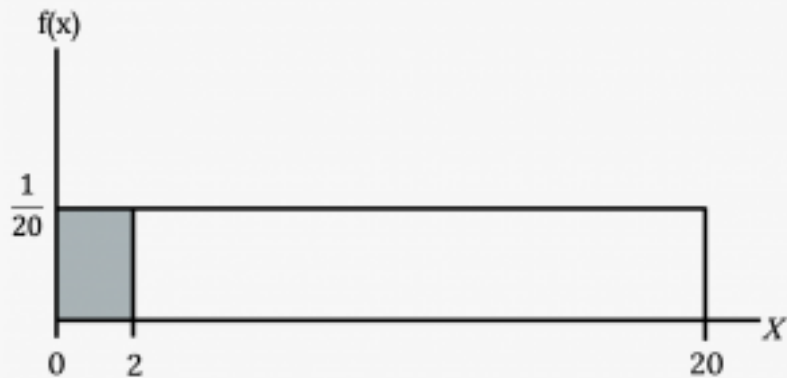


Figure 5.5: Finding Area



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=175#h5p-126>

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x -axis where $4 < x < 15$.

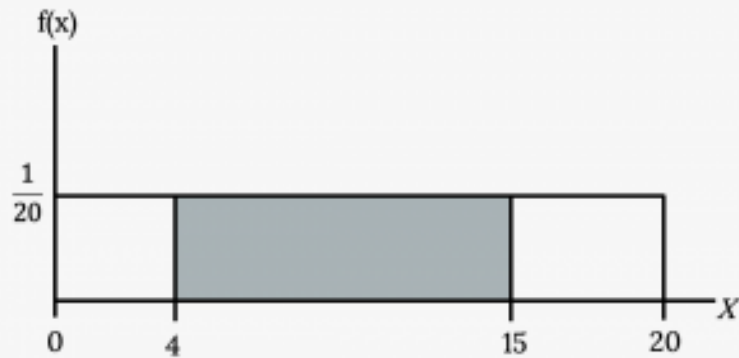


Figure 5.6: Finding Area



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=175#h5p-127>

Suppose we want to find $P(x = 15)$.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=175#h5p-128>

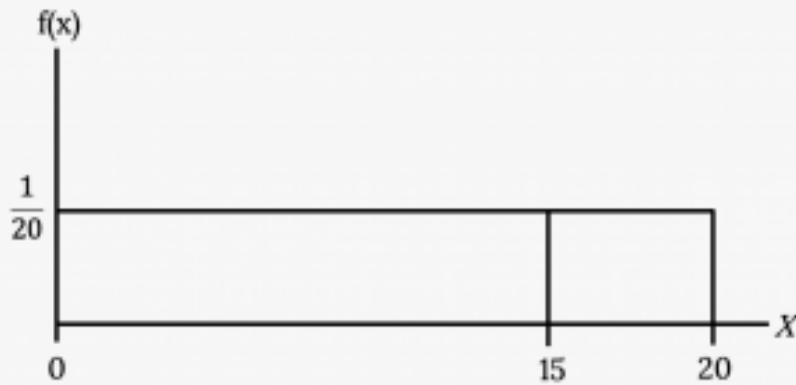


Figure 5.7: Finding a Value

$P(X \leq x)$, which can also be written as $P(X < x)$ for continuous distributions, is called the cumulative distribution function or CDF. Notice the “less than or equal to” symbol. We can also use the CDF to calculate $P(X > x)$. The CDF gives “area to the left” and $P(X > x)$ gives “area to the right.” We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.

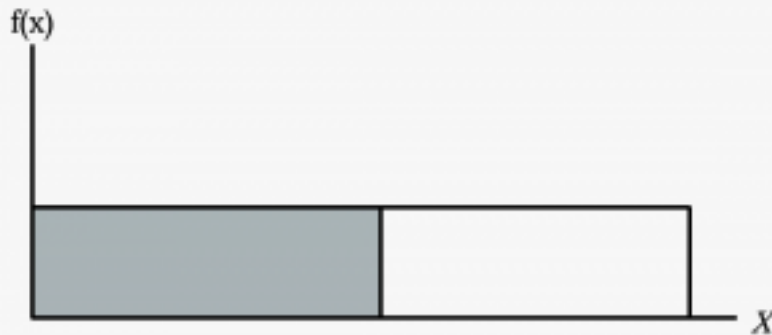


Figure 5.8: Label the Graph

- Label the graph with $f(x)$ and x . Scale the x and y axes with the maximum x and y values.



An interactive H5P element has been excluded from this version of the text. You can view it



online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=175#h5p-129>

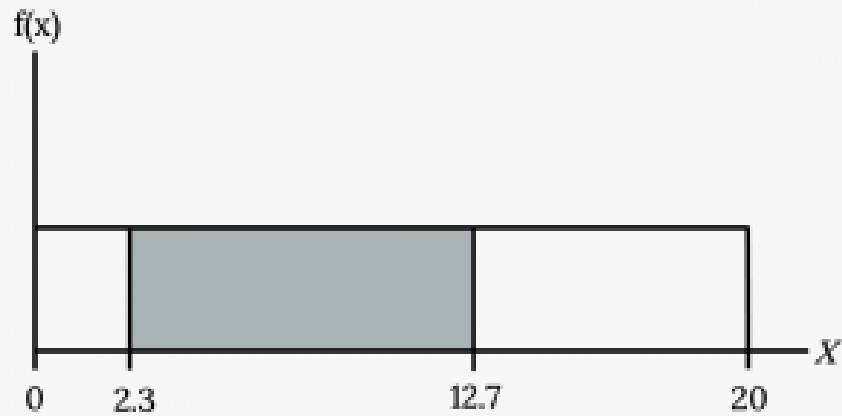


Figure 5.9: Finding Area

- To calculate the probability that x is between two values, look at the graph above. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=175#h5p-130>

Your turn!

Consider the function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Image References

Figure 5.1: Annie Spratt (2018). “Greenhouse / glasshouse.” Public domain. Retrieved from <https://unsplash.com/photos/r1yuNMUw6Lo>

Figure 5.2: Kindred Grey via Virginia Tech (2020). “Figure 5.2” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.2.png. Adaptation of Figures 5.37, 5.38, and 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 5.4: Kindred Grey (2020). “Figure 5.4.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.4.png

Figure 5.5: Kindred Grey (2020). “Figure 5.5.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.5.png

Figure 5.6: Kindred Grey (2020). “Figure 5.6.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.6.png

Figure 5.7: Kindred Grey (2020). “Figure 5.7.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.7.png

Figure 5.8: Kindred Grey (2020). “Figure 5.8.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.8.png

Figure 5.9: Kindred Grey (2020). “Figure 5.9.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.9.png

5.2 The Normal Distribution

The **normal or "Gaussian" distribution** is the most important of all the distributions, continuous or otherwise. Its graph is symmetric, bell-shaped, and unimodal. It is widely used and even more widely abused. You see this distribution in almost all disciplines including psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Many things in the natural world or dealing with humans such as IQ scores or real-estate prices fit a normal distribution.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing $X \sim N(\mu, \sigma)$.

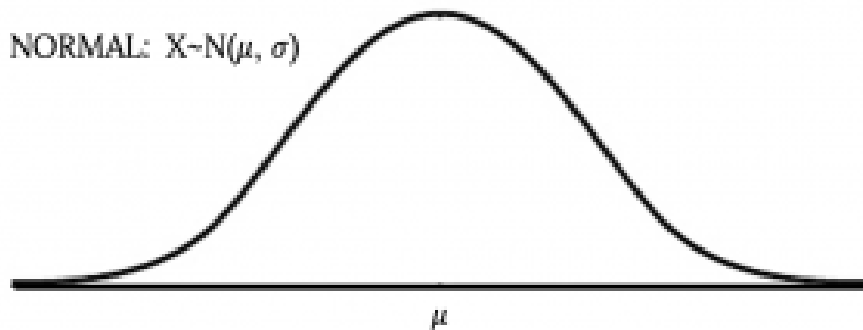


Figure 5.10: Normal Distribution

The **probability density function** of this curve is as follows:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

where:

- $-\infty < X < \infty$
- $-\infty < \mu < \infty$
- $\sigma > 0$

As you can see, the normal pdf is a rather complicated function. This could be a problem since the normal distribution is so widely used. However we will see some ways we can work around this.

The cumulative distribution function is $P(X \leq x)$. It can be calculated either by calculus, technology, or a table. Technology has made the tables almost obsolete.

The curve is symmetric about a vertical line drawn through the mean, μ . In theory, the mean is the same as

the median, because the graph is symmetric about μ . As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ , causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ . A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the standard normal distribution.

The Empirical Rule

The place to start with the normal distribution is the **Empirical Rule**. It applies to any normal distribution or data that has a bell shaped, symmetric curve. It states if X is a random variable and has a normal distribution with mean μ and standard deviation σ , then:

- Approximately 68% of the values of x are within one standard deviation of the mean. ($\pm\sigma$ or z-scores of ± 1)
- Approximately 95% of the values of x are within two standard deviations of the mean. ($\pm 2\sigma$ or z-scores of ± 2)
- Approximately 99.7% of the values of x are within three standard deviations of the mean. ($\pm 3\sigma$ or z-scores of ± 3)

The Empirical Rule is also known as the 68-95-99.7 rule.

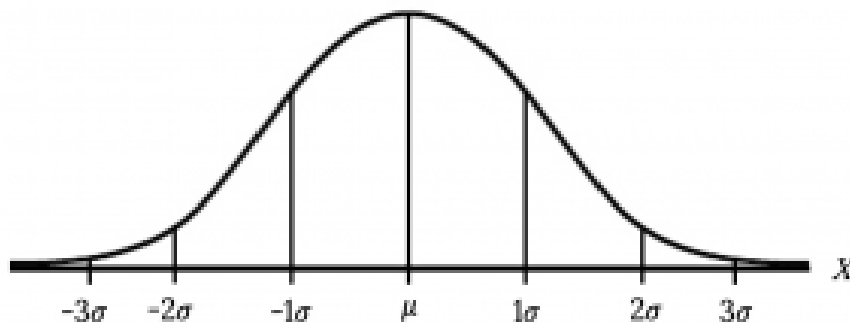


Figure 5.11: Empirical Rule

Example

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie within one standard deviation of the mean. Therefore, about 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation from the mean 50. The z -scores are -1 and $+1$ for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations from the mean 50. The z -scores are -2 and $+2$ for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ from the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z -scores are -3 and $+3$ for 32 and 68, respectively.

Your turn!

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

- About 68% of the y values lie between what two values? These values are _____ and _____. The z -scores are _____ and _____, respectively.
- About 95% of the y values lie between what two values? These values are _____ and _____. The z -scores are _____ and _____, respectively.
- About 99.7% of the y values lie between what two values? These values are _____ and _____. The z -scores are _____ and _____, respectively.

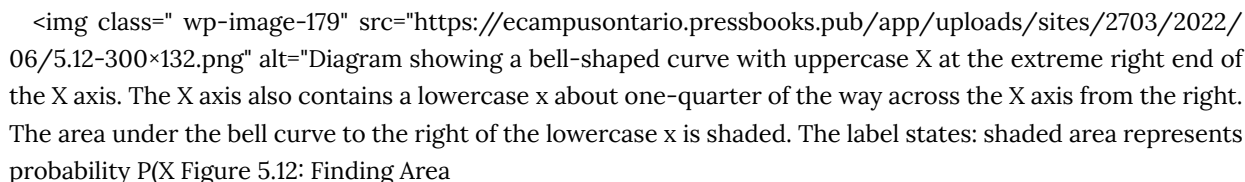


An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-131>

Finding Normal Probabilities

The shaded area in the following graph indicates the area to the left of x . This area is represented by the probability $P(X < x)$.

Figure 5.12: Finding Area

The area to the right is then $P(X > x) = 1 - P(X < x)$. Remember, $P(X < x)$ = Area to the left of the vertical line through x . $P(X > x) = 1 - P(X < x)$ = Area to the right of the vertical line through x . $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

There are 3 main ways we can find probabilities associated with the Normal Distribution. These include:

- Math (via Calculus Integration)
- The Standardizing Process
- Technology

We would like to avoid complicated math if possible.

In order to avoid the math, a process called “Standardizing” can be used. This involves Z scores, the Standard Normal Distribution and Tables. Although this tried and true process is now somewhat antiquated, it is a great place to start.

There are many technologies such as calculators and various statistical software that let us skip the entire standardizing process and instantaneously provide us with a probability. Although we typically have these at our disposal to use in practice, it is good to understand the process going on behind the scenes to make sure we apply our technology correctly.

The Standard Normal Distribution

The **standard normal distribution (SND)** is the simplest form of the normal distribution you can think of. The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x in the given equation comes from a normal distribution with mean μ and standard deviation σ .

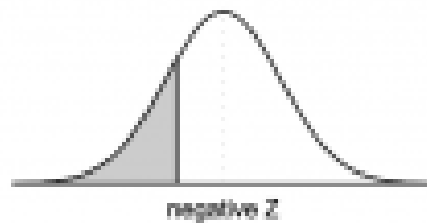
Recall our previous discussion of **z-scores**, which are converted to units of the standard deviation. If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

$$z = \frac{x - \mu}{\sigma}$$

Recall a **z-score** tells you how many standard deviations the value x is above (to the right of) or below (to the

left of) the mean, μ . Values of x that are larger than the mean have positive z -scores, and values of x that are smaller than the mean have negative z -scores. If x equals the mean, then x has a z -score of zero.

We have the Z table at our disposal with probabilities already calculated and organized. Note that most Z tables give us the left tailed, CDF, or “less than” probability. For example the area to the left of a Z score of -3.37 , $P(Z \leq -3.37) = 0.0004$.



Second decimal place of Z										Z
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	-3.3
0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0007	0.0007	-3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	-3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	-3.0

Figure 5.13: Z Table

The SND CDF value, $P(Z \leq z)$, is also denoted as $\Phi(z)$. We can then use these CDF values, $P(Z \leq z)$, and some probability rules to find greater than $[P(Z \geq z) = 1 - P(Z \leq z)]$ or in between $[P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)]$ probabilities.

Example

Use the Z table to find the following probabilities:

a. $P(Z \leq 1)$:





An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-132>

b. $P(Z \geq 1)$:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-133>

c. $P(-1 \leq Z \leq 1)$:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-134>

Your Turn

Use the Z table to find the following probabilities:

a. $P(Z \leq -0.54)$:

b. $P(Z \geq 1.2)$:

c. $P(-1.5 \leq Z \leq 0.84)$:

The standardizing process

So far we have seen the idea that we can convert any normal distribution with any mean and standard deviation to the standard normal distribution in units of z-scores. We also have the associated probabilities in our Z table. Essentially, the work has been done for us if we know how to standardize and look up the associated probability in the table. The general process is:

$$X \sim N(\mu, \sigma) \rightarrow Z \sim N(0, 1) \rightarrow \text{Probability from Z table}$$

This process, while maybe outdated in our technology age, is good for beginners to understand and useful when we do not have access to technology.

Example

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights, then find the associated probabilities.

- a. The probability that a child weighs less than 11 kg



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-135>

- b. The probability that a child weighs more than 7.9 kg



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-136>

- c. The probability that a child weighs between 11.2 and 12.2 kg \geq



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-137>

Your Turn

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

- Find the probability that a randomly selected golfer scored less than 65.
- The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

The “un-standardizing” process

Sometimes we may be given a percentile or z-score and want to work backwards through the standardizing process to find a value on the original distribution. You could call this “un-standardizing” or finding a normal **quantile**. The process looks like this:

Probability in Z table $\rightarrow Z \sim N(0,1) \rightarrow X \sim N(\mu, \sigma)$

For example, if the mean of a normal distribution is five and the standard deviation is two, what value is three standard deviations above (or to the right of) the mean (z-score = three). Rearranging the z-score formula, the calculation is as follows:

$$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$$

Often we are given a percentile to find on the original distribution. For example, what if we want to know a value on the previous distribution that corresponds to the 90th percentile? We can look up a probability of 0.9 in the Z table and find a corresponding z-score of approximately 1.28.

$$x = \mu + (z)(\sigma) = 5 + (1.28)(2) = 7.56$$

Example

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

- a. Find the 90th percentile for the diameters of mandarin oranges:



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-138>

- b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=181#h5p-139>

Your Turn

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

- a. Calculate the first and third quartile scores for this exam.
- b. The middle 50% of the exam scores are between what two values?

Image References

Figure 5.10: Kindred Grey via Virginia Tech (2020). “Figure 5.10” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.10.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 5.11: Kindred Grey via Virginia Tech (2020). “Figure 5.11” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.11.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 5.12: Kindred Grey via Virginia Tech (2020). “Figure 5.12” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.12.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 5.13: Kindred Grey via Virginia Tech (2020). “Figure 5.13” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.13.png . Adaptation of ‘Normal Probability Table’ from OpenIntro Statistics (2019) (CC BY-SA 3.0). Retrieved from https://openintro.org/go/?id=stat_prob_tables_normal_t_chisq&referrer=/book/isrs/index.php

5.3 The Normal Approximation to the Binomial

The **binomial formula** is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. Consider the following example.

Example

Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=185#h5p-140>



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=185#h5p-141>

The computations in the previous example are tedious, long, and near impossible if you do not have access to technology. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

Binomial Approximation Conditions

Consider the binomial model when the probability of a success is $p = 0.10$. The following figures show four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? What distribution does the last histogram resemble?

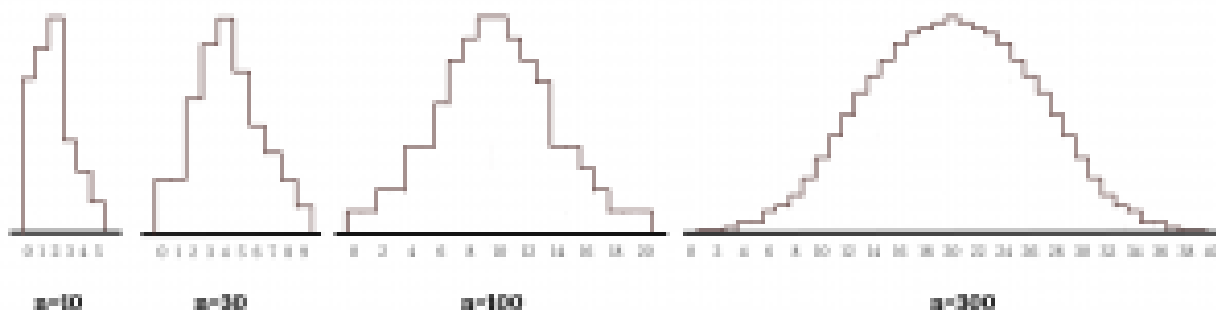


Figure 5.14: Hollow Histograms for Different Sample Sizes

It appears the distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram.

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both **at least 10**. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \text{ and } \sigma = \sqrt{np(1 - p)}$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of the previous example:

Example (Continued)

Use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$.

Already knowing that the binomial model, we then verify that both np and $n(1 - p)$ are at least 10:

- $np = 400 \times 0.15 = 60$ $n(1 - p) = 400 \times 0.85 = 340$

With these conditions met, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

- $\mu = np = 60$ and $\sigma = np(1 - p) = 7.14$

We want to find the probability of observing 42 or fewer smokers using this model. Use the normal model $N(\mu = 60, \sigma = 7.14)$ and standardize to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution we found in the previous of example, 0.0054.

Compute the Z-score first:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=185#h5p-142>

The Continuity Correction

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$. With such a large sample, we might be tempted to apply the normal approximation and use the range 49 to 51. However, we would find that the binomial solution and the normal approximation notably differ:

- Binomial: 0.0649
- Normal: 0.0421

We can identify the cause of this discrepancy in the next figure which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

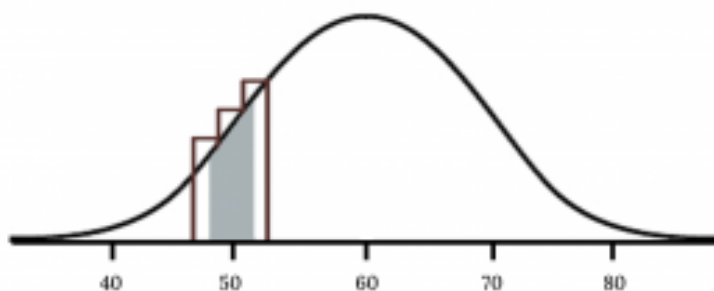


Figure 5.15: Continuity Correction

The normal approximation to the binomial distribution for intervals of values can usually be improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5. This is called the **continuity correction**.

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

Image References

Figure 5.14: Kindred Grey (2020). "Figure 5.14." CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.14.png

Figure 5.15: Kindred Grey via Virginia Tech (2020). "Figure 5.15" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_5.15.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Chapter 5 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=190#h5p-143>

Section Reviews

5.1 Introduction

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.

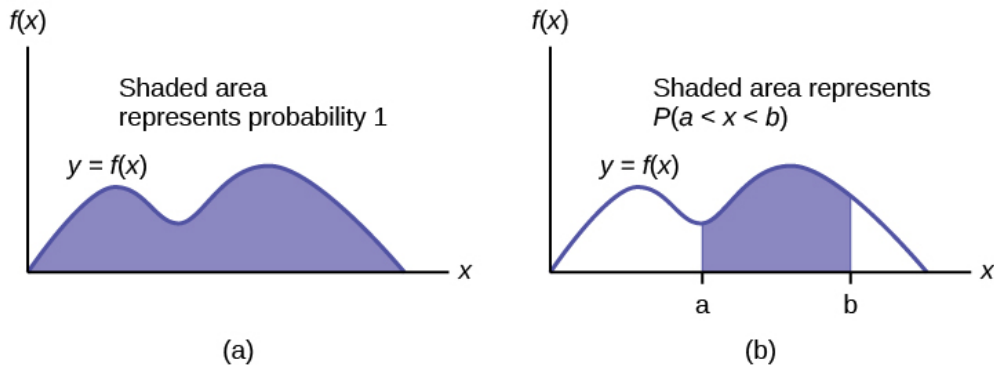


Figure 5.16: Area Under the Density Curve

The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x .

Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

5.2 Normal Distribution

A z -score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z -scores is zero and the standard deviation is one. If z is the z -score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ .

z = a standardized value (z -score)

mean = 0; standard deviation = 1

To find the k^{th} percentile of X when the z -scores is known:

$$k = \mu + (z)\sigma$$

$$\text{z-score: } z = \frac{x - \mu}{\sigma}$$

Z = the random variable for z -scores

5.3 Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n , you had to use the binomial formula, which could be very complicated. Using the normal approximation to the binomial distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a binomial distribution:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor and is used in the following example.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

5.1 Introduction

- **Continuous random variable**
- **Probability density function (PDF)**
- **Cumulative distribution function (CDF)**
- **Uniform distribution**

5.2 Normal Distribution

- Normal (Gaussian) distribution
- Probability density function (PDF)
- Empirical rule
- Standard normal distribution (SND)
- Z-score
- Quantile

5.3 Normal Approximation to the Binomial

- Binomial formula
- Continuity correction

Extra Practice

5.1 Introduction

1. What does the shaded area represent? $P(______ < x < ______)$

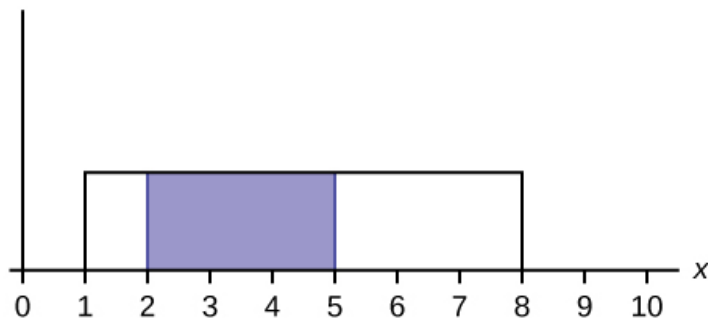


Figure 5.17

2. What does the shaded area represent? $P(______ < x < ______)$

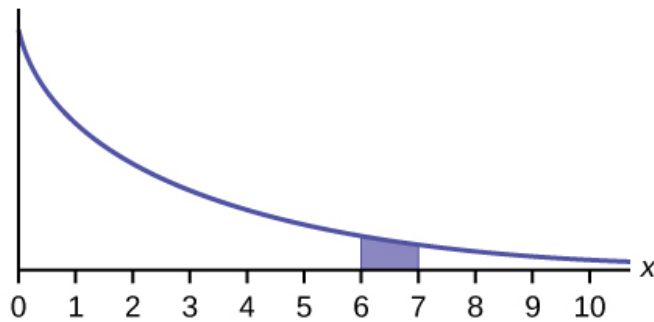


Figure 5.18

- $P(6 < x < 7)$

3. For a continuous probability distribution, $0 \leq x \leq 15$. What is $P(x > 15)$?

4. What is the area under $f(x)$ if the function is a continuous probability density function?

- One

5. For a continuous probability distribution, $0 \leq x \leq 10$. What is $P(x = 7)$?

6. A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

- Zero

7. $f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \leq x \leq 5$. What is $P(x < 0)$?

8. $f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \leq x \leq 12$. What is $P(0 < x < 12)$?

- One
-

9. Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer “yes” or “no.” You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.

- a. What part of the experiment will yield discrete data?
- b. What part of the experiment will yield continuous data?

When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

Age is a measurement, regardless of the accuracy used.

5.2 Normal Distribution

1. The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm.¹ Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when $x = 168$ cm is $z = \underline{\hspace{2cm}}$. This z-score tells you that $x = 168$ is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean $\underline{\hspace{2cm}}$ (What is the mean?).

- -0.32, 0.32, left, 170

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean.

- 177.98 cm, 1.27, right
-

1. Data from The World Almanac and Book of Facts.

2. Use the information in Number 1 to answer the following questions.

- a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z = \underline{\hspace{2cm}}$. This z-score tells you that $x = 176$ cm is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean $\underline{\hspace{2cm}}$ (What is the mean?).
 - b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = -2$. What is the male's height? The z-score ($z = -2$) tells you that the male's height is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean.
-

3. In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$.² Let $X =$ a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$. Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$ as they compare to their respective means and standard deviations?

4. What is the z-score of x , when $x = 1$ and $X \sim N(12, 3)$?

5. Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently.³ Suppose weight loss has a normal distribution. Let $X =$ the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean $\underline{\hspace{2cm}}$ (What is the mean?).

- This z-score tells you that $x = 10$ is **2.5** standard deviations to the **right** of the mean **five**.

b. Suppose a person **gained** three pounds (a negative weight loss). Then $z = \underline{\hspace{2cm}}$. This z-score tells you that $x = -3$ is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean.

- $z = -4$. This z-score tells you that $x = -3$ is **four** standard deviations to the **left** of the mean.

2. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).
3. McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

c. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z ?

- $z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.
 - The z -score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.
-

6. Fill in the blanks. Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z -score when $x = 10$ is -1.5 . This score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

7. From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm.⁴ Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Find the z -scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z -score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm as they compare to their respective means and standard deviations?

- The z -score for $x = 160.58$ is $z = -1.5$.
 - The z -score for $y = 162.85$ is $z = -1.5$.
 - Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.
-

8. In 2012, 1,664,479 students took the SAT exam.⁵ The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$. Find the z -scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z -score. What can you say about $x_1 = 325$ and $x_2 = 366.21$ as they compare to their respective means and standard deviations?

4. Data from The World Almanac and Book of Facts.

5. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).

9. Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

10. The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

- About 68% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
 - About 95% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
 - About 99.7% of the y values lie between what two values? These values are _____. The z -scores are _____, respectively.
-

11. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ _____.

- ounces of water in a bottle
-

12. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

- 61
-

13. $X \sim N(1, 2)$. $\sigma =$ _____

- 2
-

14. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ _____.

- diameter of a rubber ball
-

15. $X \sim N(-4, 1)$. What is the median?

- -4
-

16. $X \sim N(3, 5)$. $\sigma =$ _____

- 5
-

17. $X \sim N(-2, 1)$. $\mu =$ _____

- -2
-

18. What does a z-score measure?

- The number of standard deviations a value is from the mean.
-

19. What does standardizing a normal distribution do to the mean?

- The mean becomes zero.
-

20. Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

- Yes because the mean is zero, and the standard deviation is one.
-

21. What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

- $z = 2$
-

22. What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

- $z = -1.5$
-

23. What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

- $z = 2.78$
-

24. What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

- $z = -0.133$
-

25. Suppose $X \sim N(2, 6)$. What value of x has a z-score of three?

- $x = 20$
-

26. Suppose $X \sim N(8, 1)$. What value of x has a z-score of -2.25 ?

- $x = 5.75$
-

27. Suppose $X \sim N(9, 5)$. What value of x has a z-score of -0.5 ?

- $x = 6.5$
-

28. Suppose $X \sim N(2, 3)$. What value of x has a z-score of -0.67 ?

- $x = -0.01$
-

29. Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

- $x = 1$
-

30. Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

- $x = 8$
-

31. Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

- $x = 1.97$
-

32. Suppose $X \sim N(-1, 2)$. What is the z -score of $x = 2$?

- $z = 1.5$
-

33. Suppose $X \sim N(12, 6)$. What is the z -score of $x = 2$?

- $z = -1.67$
-

34. Suppose $X \sim N(9, 3)$. What is the z -score of $x = 9$?

- $z = 0$
-

35. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z -score of $x = 5.5$?

- $z \approx -0.33$
-

36. In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is _____ standard deviations to the _____ (right or left) of the mean.

- 1.25, left
-

37. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is _____ standard deviations to the _____ (right or left) of the mean.

- 0.67, right
-

38. In a normal distribution, $x = -2$ and $z = 6$. This tells you that $x = -2$ is _____ standard deviations to the _____ (right or left) of the mean.

- six, right
-

39. In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is _____ standard deviations to the _____ (right or left) of the mean.

- 3.14, left
-

40. In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is _____ standard deviations to the _____ (right or left) of the mean.

- 1.7, left
-

41. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

- about 68%
-

42. About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

- about 95.45%
-

43. About what percent of x values lie between the second and third standard deviations (both sides)?

- about 4%
-

44. Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

- between 12 and 18
-

45. Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution (i.e., -3).

- between -5 and -1
-

46. Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

- between -4 and -3 or between -3 and -2
-

47. About what percent of x values lie between the mean and three standard deviations?

- about 50%
-

48. About what percent of x values lie between the mean and one standard deviation?

- about 34.14%
-

49. About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

- about 27%
-

50. About what percent of x values lie between the first and third standard deviations (both sides)?

- about 34.46%
-

51. The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

a. Define the random variable X in words. X = _____.

- The lifetime of a Sunshine CD player measured in years.

b. $X \sim \text{_____}(\text{_____,} \text{_____})$

- $X \sim N(4.1, 1.3)$
-

52. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

a. What is the median recovery time?

- a. 2.7
- b. 5.3
- c. 7.4
- d. 2.1

- B

b. What is the z-score for a patient who takes ten days to recover?

- a. 1.5
- b. 0.2
- c. 2.2
- d. 7.3

- C
-

53. The length of time to find it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

- I. The data cannot follow the uniform distribution.
- II. The data cannot follow the exponential distribution..
- III. The data cannot follow the normal distribution.

- a. I only
- b. II only
- c. III only
- d. I, II, and III

- B

54. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches.⁶ For each of the following heights, calculate the z-score and interpret it using complete sentences.

- a. 77 inches
 - b. 85 inches
 - c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.
- a. Use the z-score formula. $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
 - b. Use the z-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
 - c. Height = $79 + 3.5(3.89) = 92.615$ inches, which is taller than 7 feet, 8 inches. There are very few NBA players this tall so the answer is no, not likely.

55. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.⁷

- a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
 - b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?
- Use the z-score formula. $100 - 125/14 \approx -1.8$ and $150 - 125/14 \approx 1.8$ I would tell him that 2.5 standard deviations below the mean would give him a blood pressure reading of 90, which is below the range of 100 to 150.

6. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

7. “Blood Pressure of Males and Females.” StatCrunch, 2013. Available online at <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960> (accessed May 14, 2013).

56. Kyle's doctor told him that the z -score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If X = a systolic blood pressure score then $X \sim N(125, 14)$.

- a. Which answer(s) **is/are** correct?
 - i. Kyle's systolic blood pressure is 175.
 - ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
 - iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
 - iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
 - b. Calculate Kyle's blood pressure.
- a. iv
 - b. Kyle's blood pressure is equal to $125 + (1.75)(14) = 149.5$.

57. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.⁸

- a. Calculate the z -score for an SAT score of 720. Interpret it using a complete sentence.
- b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
- c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3.⁹ If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Let X = an SAT math score and Y = an ACT math score.

- a. $X = 720$ $\frac{720 - 520}{115} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.
- b. $z = 1.5$
The math SAT score is $520 + 1.5(115) \approx 692.5$. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.

8. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).
9. "Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

- c. $\frac{X - \mu}{\sigma} = \frac{700 - 514}{117} \approx 1.59$, the z-score for the SAT. $\frac{Y - \mu}{\sigma} = \frac{30 - 21}{5.3} \approx 1.70$, the z-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher z-score).
-

5.3 Normal Approximation to the Binomial

1. Suppose in a local Kindergarten through 12th grade (K – 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

- Find the probability that **at least 150** favor a charter school.
- Find the probability that **at most 160** favor a charter school.
- Find the probability that **more than 155** favor a charter school.
- Find the probability that **fewer than 147** favor a charter school.
- Find the probability that **exactly 175** favor a charter school.

Solution:

Let X = the number that favor a charter school for grades K through 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y . $Y \sim N(159, 8.6447)$.

- For part a, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.
 - For part b, you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.
 - For part c, you **exclude 155** so $P(X > 155)$ has normal approximation $P(Y > 155.5) = 0.6572$.
 - For part d, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.
 - For part e, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.
-

2. In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

References

Image References

Figure 5.16: Figure 5.21 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/5-chapter-review>

Figure 5.17: Figure 5.26 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/5-practice>

Figure 5.18: Figure 5.27 from OpenStax Introductory Business Statistics (2012) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-business-statistics/pages/5-practice>

Text

“Blood Pressure of Males and Females.” StatCrunch, 2013. Available online at <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960> (accessed May 14, 2013).

“The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores.” London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

CHAPTER 6: FOUNDATIONS OF INFERENCE

6.1 Point Estimation and Sampling Distributions

Learning Objectives

By the end of this chapter, the student should be able to:

- Understand point estimation
- Apply and interpret the Central Limit Theorem
- Construct and interpret confidence intervals for means when the population standard deviation is known
- Understand the behavior of confidence intervals
- Carry out hypothesis tests for means when the population standard deviation is known
- Understand the probabilities of error in hypothesis tests



Figure 6.1: Loose Change. If you want to figure out the distribution of the change people carry in their pockets, and your sample is large enough, you will find that the distribution follows certain patterns.

Statistical Inference

It is often necessary to “guess”, infer, or generalize about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

Statistical inference uses what we know about probability to make our best “guesses” or estimates from **samples** about the **population** they came from. The main forms of Inference are:

- 1. **Point estimation**
- 2. **confidence interval**
- 3. **Hypothesis testing**

Point Estimation

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a **point estimate** of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

The most natural way to estimate features of the population (**parameters**) is to use the corresponding summary **statistic** calculated from the sample. Some common point estimates and their corresponding parameters are found in the following table:

Figure 6.2: Parameters and Point Estimates

Parameter	Measure	Statistic
μ	Mean of a single population	\bar{x}
p	Proportion of a single population	\hat{p}
μ_D	Mean difference of two dependent populations (MP)	\bar{x}_D
$\mu_1 - \mu_2$	Difference in means of two independent populations	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S

Suppose the mean weight of a sample of 60 adults is 173.3 lbs; this sample mean is a point estimate of the population mean weight, μ . Remember this is one of many samples that we could have taken from the population. If a different random sample of 60 individuals were taken from the same population, the new sample mean would likely be different as a result of **sampling variability**. While estimates generally vary from one sample to another, the population mean is a fixed value.

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a point estimate of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the parameter of interest. When the parameter is a proportion, it is often denoted by p , and we often refer to the sample proportion as \hat{p} (pronounced "p-hat"). Unless we collect responses from every individual in the population, p remains unknown, and we use \hat{p} as our estimate of p .

How would one estimate the difference in average weight between men and women? Suppose a sample of men yields a mean of 185.1 lbs and a sample of women yields a mean of 162.3 lbs. What is a good point estimate for the difference in these two population means? We will expand on this in following chapters.

Sampling Distributions

We have established that different samples yield different statistics due to sampling variability. These statistics have their own distributions, called sampling distributions, that reflect this as a random variable. The **sampling distribution** of a sample statistic is the distribution of the point estimates based on samples of a fixed size, n , from a certain population. It is useful to think of a particular point estimate as being drawn from a sampling distribution.

Recall the sample mean weight calculated from a previous sample of 173.3 lbs. Suppose another random sample of 60 participants might produce a different value of \bar{x} , such as 169.5 lbs. Repeated random sampling could result in additional different values, perhaps 172.1 lbs, 168.5 lbs, and so on. Each sample mean can be thought of as a single observation from a random variable \bar{X} . The distribution of \bar{X} is called the sampling distribution of the sample mean, and has its own mean and standard deviation like the random variables discussed previously. We will simulate the concept of a sampling distribution using technology to repeatedly sample, calculate statistics, and graph them. However, the actual sampling distribution would only be attainable if we could theoretically take an infinite amount of samples.

Each of the point estimates in the table above have their own unique sampling distributions which we will look at in the future

Unbiased Estimation

Although variability in samples is present, there remains a fixed value for any population parameter. What makes a statistical estimate of this parameter of interest a "Good" one? It must be both accurate and precise.

The accuracy of an estimate refers to how well it estimates the actual value of that parameter. Mathematically, this is true when that the expected value your statistic is equal to the value of that parameter. This can be visualized as the center of the sampling distribution appearing to be situated at the value of that parameter.

According to the **law of large numbers**, probabilities converge to what we expect over time. Point estimates follow this rule, becoming more accurate with increasing sample size. The figure below shows the sample mean weight calculated for random samples drawn, where sample size increases by 1 for each draw until sample size equals 500. The maroon dashed horizontal line is drawn at the average weight of all adults 169.7 lbs, which represents the population mean weight according to the CDC.

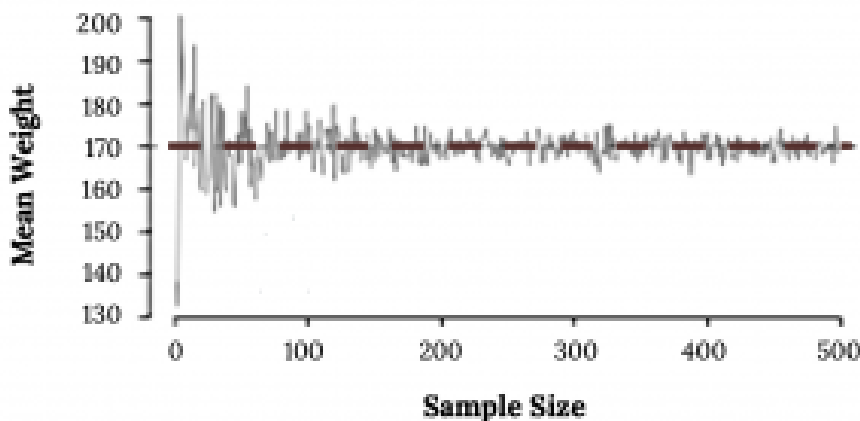


Figure 6.3: Law of Large Numbers

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

In addition to accuracy, a precise estimate is also more useful. This means when repeatedly sampling, the values of the statistics seem pretty close together. The precision of an estimate can be visualized as the spread of the sampling distribution, usually quantified by the standard deviation. The phrase “the standard deviation of a sampling distribution” is often shortened to the **standard error**. A smaller standard error means a more precise estimate and is also effected by sample size.

Image Credits

Figure 6.1: Michael Longmire (2019). “Coins spilling out of a jar.” Public domain. Retrieved from <https://unsplash.com/photos/lhltMGdohc8>

Figure 6.3: Kindred Grey (2020). “Figure 6.3.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.3.png

6.2 The Sampling Distribution of the Sample Mean (σ Known)

Let's start our foray into inference by focusing on the sample mean. Why are we so concerned with means? Two reasons: they give us a middle ground for comparison, and they are easy to calculate. In this section we will see what we can deduce about the sampling distribution of the sample mean.

The Central Limit Theorem for a Sample Mean

The **central limit theorem (CLT)** is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ , and a known standard deviation, σ . The first alternative says that if we collect samples of size n with a “large enough n ,” then the resulting distribution can be approximated by the normal distribution.

Applying the **law of large numbers** here, we could say that if you take larger and larger samples from a population, then the mean \bar{X} of the sample tends to get closer and closer to μ . From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for \bar{X} is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean \bar{X} must be close to the population mean μ . We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

The size of the sample, n , that is required in order to be “large enough” depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. Sampling is done with replacement.

The following images look at sampling distributions of the sample mean built from taking 1000 samples of different sample sizes from a normal Population. What pattern do you notice?

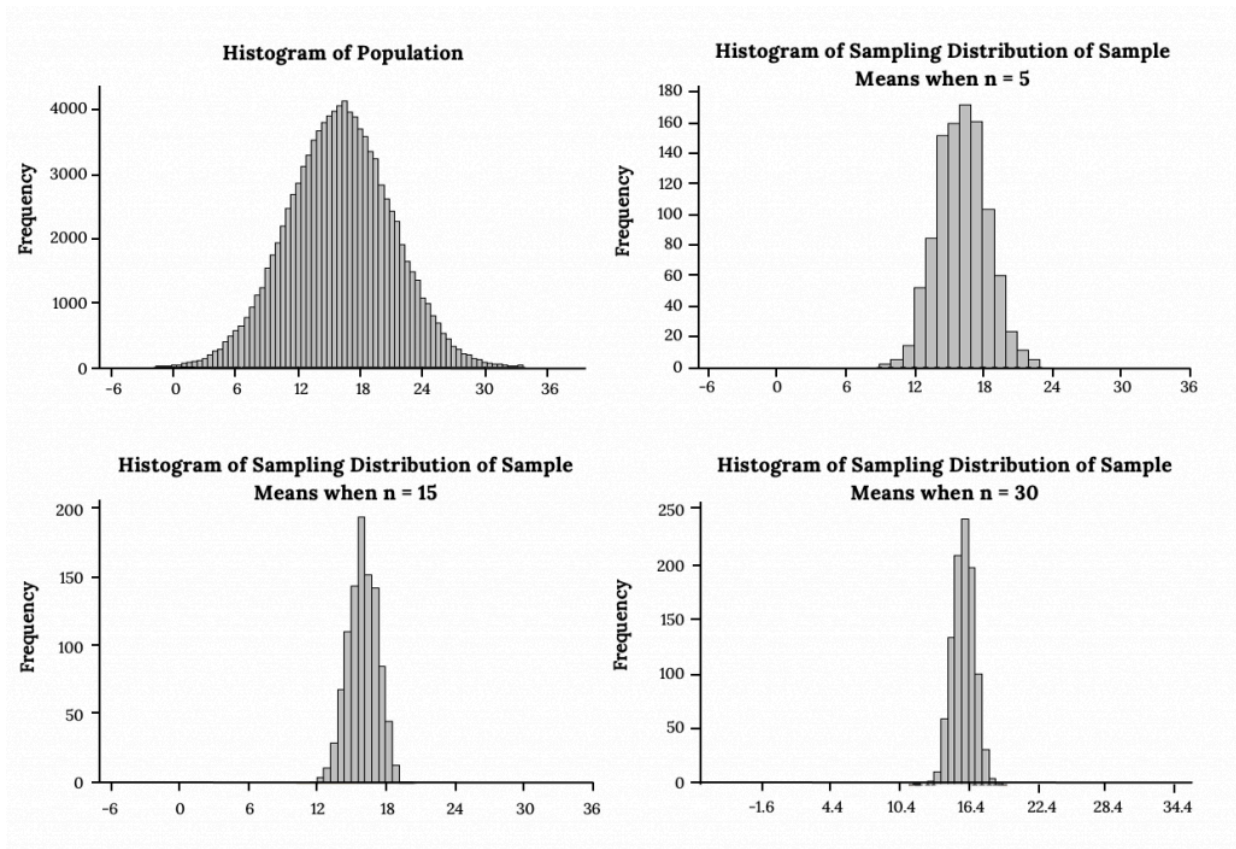


Figure 6.4: Sampling Distributions of the Sample Mean from a Normal Population

The following images look at sampling distributions of the sample mean built from taking 1000 samples of different sample sizes from a non-normal Population (in this case it happens to be exponential). What pattern do you notice?

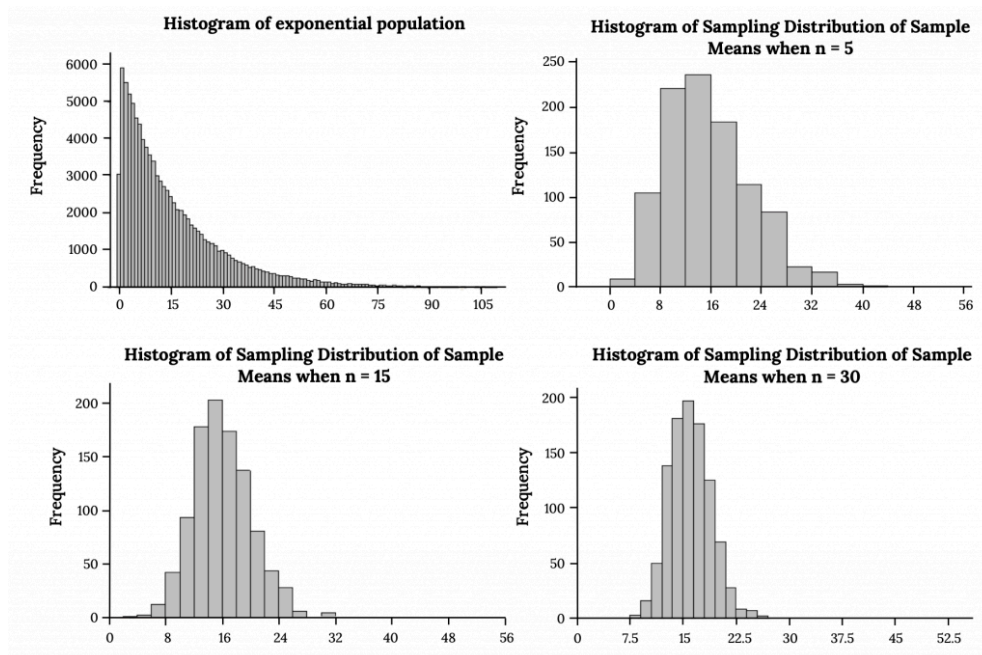


Figure 6.5: Sampling Distributions of the Sample Mean from a Non-Normal Population

What differences do you notice when sampling from a normal population vs. Non normal?

Example

Suppose:

- eight students roll one fair die ten times
- seven roll two fair dice ten times
- nine roll five fair dice ten times
- 11 roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample mean of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2 + 2 + 3 + 4 + 6}{5} = 3.4$. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

As the number of dice rolled increases from one to two to five to ten, the following would happen:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

We have just demonstrated the idea of central limit theorem (clt) for means, that as you increase the sample size, the sampling distribution of the sample mean tends toward a normal distribution.

To summarize, the central limit theorem for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, the sample means form their own normal distribution (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. Standard deviation is the square root of variance, so the standard deviation of the sampling distribution (aka standard error) is the standard deviation of the original distribution divided by the square root of n . The variable n is the number of values that are averaged together, not the number of times the experiment is done.

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool. We can simulate this idea using technology.

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, let:

- μ_X = the mean of X
- σ_X = the standard error of X

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ = standard deviation of \bar{x} and is called the **standard error** of the mean. Note here we are assuming we know the population standard deviation.

If you draw random samples of size n , then as n increases, the random variable \bar{X} which consists of sample means, tends to be normally distributed and

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right).$$

To put it more formally, if you draw random samples of size n , the distribution of the random variable \bar{X} , which consists of sample means, is called the sampling distribution of the sample mean. The sampling distribution of the mean approaches a normal distribution as n , the sample size, increases.

Using the CLT

It is important to understand when to use the central limit theorem:

- If you are being asked to find the probability of an individual value, do not use the CLT. Use the

distribution of its random variable.

- If you are being asked to find the probability of the mean of a sample, then use the CLT for the mean.

The random variable \bar{X} has a different z-score formula associated with it from that of a single observation. Remember, The mean \bar{x} is the mean of one sample and μ_X is the average, or center, of both X (The original distribution) and \bar{X} .

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

We can use our Z table and standardize just as we are already familiar with, or can use your technology of choice

Example

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

a. Find the probability that the sample mean is between 85 and 92.

- Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=200#h5p-144>

- Find $P(85 < \bar{x} < 92)$. Draw a graph.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=200#h5p-145>

<img class=" wp-image-199" src="https://ecampusontario.pressbooks.pub/app/uploads/sites/

2703/2022/06/6.4-1-300×146.png" alt="Normal distribution curve where the peak of the curve coincides with the point 90 on the horizontal axis. The points 85 and 92 are labeled on the axis. Vertical lines are drawn from these points to the curve and the area between the lines is shaded. The shaded region represents the probability that $85 < x$ Figure 6.6: Area Under the Curve

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=200#h5p-146>

Your turn!

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size $n = 30$ are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Image References

Figure 6.4: Kindred Grey via Virginia Tech (2021). “Sampling Distributions of the Sample Mean from a Normal Population.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Sampling_Distributions_of_the_Sample_Mean_from_a_Normal_Population.png

Figure 6.5: Kindred Grey via Virginia Tech (2021). “Sampling Distributions of the Sample Mean from a Non-Normal Population.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Sampling_Distributions_of_the_Sample_Mean_from_a_Non-Normal_Population.png

Figure 6.6: Kindred Grey via Virginia Tech (2020). “Figure 6.4” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.4.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

6.3 Introduction to Confidence Intervals



Figure 6.7: M&Ms. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question.

We use **inferential statistics** to make generalizations about an unknown **population**. The simplest way of doing this is to use the sample data help us to make a **point estimate** of a population **parameter**. We realize that due to **sampling variability** the point estimate is most likely not the exact value of the population parameter, but should be close to it. After calculating point estimates, we can build off of them to construct interval estimates, called confidence intervals.

Confidence Intervals

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall. Essentially the idea is that since a point estimate may not be perfect due to variability, we will build an interval based on a point estimate to hopefully capture the parameter of interest in the interval. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.

It is important to keep in mind that the confidence interval itself is a random variable, while the population parameter is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, \bar{x} . You would use \bar{x} to estimate the population mean. The sample mean, \bar{x} , is the point estimate for the population mean, μ .

Suppose, for the iTunes example, we do not know the population mean μ , but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, \bar{x} , will be within two standard deviations of the population mean μ . For our iTunes example, two standard deviations is $(2)(0.1) = 0.2$. The sample mean \bar{x} is likely to be within 0.2 units of μ .

Because \bar{x} is within 0.2 units of μ , which is unknown, then μ is likely to be within 0.2 units of \bar{x} in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $(2)(0.1)$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \text{ and } \bar{x} + 0.2 = 2 + 0.2 = 2.2$$

We can say that we are about 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The approximate 95% confidence interval is (1.8, 2.2).

This approximate 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an \bar{x} that is not within 0.2 units of the true mean μ . The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence intervals are created for an unknown population parameter. Confidence intervals for most parameters have the form:

$$(\text{Point Estimate} \pm \text{Margin of Error}) = (\text{Point Estimate} - \text{Margin of Error}, \text{Point Estimate} + \text{Margin of Error})$$

The **margin of error (MoE)** depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase “margin of error.” Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where MoE = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , where the population standard deviation is known, we need \bar{x} as an estimate for μ and we need the margin of error. Here, the margin of error (MoE). The sample mean \bar{x} is the point estimate of the unknown population mean μ .

Since a confidence interval estimate will have the form:

$$(PE - MoE, PE + MoE)$$

Then a confidence Interval for the unknown population mean μ in symbols would look like:

$$(\bar{x} - MoE, \bar{x} + MoE)$$

Remember, the margin of error (MoE) depends mainly on the confidence level (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL and represents the chance that the interval does not contain the unknown population parameter.

Mathematically, $\alpha + CL = 1$.

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean \bar{x} from the sample data. Remember, in this section we already know the population standard deviation σ .
- Find the z-score (Critical Value) that corresponds to the confidence level.
- Calculate the margin of error (MoE).
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Example

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5. Find the confidence interval and interpret.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=205#h5p-147>

Your turn!

Suppose we have data from a sample. The sample mean is 15, and the margin of error for the mean is 3.2. What is the confidence interval estimate for the population mean?

Changing the Confidence Level

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $\text{MoE} = 5$.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10\%$ in both tails, or 5% in each tail, of the normal distribution.

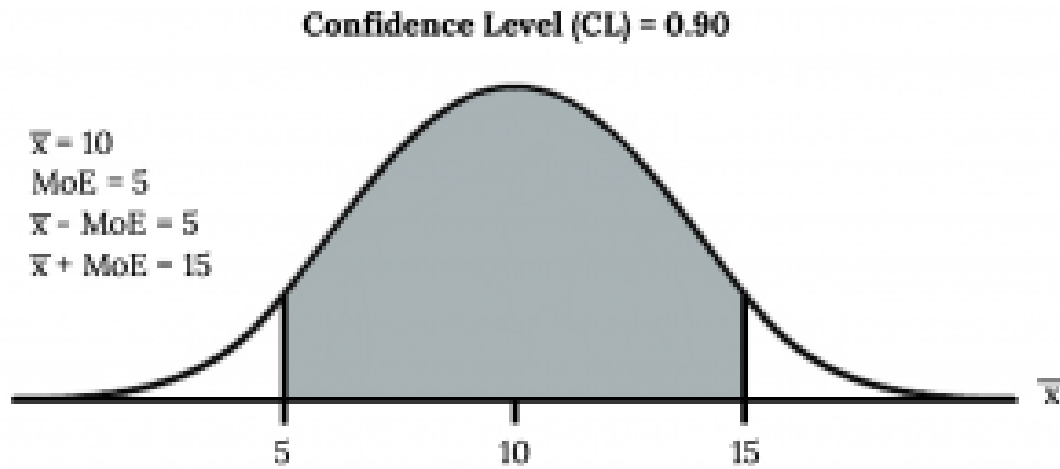


Figure 6.8: 90% Confidence Level

To capture the central 90%, we must go out 1.645 “standard deviations” on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the “standard deviation” used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the “standard error of the mean” in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ .

In summary, as a result of the central limit theorem:

- \bar{X} is normally distributed, that is, $\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the margin of error.

Finding the Critical Value

When we know the population standard deviation σ , we use a standard normal distribution to calculate the margin of error (MoE) and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$. This Z score is also called a **critical value**.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

Note: Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$

Example

Find the critical value for a 95% Confidence Interval:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=205#h5p-148>

Examples

Find the critical value for a 90% Confidence Interval

Calculating the Margin of Error (MoE)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$\text{MoE} = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$$

Constructing the Confidence Interval

A confidence interval estimate has the format:

$$(\bar{x} - MoE, \bar{x} + MoE).$$

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

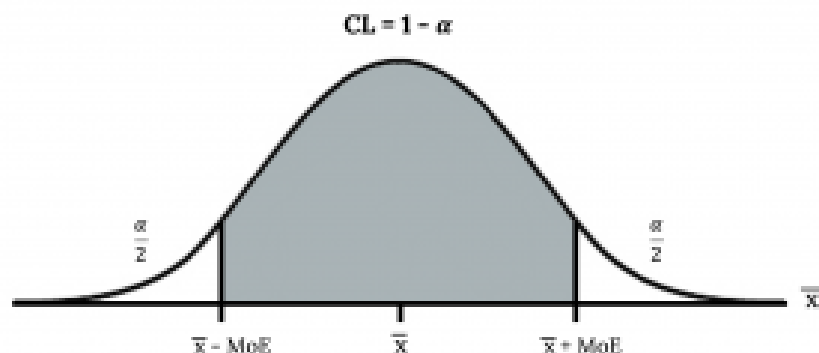


Figure 6.9: Constructing the Confidence Interval

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here the population mean), and state the confidence interval (both endpoints). “We can be ____ % confident that the interval we created, ____ to ____ captures the true population mean (include the context of the problem and appropriate units).”

Be careful that you do not associate the confidence level with the parameter itself. Your parameter is a fixed value, what is changing is the sample you take and the interval you calculate. We always want to associate the CL% with the sampling process and the interval.

Example

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a

sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

The step-by-step solution is shown below. If you are comfortable using software, you can use it to calculate the confidence interval directly.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=205#h5p-149>

Your turn!

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time and interpret.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=205#h5p-150>

Image Credits

Figure 6.7: Sebastian Gomez (2020). “Yellow green and red candies” Public domain. Retrieved from <https://unsplash.com/photos/w9pT3v9z1CM>

Figure 6.8: Kindred Grey via Virginia Tech (2020). “Figure 6.6” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.6.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 6.9: Kindred Grey via Virginia Tech (2020). “Figure 6.7” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.7.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

6.4 The Behavior of Confidence Intervals

Once we know the basics of how to calculate a **confidence interval**, we also need to know how they behave. In other words how does tweaking certain parts of the equation effect the interval? Keep in mind one of the criteria that makes something a “good” statistical estimate is precision. A smaller, or more narrow interval gives us a more precise and therefore useful estimate.

Changing the Confidence Level or Sample Size

Example

Recall the previous example:

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

The 90% confidence interval is **(67.1775, 68.8225)**.

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

To find the confidence interval, you need the sample mean, \bar{x} , and the MoE.

- $\bar{x} = 68$
- $\text{MoE} = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; $n = 36$; The confidence level is 95% ($CL = 0.95$).

$CL = 0.95$ so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$ $z_{\frac{\alpha}{2}} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

$$\text{MoE} = (1.96) \left(\frac{3}{\sqrt{36}} \right) = 0.98$$

$$\bar{X} - \text{MoE} = 68 - 0.98 = 67.02$$

$$\bar{X} + \text{MoE} = 68 + 0.98 = 68.98$$

Notice that the MoE is **larger** for a 95% confidence level in the original problem, creating a less precise interval.

Interpretation: We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Alternative Interpretation

Recall the interpretation of a CI above. Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results:

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Figure 6.10: Confidence Level Comparisons

In conclusion, increasing the confidence level increases the error bound, making the confidence interval wider.

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $\text{MoE} = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: $\text{MoE} = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Your turn!

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Calculating the Sample Size needed.

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is

$$\text{MoE} = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right).$$

The formula for sample size is $n = \frac{z^2 \sigma^2}{\text{MoE}^2}$, found by solving the error bound formula for n .

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $\text{MoE} = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2 \sigma^2}{\text{MoE}^2} = \frac{(1.96)^2 (15)^2}{2^2} = 216.09$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Example

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between \$69,720 and \$69,922.¹ Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

Use the formula for MoE, solved for n :

$$n = \frac{z^2 \sigma^2}{MoE^2}$$

From the statement of the problem, you know that $\sigma = 2.5$, and you need $MoE = 1$.

$$z = z_{0.035} = 1.812$$

(This is the value of z for which the area under the density curve to the **right** of z is 0.035.)

$$n = \frac{z^2 \sigma^2}{MoE^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$$

You need to measure at least 21 male students to achieve your goal.

Your turn!

1. American Fact Finder." U.S. Census Bureau. Available online at <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> (accessed July 2, 2013).

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Image References

Figure 6.10: Kindred Grey via Virginia Tech (2020). “Figure 6.8” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.8.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

6.5 Introduction to Hypothesis Tests

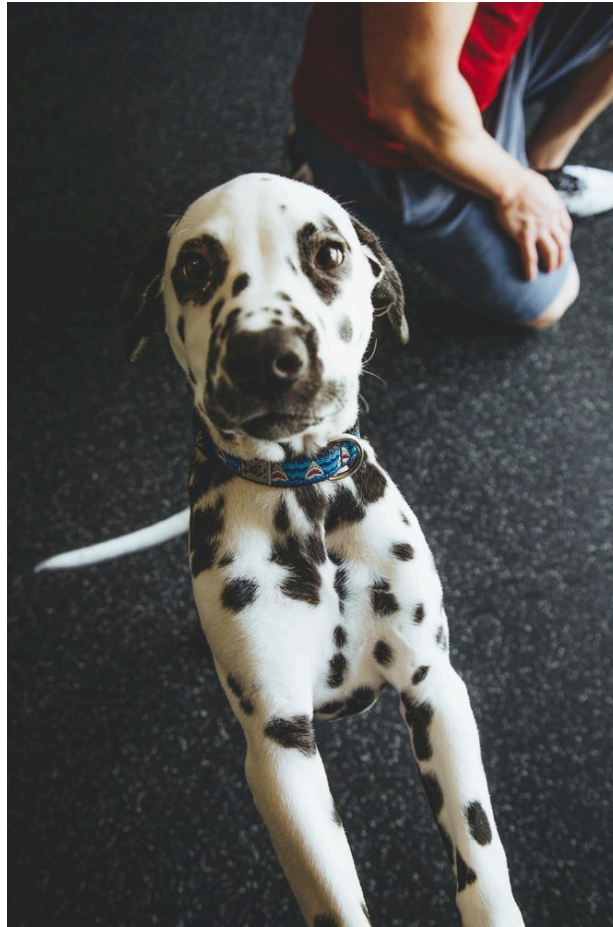


Figure 6.11: Dalmation Spots. You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter.

Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year. A statistician may want to make a decision about or evaluate these claims. A **hypothesis test** can be used to do this.

A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes

a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this section you will conduct hypothesis tests on single means when the population standard deviation is known.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will perform some variation of these steps:

1. Define hypotheses.
2. Collect and/OR use the sample data to determine the correct distribution to use.
3. Calculate Test Statistic.
4. Make a decision
5. Write a conclusion.

Defining your hypotheses

The actual test begins by considering two hypotheses. They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

The null hypothesis (H_0): It is often a statement of the accepted historical value or norm. This is your starting point that you must assume from the beginning in order to show an effect exists.

The alternative hypothesis (H_a): It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0 .

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a decision. There are two options for a decision. They are “reject H_0 ” if the sample information favors the alternative hypothesis or “do not reject H_0 ” or “decline to reject H_0 ” if the sample information is insufficient to reject the null hypothesis.

Mathematical symbols used in H_0 and H_a :

Figure 6.12: Null and Alternative Hypotheses

H_0	H_a
equal (=)	not equal (\neq) or greater than ($>$) or less than ($<$)
greater than or equal to (\geq)	less than ($<$)
less than or equal to (\leq)	more than ($>$)

Note: H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with $>$ or $<$ as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

Example

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null hypothesis is: $H_0: \mu = 2.0$. What is the alternative hypothesis?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=212#h5p-151>

Your turn!

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Using the Sample to Test the Null Hypothesis

Once you have defined your hypotheses the next step in the process, is to collect sample data. In a classroom context most of the time the data or summary statistics will be given to you.

Then you will have to determine the correct distribution to perform the hypothesis test, given the assumptions you are able to make about the situation. Right now we are demonstrating these ideas in a test for a mean when the population standard deviation is known using the Z distribution. We will see other scenarios in the future.

Calculating a Test Statistic

Next, you will start evaluating the data. This begins with calculating your **test statistic**, which is a measure of how far what you observed is from what you are assuming to be true. In this context, your test statistic, z_0 , quantifies the number of standard deviations between the sample mean \bar{x} and the population mean μ . Calculating the test statistic is analogous to standardizing observations with Z-scores as discussed previously:

$$z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

where μ_0 is the value assumed to be true in the null hypothesis.

Making a Decision

Once you have your test statistic there are two methods to use it to make your decision:

1. Critical value method – This is one way you can make a decision, but will not be discussed in detail at this time.
2. P-Value method – This is the preferred method we will focus on.

P-Value Method

To find a **p-value** we use the test statistic to calculate the actual probability of getting the test result. Formally, the p -value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p -value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p -value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p -value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a

hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

The null hypothesis could be $H_0: \mu \leq 15$

The alternate hypothesis is $H_a: \mu > 15$

The words “is more than” translates as a “>” so “ $\mu > 15$ ” goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ is known ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how unlikely the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p -value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p -value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

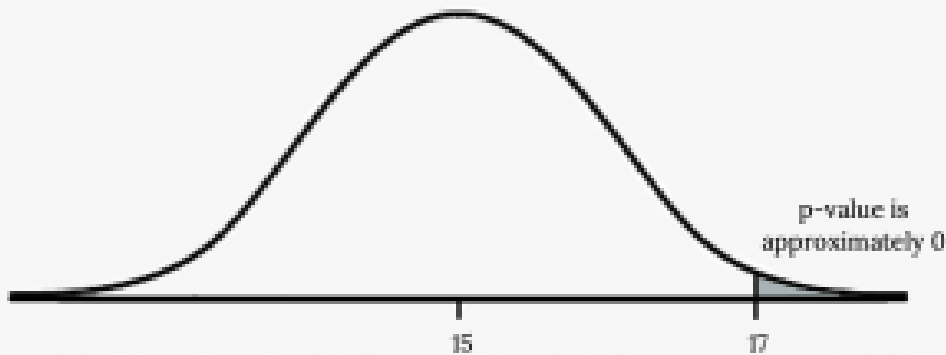


Figure 6.13: Bread Height Probability

$p\text{-value} = P(\bar{X} > 17)$ which is approximately zero.

A p -value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. purely by CHANCE had the population mean height really been 15 cm. Because the outcome of 17 cm. is so unlikely (meaning it is happening NOT by chance alone), we conclude that the evidence is strongly

against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Your turn!

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

Find The P-value:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=212#h5p-152>

Decision and conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p -value and a preset or preconceived α (also called a **significance level**). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem. If there is no given preconceived α , then use $\alpha = 0.05$.

When you make a decision to reject or not reject H_0 , do as follows:

- If $\alpha > p$ -value, reject H_0 . The results of the sample data are **statistically significant**. You can say there is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a , may be correct.
- If $\alpha \leq p$ -value, fail to reject H_0 . The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a , may be correct.

After you make your decision, write a thoughtful conclusion in the context of the scenario incorporating the hypotheses.

NOTE: When you “do not reject H_0 ”, it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

Example

When using the p -value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

If the p -value is low, the null must go.

If the p -value is high, the null must fly.

This memory aid relates a p -value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p -value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=212#h5p-153>

Your turn!

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

$$H_0: p = 0.50, H_a: p > 0.50$$

$$\alpha = 0.01$$

$$p\text{-value} = 0.025$$

Interpret the results and state a conclusion in simple, non-technical terms.

Image Credits

Figure 6.11: Alora Griffiths (2019). "Dalmation puppy near man..." Public domain. Retrieved from <https://unsplash.com/photos/7aRQZtLsvqw>

Figure 6.13: Kindred Grey via Virginia Tech (2020). "Figure 6.11" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_6.11.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

6.6 Hypothesis Tests In-Depth

Establishing the parameter of interest, type of distribution to use, the test statistic and p-value can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when interpreting the results.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very unlikely to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a \$100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a \$100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more \$100 bills in the basket. A “rare event” has occurred (Didi getting the \$100 bill) so Ali doubts the assumption about only one \$100 bill being in the basket.

Errors in Hypothesis Tests

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

Figure 6.14: Type 1 and Type 2 Errors

ACTION	H ₀ IS ACTUALLY	
	True	False
Do not reject H ₀	Correct Outcome	Type II error
Reject H ₀	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is not to reject H_0 when H_0 is true (correct decision).

2. The decision is to reject H_0 when H_0 is true (incorrect decision known as a **Type I error**).
3. The decision is not to reject H_0 when, in fact, H_0 is false (incorrect decision known as a **Type II error**).
4. The decision is to reject H_0 when H_0 is false (correct decision whose probability is called the **power** of the test).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a **Type I error** = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a **Type II error** = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

The **power of a test** is $1 - \beta$.

Ideally, α and β should be as small as possible because they are probabilities of errors, but rarely are they zero. We want a high power that is as close to one as well. Increasing the sample size can help us achieve these by reducing both α and β , and therefore increasing the power of the test.

Example

Suppose the null hypothesis, H_0 , is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error:** Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = probability that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Your turn!

Suppose the null hypothesis, H_0 , is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Statistical Significance Versus Practical Significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected, even differences where there is no practical value. In such cases, we still say the difference is **statistically significant**, but it is not practically significant.

For example, an online experiment might identify that placing additional ads on a movie review website statistically significantly increases viewership of a TV show by 0.001%, but this increase might not have any practical value.

One role of a data scientist in conducting a study often includes planning the size of the study. The data scientist might first consult experts or scientific literature to learn what would be the smallest meaningful difference from the null value. She also would obtain other information, such as a very rough estimate of the true proportion p , so that she could roughly estimate the standard error. From here, she can suggest a sample size that is sufficiently large that, if there is a real difference that is meaningful, we could detect it. While larger sample sizes may still be used, these calculations are especially helpful when considering costs or potential risks, such as possible health impacts to volunteers in a medical study.

Chapter 6 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=226#h5p-154>

Section Reviews

6.1 Point Estimation and Sampling Distributions

Since Populations are typically large and a census may not be feasible, we often use sample statistics to estimate population parameters. Some examples of point estimates are:

- \bar{x} is a point estimate for μ
- \hat{p} is a point estimate for ρ
- s is a point estimate for σ

However, we know sampling variability exists, so each statistic has its own probability distribution called a sampling distribution. In order for the statistic to be unbiased, the center of this sampling distribution should be equal to the parameter of interest (accurate), and the standard error tells us about the precision of the estimate.

6.2 Sampling Distribution of the Sample Mean

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

The Central Limit Theorem for Sample Means: $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

The Mean \bar{X} : μ_x

Central Limit Theorem for Sample Means z-score and standard error of the mean: $z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$

Standard Error of the Mean (Standard Deviation (\bar{x})): $\frac{\sigma_x}{\sqrt{n}}$

6.3 Intro to Confidence Intervals

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. A confidence interval is made up of the point estimate with a Margin of Error built in (MoE). A CI has the general form:

(lower bound, upper bound) = (point estimate - MoE, point estimate + MoE)

The calculation of the MoE depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding MoE increases as well. As the sample size increases, the MoE decreases. By the central limit theorem,

$$MoE = z \frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (MoE) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

For a Single Population Mean with known Standard Deviation we can use the Normal Distribution. assuming the CLT holds

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of "MoE where $\alpha = 1 - CL$.

6.4 Behavior of Confidence Intervals

If you want to increase your confidence level, the interval will increase in width all other things held constant. In order to make your interval smaller (more precise) you have to increase the sample size.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the MoE formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \left(\frac{z * \sigma}{MoE} \right)^2$$

6.5 Intro to Hypothesis Tests

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p -value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. $\alpha > p$ -value, reject the null hypothesis
2. $\alpha \leq p$ -value, do not reject the null hypothesis

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0 . The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality ($=$, \leq or \geq)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1 , using less than, greater than, or not equals symbols, i.e., (\neq , $>$, or $<$).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

H_0 and H_a are contradictory.

Figure 6.15

If H_0 has:	equal ($=$)	greater than or equal to (\geq)	less than or equal to (\leq)
then H_a has:	not equal (\neq) or greater than ($>$) or less than ($<$)	less than ($<$)	greater than ($>$)

If $\alpha \leq p$ -value, then do not reject H_0 .

If $\alpha > p$ -value, then reject H_0 .

α is preconceived. Its value is set before the hypothesis test starts. The p -value is calculated from the data.

6.6 Hypothesis Tests in Depth

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I error** occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β , for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

α = probability of a Type I error = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

6.1 Point Estimation and Sampling Distributions

- Statistical inference
- Sample
- Population
- Point estimation
- Hypothesis testing
- Point estimate
- Parameter
- Statistic
- Sampling variability
- Sampling distribution
- Law of large numbers

- **Standard error**

6.2 Sampling Distribution of the Sample Mean

- **Central limit theorem (CLT)**
- **Law of large numbers**
- **Standard error**

6.3 Intro to Confidence Intervals

- **Inferential statistics**
- **Population**
- **Point estimate**
- **Parameter**
- **Sampling variability**
- **Confidence interval**
- **Empirical rule**
- **Margin of error**
- **Critical value**

6.4 Behavior of Confidence Intervals

- **Confidence interval**

6.5 Intro to Hypothesis Tests

- **Hypothesis test**
- **Null hypothesis**
- **Alternative hypothesis**
- **Test statistic**
- **P-value**
- **Significance level**
- **Statistically significant**

6.6 Hypothesis Tests in Depth

- **Type I error**
- **Type II error**
- **Power**
- **Statistically significant**

Extra Practice

6.1 Point Estimation and Sampling Distributions

1. The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. The figure below shows the highest SAR level for a random selection of cell phone models as measured by the FCC.¹ Find a point estimate of the true (population) mean of the Specific Absorption Rates (SARs) for cell phones.

1. La, Lynn, Kent German. "Cell Phone Radiation Levels." part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).

Figure 6.16

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Solution:

The sample mean is: $\bar{x} = 1.024$

2. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Find a point estimate of the true (population) proportion of students in the school district who are against the new legislation.

Solution:

The sample proportion is: $\hat{p} = \frac{480}{600} = 0.8$

6.2 Sampling Distribution of the Sample Mean

1. The length of time, in hours, it takes an “over 40” group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size $n = 50$** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

- Let X = the time, in hours, it takes to play one soccer match.
- The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.
- Let \bar{X} = the mean time, in hours, it takes to play one soccer match.

a. If $\mu_X = \underline{\hspace{2cm}}$, $\sigma_X = \underline{\hspace{2cm}}$, and $n = \underline{\hspace{2cm}}$, then $X \sim N(\underline{\hspace{2cm}}, \underline{\hspace{2cm}})$ by the central limit theorem for means.

- $\mu_X = 2$, $\sigma_X = 0.5$, $n = 50$, and $X \sim N(2, \frac{0.5}{\sqrt{50}})$

b. Find $P(1.8 < \bar{X} < 2.3)$. Draw a graph.

- $P(1.8 < \bar{X} < 2.3) = 0.9977$
- $\text{normalcdf}\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$
- The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

2. The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours.² A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

3. In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years.³ Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

- What are the mean and standard deviation for the sample mean ages of tablet users?
- What does the distribution look like?
- Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
- Find the 95th percentile for the sample mean age (to one decimal place).

Solutions:

- Since the sample mean tends to target the population mean, we have $\mu_{\bar{X}} = \mu = 34$. The sample standard deviation is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$
- The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.
- The probability that the sample mean age is more than 30 is given by $P(\bar{X} > 30) = \text{normalcdf}(30, E99, 34, 1.5) = 0.9962$
- Let k = the 95th percentile.
 $k = \text{invNorm}\left(0.95, 34, \frac{15}{\sqrt{100}}\right) = 36.5$

- “2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).
- Data from The Flurry Blog, 2013. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

4. In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified.⁴ You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

- What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
 - What is the standard error of the mean?
 - Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
 - Find the probability that the sample mean is between eight minutes and 8.5 minutes.
-

5. Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\bar{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

6. Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let \bar{X} be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

- What is the mean, standard deviation, and sample size?

- Solution: mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

- Complete the distributions.

$$\begin{aligned} X &\sim \text{-----}(\text{-----}, \text{-----}) \\ \bar{X} &\sim \text{-----}(\text{-----}, \text{-----}) \end{aligned}$$

- Solution: $X \sim N(4, 1.2)$.

4. Data from The Flurry Blog, 2013. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

- Solution: $X \sim N(4, 1.216)$

c. Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.

Figure 6.17



- $P(\text{-----} < x < \text{-----}) = \text{-----}$

- Solution: 3.5, 4.25, 0.2441

d. Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.

Figure 6.18



a.

b. $P(\text{-----}) = \text{-----}$

- Solution: 0.7499

e. What causes the probabilities in C and D to be different?

- Solution: The fact that the two distributions are different accounts for the different probabilities.

f. Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.

Figure 6.19



a.

b. The 95th Percentile = _____

- Solution: $P(3.5 < x < 4.25) = \text{invNorm}(95, 4, 1.216) = 4.49$

7. Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of 88 cents. Suppose that we randomly pick 25 daytime statistics students.

- In words, $X =$ _____
- $X \sim$ _____ (_____, _____)
- In words, $\bar{X} =$ _____
- $\bar{X} \sim$ _____ (_____, _____)
- Find the probability that an individual had between \$0.80 and \$1.00. Graph the situation, and shade in the area to be determined.
- Find the probability that the average of the 25 students was between \$0.80 and \$1.00. Graph the situation, and shade in the area to be determined.
- Explain why there is a difference in part e and part f.

Solutions:

- $X =$ amount of change students carry
- $X \sim E(0.88, 0.88)$
- $\bar{X} =$ average amount of change carried by a sample of 25 students.

- d. $\bar{X} \sim N(0.88, 0.176)$
 - e. 0.0819
 - f. 0.1882
 - g. The distributions are different. Part a is exponential and part b is normal.
-

8. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

- a. If \bar{X} = average distance in feet for 49 fly balls, then $\bar{X} \sim \text{-----}(\text{-----}, \text{-----})$
- b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for \bar{X} . Shade the region corresponding to the probability. Find the probability.
- c. Find the 80th percentile of the distribution of the average of 49 fly balls.

Solutions:

- $N(250, 50/49)$
 - 0.0808
 - 256.01 feet
-

9. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

- a. In words, $X = \text{-----}$
- b. In words, $\bar{X} = \text{-----}$
- c. $\bar{X} \sim \text{-----}(\text{-----}, \text{-----})$
- d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
- e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

Solutions:

- a. length of time for an individual to complete IRS form 1040, in hours.
- b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
- c. $N(10.53, \frac{1}{3})$
- d. Yes. I would be surprised, because the probability is almost 0.
- e. No. I would not be totally surprised because the probability is 0.2312

10. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let \bar{X} the average of the 49 races.

- $\bar{X} \sim \text{-----}(\text{-----}, \text{-----})$
- Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
- Find the 80th percentile for the average of these 49 marathons.
- Find the median of the average running times.

Solutions:

- $N(145, 14/49)$
 - 0.6247
 - 146.68
 - 145 minutes
-

11. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

- In words, $X = \text{-----}$
- $X \sim \text{-----}$
- In words, $\bar{X} = \text{-----}$
- $\bar{X} \sim \text{-----}(\text{-----}, \text{-----})$
- Find the first quartile for the average song length, \bar{X} .
- The IQR (interquartile range) for the average song length, \bar{X} , is from ____ - ____.

Solutions:

- the length of a song, in minutes, in the collection
 - $U(2, 3.5)$
 - the average length, in minutes, of the songs from a sample of five albums from the collection
 - $N(2.75, 0.0660)$
 - 2.71 minutes
 - 0.09 minutes
-

12. In 1940 the average size of a U.S. farm was 174 acres.⁵ Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

5. Data from the United States Department of Agriculture.

- a. In words, $X =$ _____
- b. In words, $\bar{X} =$ _____
- c. $\bar{X} \sim$ _____ (_____, _____)
- d. The IQR for \bar{X} is from _____ acres to _____ acres.

Solutions:

- the size of a U.S. farm in 1940 the average size of a U.S. farm, in acres
 - $N(174, 55.38)$
 - 168.0, 180.0
-

13. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

- a. When the sample size is large, the mean of \bar{X} is approximately equal to the mean of X .
- b. When the sample size is large, \bar{X} is approximately normally distributed.
- c. When the sample size is large, the standard deviation of \bar{X} is approximately the same as the standard deviation of X .

Solutions:

- a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
 - b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
 - c. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.
-

14. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten.⁶ Suppose that 16 individuals are randomly chosen. Let \bar{X} = average percent of fat calories.

- a. $\bar{X} \sim$ _____ (_____, _____)
- b. For the group of 16, find the probability that the average percent of fat calories consumed is more than

6. "National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed May 17, 2013).

five. Graph the situation and shade in the area to be determined.

- c. Find the first quartile for the average percent of fat calories.

Solutions:

- $N(36, 10.16)$
 - 1
 - 34.31
-

15. The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be \$2,000 per year with a standard deviation of \$8,000. We randomly survey 1,000 residents of that country.

- a. In words, $X =$ _____
- b. In words, $\bar{X} =$ _____
- c. $\bar{X} \sim$ _____(_____, _____)
- d. How is it possible for the standard deviation to be greater than the average?
- e. Why is it more likely that the average of the 1,000 residents will be from \$2,000 to \$2,100 than from \$2,100 to \$2,200?

Solutions:

- a. X = the yearly income of someone in a third world country
 - b. the average salary from samples of 1,000 residents of a third world country
 - c. $\bar{X} \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
 - d. Very wide differences in data values can have averages smaller than standard deviations.
 - e. The distribution of the sample mean will have higher probabilities closer to the population mean.
 $P(2000 < \bar{X} < 2100) = 0.1537$
 $P(2100 < \bar{X} < 2200) = 0.1317$
-

16. Which of the following is NOT TRUE about the distribution for averages?

- a. The mean, median, and mode are equal.
- b. The area under the curve is one.
- c. The curve never touches the x-axis.
- d. The curve is skewed to the right.

Solution: d

17. The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$4.59 cents and a standard deviation of 10 cents. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations. The distribution to use for the average cost of gasoline for the 16 gas stations is:

- a. $\bar{X} \sim N(4.59, 0.10)$
- b. $\bar{X} \sim N(4.59, \frac{0.10}{\sqrt{16}})$
- c. $\bar{X} \sim N(4.59, \frac{16}{0.10})$
- d. $\bar{X} \sim N(4.59, \frac{\sqrt{16}}{0.10})$

Solution: b

6.3 Intro to Confidence Intervals

1. The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. The figure below shows the highest SAR level for a random selection of cell phone models as measured by the FCC.⁷ Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

7. La, Lynn, Kent German. "Cell Phone Radiation Levels." part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).

Figure 6.20

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

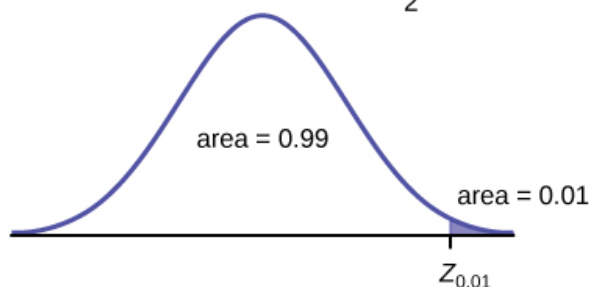
Solution:

To find the confidence interval, start by finding the point estimate: the sample mean, $\bar{x} = 1.024$. Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

$$\alpha = 1 - CL = 1 - 0.98 = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

Figure 6.21



You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01}) \frac{\sigma}{\sqrt{n}} = (2.326) \frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$$

$$\bar{x} + EBM = 1.024 + 0.1431 = 1.1671$$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

2. The figure below shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States.⁸ As previously, assume that the population standard deviation is $\sigma = 0.337$.

Figure 6.22

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

3. The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

a. Identify the following:

- \bar{x} = _____
- σ = _____
- n = _____

Solutions:

- 244
- 15
- 50

b. In words, define the random variables X and \bar{X} .

8. La, Lynn, Kent German. "Cell Phone Radiation Levels." part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).

c. Which distribution should you use for this problem?

- Solution: $N\left(244, \frac{15}{\sqrt{50}}\right)$

d. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

e. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

- Solution: As the sample size increases, there will be less variability in the mean, so the interval size decreases.

4. The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.⁹

a. Identify the following:

- \bar{x} = _____
- σ = _____
- n = _____

b. In words, define the random variables X and \bar{X} .

- Solution: X is the time in minutes it takes to complete the U.S. Census short form. \bar{X} is the mean time it took a sample of 200 people to complete the U.S. Census short form.

c. Which distribution should you use for this problem?

d. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

- Solution: CI: (7.9441, 8.4559)

9. "American Fact Finder." U.S. Census Bureau. Available online at <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> (accessed July 2, 2013).

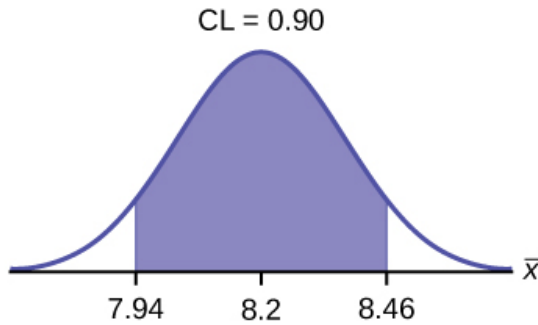


Figure 6.23

- Solution: $EBM = 0.26$

e. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

f. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

- Solution: The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

g. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

5. A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

a. Identify the following:

- $\bar{x} = \underline{\hspace{2cm}}$
- $\sigma = \underline{\hspace{2cm}}$
- $n = \underline{\hspace{2cm}}$

Solutions:

- $\bar{x} = 2.2$
- $\sigma = 0.2$
- $n = 20$

b. In words, define the random variable X .

c. In words, define the random variable \bar{X} .

- Solution: \bar{X} is the mean weight of a sample of 20 heads of lettuce.

d. Which distribution should you use for this problem?

e. Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

- Solution: EBM = 0.07
- Solution: CI: (2.1264, 2.2736)

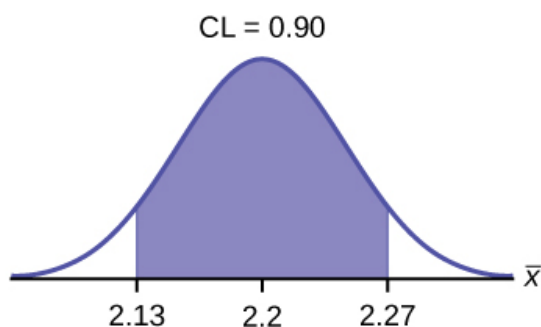


Figure 6.24

f. Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

g. In complete sentences, explain why the confidence interval in F is larger than in E.

- Solution: The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

h. In complete sentences, give an interpretation of what the interval in E means.

i. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

- Solution: The confidence level would increase.

j. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

6. The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students.¹⁰ Let X = the age of a Winter Foothill College student.

a. \bar{x} = _____

- Solution: 30.4

b. n = _____

c. _____ = 15

- Solution: σ

d. In words, define the random variable \bar{X} .

e. What is \bar{x} estimating?

- Solution: μ

f. Is σ_x known?

g. As a result of your answer to E, state the exact distribution to use when calculating the confidence interval.

- Solution: normal

h. *Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven bullet points.*

- How much area is in both tails (combined)? α = _____

- How much area is in each tail? $\frac{\alpha}{2}$ = _____

- Solution: 0.025

- Identify the following specifications:

10. "Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30,2013).

- a. lower limit
 - b. upper limit
 - c. error bound
- The 95% confidence interval is:_____.
 - Solution: (24.52,36.28)
 - Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

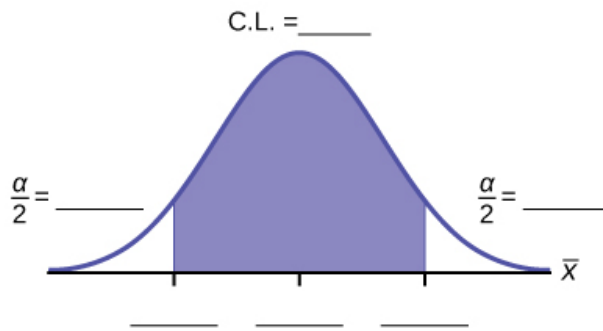


Figure 6.25

- In one complete sentence, explain what the interval means.
- Solution: We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.
- Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?
- Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?
- Solution: The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

7. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

- a. \bar{x} = _____
- b. σ = _____

c. $n =$ _____

d. In words, define the random variables X and \bar{X} .

e. Which distribution should you use for this problem? Explain your choice.

f. Construct a 95% confidence interval for the population mean height of male Swedes.

- i. State the confidence interval.
- ii. Sketch the graph.
- iii. Calculate the error bound.

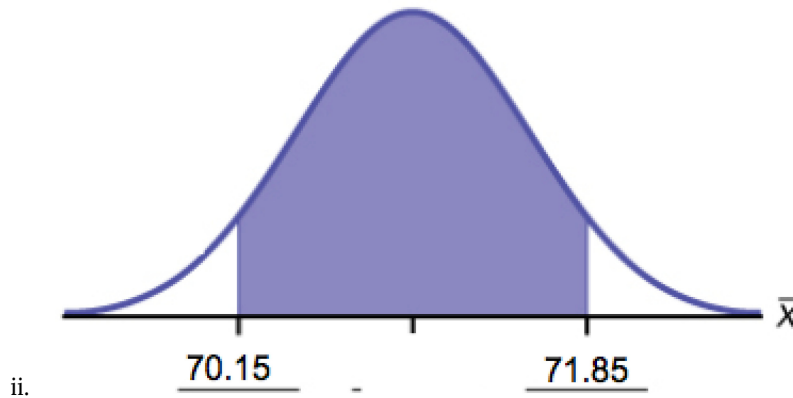
g. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

- i. 71
- ii. 3
- iii. 48

Solutions:

- a. X is the height of a Swiss male, and \bar{X} is the mean height from a sample of 48 Swiss males.
- b. Normal. We know the standard deviation for the population, and the sample size is greater than 30.
 - i. CI: (70.151, 71.49)

Figure 6.26



iii. $EBM = 0.849$

- c. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

8. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

- a. In words, define the random variables X and \bar{X} .
 - b. Which distribution should you use for this problem? Explain your choice.
 - c. Construct a 95% confidence interval for the population mean length of engineering conferences.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
-

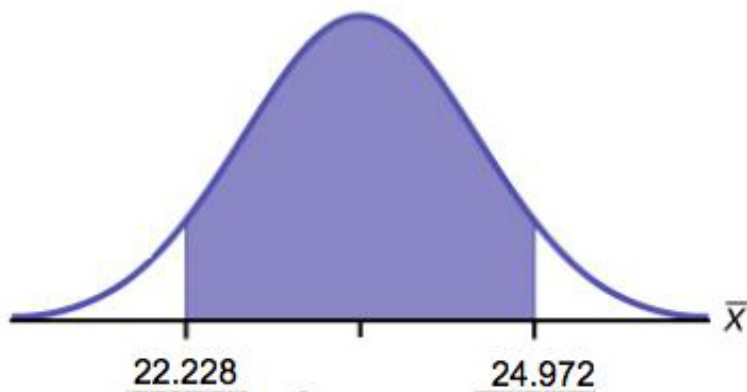
9. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

- i. \bar{x} = _____
- ii. σ = _____
- iii. n = _____
- b. In words, define the random variables X and \bar{X} .
- c. Which distribution should you use for this problem? Explain your choice.
- d. Construct a 90% confidence interval for the population mean time to complete the tax forms.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
- f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
- g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

Solutions:

- i. $\bar{x} = 23.6$
- ii. $\sigma = 7$
- iii. $n = 100$
- b. X is the time needed to complete an individual tax form. \bar{X} is the mean time to complete tax forms from a sample of 100 customers.
- c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.
 - i. (22.228, 24.972)

Figure 6.27



ii.

iii. EBM = 1.372

- d. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
- e. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
- f. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

10. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

- i. \bar{x} = _____
- ii. σ = _____
- iii. s_x = _____
- b. In words, define the random variable X .
- c. In words, define the random variable \bar{X} .
- d. Which distribution should you use for this problem? Explain your choice.
- e. Construct a 90% confidence interval for the population mean weight of the candies.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- f. Construct a 98% confidence interval for the population mean weight of the candies.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval

in part e.

- h. In complete sentences, give an interpretation of what the interval in part f means.

11. A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

- i. \bar{x} = _____
- ii. σ = _____
- iii. n = _____
- b. Define the random variables X and \bar{X} in words.
- c. Which distribution should you use for this problem? Explain your choice.
- d. Construct a 90% confidence interval for the population mean number of letters campers send home.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

Solutions:

- i. 7.9
- ii. 2.5
- iii. 20
- b. X is the number of letters a single camper will send home. \bar{X} is the mean number of letters sent home from a sample of 20 campers.
- c. $N7.9\left(\frac{2.5}{\sqrt{20}}\right)$
 - i. CI: (6.98, 8.82)

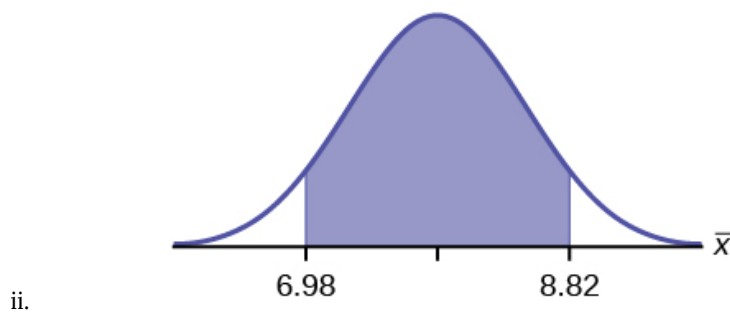


Figure 6.28

iii. EBM: 0.92

d. The error bound and confidence interval will decrease.

12. What is meant by the term “90% confident” when constructing a confidence interval for a mean?

- a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
 - b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
 - c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
 - d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.
-

13. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. The figure below shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest \$100. The standard deviation for this data to the nearest hundred is $\sigma = \$909,200$.

Figure 6.29

\$3,600	\$1,243,900	\$10,900	\$385,200	\$581,500
\$7,400	\$2,900	\$400	\$3,714,500	\$632,500
\$391,000	\$467,400	\$56,800	\$5,800	\$405,200
\$733,200	\$8,000	\$468,700	\$75,200	\$41,000
\$13,300	\$9,500	\$953,800	\$1,113,500	\$1,109,300
\$353,900	\$986,100	\$88,600	\$378,200	\$13,200
\$3,800	\$745,100	\$5,800	\$3,072,100	\$1,626,700
\$512,900	\$2,309,200	\$6,600	\$202,400	\$15,800

- a. Find the point estimate for the population mean.
- b. Using 95% confidence, calculate the error bound.
- c. Create a 95% confidence interval for the mean total individual contributions.
- d. Interpret the confidence interval in the context of the problem.

Solution:

- $\bar{x} = \$568,873$
 - $CL = 0.95 \quad \alpha = 1 - 0.95 = 0.05 \quad z_{\frac{\alpha}{2}} = 1.96$
 $EBM = z_{0.025} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{909,200}{\sqrt{40}} = \$281,764$
 - $\bar{x} - EBM = 568,873 - 281,764 = 287,109$
 $\bar{x} + EBM = 568,873 + 281,764 = 850,637$
 - We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.
-

6.4 Behavior of Confidence Intervals

- Refer back to the pizza-delivery exercise:

“Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 min.”

The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

- Suppose we change the original problem in the pizza-delivery exercise to see what happens to the error bound if the sample size is changed. Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; The confidence level is 90% ($CL=0.90$); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution A: If we **increase** the sample size n to 100, we **decrease** the error bound.

$$\text{When } n = 100: EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.645) \left(\frac{3}{\sqrt{100}} \right) = 0.4935.$$

Solution B: If we **decrease** the sample size n to 25, we **increase** the error bound.

$$\text{When } n = 25: EBM = \left(z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.645) \left(\frac{3}{\sqrt{25}} \right) = 0.987.$$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
 - Decreasing the sample size causes the error bound to increase, making the confidence interval wider
-

3. Refer back to the pizza-delivery exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

6.5 Intro to Hypothesis Tests

1. When do you reject the null hypothesis?

2. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

- Solution: The outcome of winning is very unlikely.
-

3. The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

4. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p -value is almost zero. State the null and alternative hypotheses and interpret the p -value.

$$H_0: \mu \geq 73$$

$$H_a: \mu < 73$$

The p -value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

5. The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p -value is 0.0264. State the null and alternative hypotheses and interpret the p -value.

6. Does the shaded region represent a low or a high p -value compared to a level of significance of 1%?

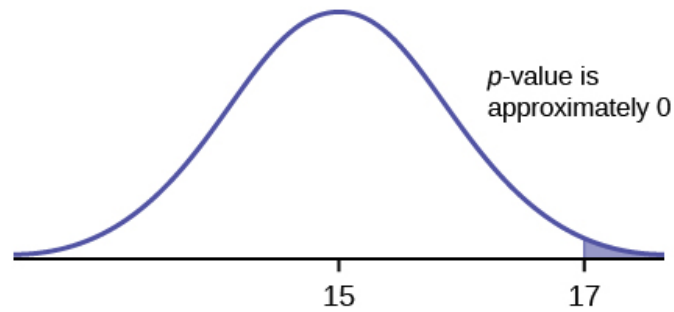


Figure 6.30

- Solution: The shaded region shows a low p -value.

7. What should you do when $\alpha > p$ -value?

8. What should you do if $\alpha = p$ -value?

- Solution: Do not reject H_0 .

9. If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

10. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

a. Is this a test of means or proportions?

- means

b. What symbol represents the random variable for this test?

c. In words, define the random variable for this test.

- the mean time spent in jail for 26 first time convicted burglars

d. Is σ known and, if so, what is it?

e. Calculate the following:

a. \bar{x} _____

b. σ _____

c. s_x _____

d. n _____

- 3
- 1.5
- 1.8
- 26

f. Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

g. State the distribution to use for the hypothesis test.

- $\bar{X} \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

11. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?

b. State the null and alternative hypotheses.

H_0 : _____ H_a : _____

c. Is this a right-tailed, left-tailed, or two-tailed test?

d. What symbol represents the random variable for this test?

e. In words, define the random variable for this test.

f. Is the population standard deviation known and, if so, what is it?

g. Calculate the following:

i. \bar{x} = _____

ii. s = _____

iii. n = _____

- h. Which test should be used?
 - i. State the distribution to use for the hypothesis test.
 - j. Find the p -value.
 - k. At a pre-conceived $\alpha = 0.05$, what is your:
 - i. Decision:
 - ii. Reason for the decision:
 - iii. Conclusion (write out in a complete sentence):
-

12. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness.¹¹ Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

- a. Is this a test of one mean or proportion?
 - b. State the null and alternative hypotheses.
 H_0 : _____ H_a : _____
 - c. Is this a right-tailed, left-tailed, or two-tailed test?
 - d. What symbol represents the random variable for this test?
 - e. In words, define the random variable for this test.
 - f. Calculate the following:
 - i. \bar{x} = _____
 - ii. n = _____
 - iii. p' = _____
 - g. Calculate σ_x = _____. Show the formula set-up.
 - h. State the distribution to use for the hypothesis test.
 - i. Find the p -value.
 - j. At a pre-conceived $\alpha = 0.05$, what is your:
 - i. Decision:
 - ii. Reason for the decision:
 - iii. Conclusion (write out in a complete sentence):
-

13. We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol ($=$, \neq , \geq , $<$, \leq , $>$) for the null and alternative hypotheses.

11. Data from the National Institute of Mental Health. Available online at <http://www.nimh.nih.gov/publicat/depression.cfm>.

- a. $H_0: \mu __ 66$
b. $H_a: \mu __ 66$
-

14. We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

$H_0: \mu \geq 5$
 $H_a: \mu < 5$

15. We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol ($=$, \neq , \geq , $<$, \leq , $>$) for the null and alternative hypotheses.

- a. $H_0: \mu __ 45$
b. $H_a: \mu __ 45$
-

16. In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

$H_0: p \leq 0.066$
 $H_a: p > 0.066$

17. On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol ($=$, \neq , \geq , $<$, \leq , $>$) for the null and alternative hypotheses.

- a. $H_0: p __ 0.40$
b. $H_a: p __ 0.40$
-

18. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

- The random variable is the mean Internet speed in Megabits per second.
-

19. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

20. The American family has an average of two children. What is the random variable? Describe in words.

- The random variable is the mean number of children an American family has.
-

21. The mean entry level salary of an employee at a company is \$58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

22. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

- The random variable is the proportion of people picked at random in Times Square visiting the city.
-

23. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

24. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

- $H_0: p = 0.42$
 - $H_a: p < 0.42$
-

25. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

- a. H_0 : _____
b. H_a : _____
-

26. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

- a. H_0 : _____
b. H_a : _____

Solutions:

- $H_0: \mu = 15$
 - $H_a: \mu \neq 15$
-

27. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness.¹² Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

- a. H_0 : _____
b. H_a : _____
-

28. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis. State the null hypothesis, H_0 , and the alternative hypothesis, H_a , in terms of the appropriate parameter (μ or p).

- The mean number of years Americans work before retiring is 34.
- At most 60% of Americans vote in presidential elections.
- The mean starting salary for San Jose State University graduates is at least \$100,000 per year.
- Twenty-nine percent of high school seniors get drunk each month.
- Fewer than 5% of adults ride the bus to work in Los Angeles.
- The mean number of cars a person owns in her lifetime is not more than ten.
- About half of Americans prefer to live away from cities, given the choice.

12. Data from the National Institute of Mental Health. Available online at <http://www.nimh.nih.gov/publicat/depression.cfm>.

- h. Europeans have a mean paid vacation each year of six weeks.
- i. The chance of developing breast cancer is under 11% for women.
- j. Private universities' mean tuition cost is more than \$20,000 per year.

Solutions:

- a. $H_0: \mu = 34$; $H_a: \mu \neq 34$
 - b. $H_0: p \leq 0.60$; $H_a: p > 0.60$
 - c. $H_0: \mu \geq 100,000$; $H_a: \mu < 100,000$
 - d. $H_0: p = 0.29$; $H_a: p \neq 0.29$
 - e. $H_0: p = 0.05$; $H_a: p < 0.05$
 - f. $H_0: \mu \leq 10$; $H_a: \mu > 10$
 - g. $H_0: p = 0.50$; $H_a: p \neq 0.50$
 - h. $H_0: \mu = 6$; $H_a: \mu \neq 6$
 - i. $H_0: p \geq 0.11$; $H_a: p < 0.11$
 - j. $H_0: \mu \leq 20,000$; $H_a: \mu > 20,000$
-

29. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

- a. $p < 0.30$
 - b. $p \leq 0.30$
 - c. $p \geq 0.30$
 - d. $p > 0.30$
-

30. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

- a. $p = 0.20$
- b. $p > 0.20$
- c. $p < 0.20$
- d. $p \leq 0.20$

Solution: c

31. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

- a. $H_0: \bar{x} = 4.5, H_a: \bar{x} > 4.5$
 - b. $H_0: \mu \geq 4.5, H_a: \mu < 4.5$
 - c. $H_0: \mu = 4.75, H_a: \mu > 4.75$
 - d. $H_0: \mu = 4.5, H_a: \mu > 4.5$
-

6.6 Hypothesis Tests in Depth

1. Suppose the null hypothesis, H_0 , is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error:** The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = $P(\text{Type I error})$. β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.

Which is the error with the greater consequence?

Solution: Type I error (If the emergency crew thinks the victim is dead, they will not treat him.)

2. Suppose the null hypothesis, H_0 , is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

3. It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0 , is: It's a Boy Genetic Labs has no effect on gender outcome.

Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α .

Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β .

What is the error of greater consequence?

Solution: Type I error (couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy)

4. "Red tide" is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

5. A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

Solution: In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

6. Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0 , that states the percentage of adults with jobs is at least 88%.

7. Identify the Type I and Type II errors from these four statements.

- Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
 - Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
 - Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
 - Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.
-

8. The mean price of mid-sized cars in a region is \$32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Solution:

Type I: The mean price of mid-sized cars is \$32,000, but we conclude that it is not \$32,000.

Type II: The mean price of mid-sized cars is not \$32,000, but we conclude that it is \$32,000.

9. A sleeping bag is tested to withstand temperatures of -15°F . You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

10. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0 , is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences. Which is the error with the greater consequence?

Solution:

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

11. The power of a test is 0.981. What is the probability of a Type II error?

0.019

12. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0 , is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

13. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0 , is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

- 0.998
-

14. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0 , is: the sample contains E-coli. Which is the error with the greater consequence?

15. State the Type I and Type II errors in complete sentences given the following statements.

- a. The mean number of years Americans work before retiring is 34.
- b. At most 60% of Americans vote in presidential elections.
- c. The mean starting salary for San Jose State University graduates is at least \$100,000 per year.
- d. Twenty-nine percent of high school seniors get drunk each month.
- e. Fewer than 5% of adults ride the bus to work in Los Angeles.
- f. The mean number of cars a person owns in his or her lifetime is not more than ten.
- g. About half of Americans prefer to live away from cities, given the choice.
- h. Europeans have a mean paid vacation each year of six weeks.
- i. The chance of developing breast cancer is under 11% for women.
- j. Private universities mean tuition cost is more than \$20,000 per year.

Solutions:

- a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
- b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%. Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
- c. Type I error: We conclude that the mean starting salary is less than \$100,000, when it really is at least \$100,000. Type II error: We conclude that the mean starting salary is at least \$100,000 when, in fact, it is less than \$100,000.
- d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
- e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer than 5% do.
- f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
- g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
- h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.

- i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
 - j. Type I error: We conclude that the average tuition cost at private universities is more than \$20,000, though in reality it is at most \$20,000. Type II error: We conclude that the average tuition cost at private universities is at most \$20,000 when, in fact, it is more than \$20,000.
-

16. For statements a-j in the previous question answer the following in complete sentences.

- a. State a consequence of committing a Type I error.
 - b. State a consequence of committing a Type II error.
-

17. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

- a. To conclude the drug is safe when in, fact, it is unsafe.
- b. Not to conclude the drug is safe when, in fact, it is safe.
- c. To conclude the drug is safe when, in fact, it is safe.
- d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

Solution: b

18. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

- a. at least 20%, when in fact, it is less than 20%.
 - b. 20%, when in fact, it is 20%.
 - c. less than 20%, when in fact, it is at least 20%.
 - d. less than 20%, when in fact, it is less than 20%.
-

19. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean

of 7.24 hours with a standard deviation of 1.93 hours.¹³ At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

- a. is more than seven hours.
- b. is at most seven hours.
- c. is at least seven hours.
- d. is less than seven hours.

Solution: d

20. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

- a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
- b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
- c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
- d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

References

Image References

Figure 6.17: Figure 7.16 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/7-practice>

13. King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

Figure 6.18: Figure 7.17 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/7-practice>

Figure 6.19: Figure 7.17 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/7-practice>

Figure 6.21: Figure 8.4 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-1-a-single-population-mean-using-the-normal-distribution>

Figure 6.23: Figure 8.12 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/8-solutions#eip-773-solution>

Figure 6.24: Figure 8.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/8-solutions#eip-773-solution>

Figure 6.25: Figure from Lumen Learning Introduction to Statistics (CC BY 4.0). Retrieved from <https://courses.lumenlearning.com/introstats1/chapter/section-exercises-7/>

Figure 6.26: Figure 8.18 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/8-solutions#eip-773-solution>

Figure 6.27: Figure 8.19 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/8-solutions#eip-773-solution>

Figure 6.28: Figure 8.20 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/8-solutions#eip-773-solution>

Figure 6.30: Figure 9.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/9-practice>

Text

Baran, Daya. “20 Percent of Americans Have Never Used Email.” WebGuild, 2010. Available online at <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email> (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

Data from the United States Department of Agriculture.

Image credit: comedy_nose/flickr

“American Fact Finder.” U.S. Census Bureau. Available online at <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> (accessed July 2, 2013).

“Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).

“Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall.” Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30, 2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. “2000 CDC Growth Charts for the

United States: Methods and Development.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/growthcharts/2000growthchart-us.pdf> (accessed July 2, 2013).

La, Lynn, Kent German. “Cell Phone Radiation Levels.” c|net part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).

“Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates.” American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

“Metadata Description of Candidate Summary File.” U.S. Federal Election Commission. Available online at <http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml> (accessed July 2, 2013).

“National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed July 2, 2013).

(Credit: Robert Neff)

Data from the National Institute of Mental Health. Available online at <http://www.nimh.nih.gov/publicat/depression.cfm>.

CHAPTER 7: INFERENCE FOR ONE SAMPLE

7.1 The Sampling Distribution of the Sample Mean (σ Un-known)

Learning Objectives

By the end of this chapter, the student should be able to:

- Construct and interpret confidence intervals for means when the population standard deviation is unknown
- Carry out hypothesis tests for means when the population standard deviation is unknown
- Construct and interpret confidence intervals for a proportion
- Understand the behavior of confidence intervals for a proportion
- Carry out hypothesis tests for a proportion

We have discussed the **sampling distribution** of the sample mean when the population standard deviation, σ , is known. However, in practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size can cause inaccuracies in the confidence interval.

Student's t Distribution



Figure 7.1: William Gosset (Student). William Sealy Gosset wrote under the pseudonym, “Student,” so that readers would not know he was a scientist at Guinness Brewery.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not always produce accurate results when he tried to use existing inference techniques. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This is because s is a more reliable estimate of σ as samples get bigger. This problem led him to “discover” what is called **Student's t -distribution**. The name “Student's T distribution” comes from the fact that Gosset wrote under the pen name “Student.”

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and used the Student's t -distribution only for sample sizes of at most 30. In our current age of technology, the accepted practice now is to simply use the Student's t -distribution whenever s is used as an estimate for σ .

In summary, if you draw a simple random sample of size n from a population that has an approximately normal

distribution with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a Student's t-distribution with $n - 1$ degrees of freedom. The t-score has the same interpretation as the z-score. It measures how far \bar{x} is from its mean μ . For each sample size n , there is a different Student's t-distribution.

The following images compare the Z (Standard Normal) and t (Student's T). What differences do you notice?

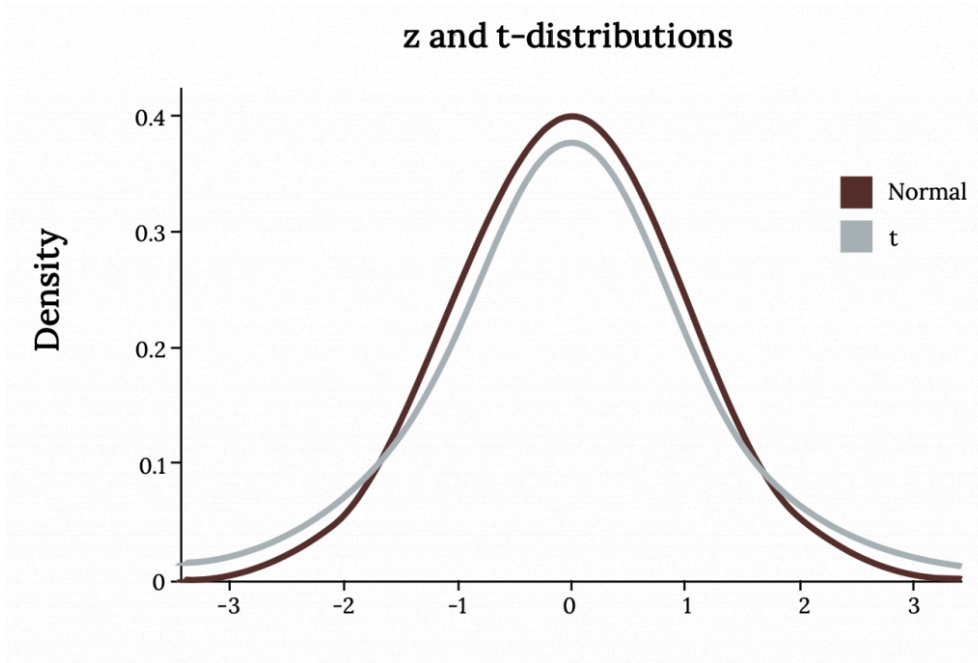


Figure 7.2: Comparing the Standard Normal Distribution and Student's T Distribution

Degrees of Freedom

The **degrees of freedom**, $n - 1$, come from the calculation of the sample standard deviation s . Remember when we calculated a sample standard deviation we divided the sum of the squared deviations by $n - 1$, but we used n deviations $(x - \bar{x})$ to calculate s . Because the sum of the deviations is zero, we can find the last deviation once we know the other $n - 1$ deviations. The other $n - 1$ deviations can change or vary freely. We call the number $n - 1$ the degrees of freedom (df).

For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

The following image shows what happens to the t distribution as you change the degrees of freedom. What happens as the df increases? What happens once n becomes around 30 and how does that relate to what you already know about the CLT?

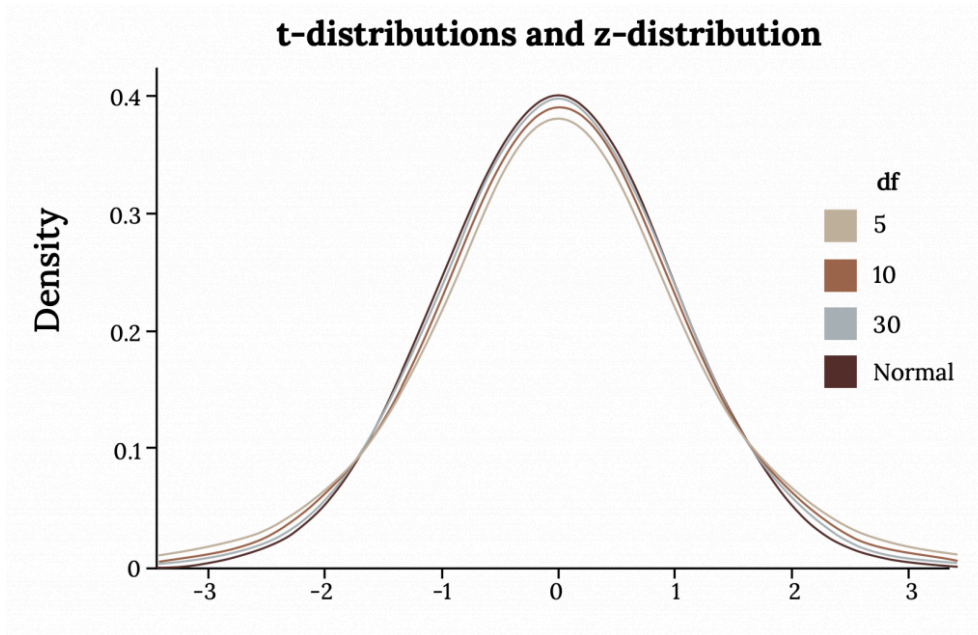


Figure 7.3: T Distribution with Different Degrees of Freedom

Properties of the Student's t-Distribution

To summarize the properties of the t distribution:

- The graph for the Student's t-distribution is similar to the standard normal curve, in that it is symmetric about a mean of zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.
- The notation for the Student's t-distribution (using T as the random variable) is $T \sim t_{df}$ where $df = n - 1$.

Example

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Plots of the data show no skewness or outliers. Which distribution is appropriate to use here?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=232#h5p-155>

Your turn!

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects and plots of the data show no skewness or outliers. Which distribution is appropriate to use here?

Finding T Distribution Probabilities

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that most t tables gives t-scores given the degrees of freedom and the right-tailed probability.

You'll find that the t table adequate for finding critical values, but is very limited when trying to find **p-values**. Calculators and computers can easily calculate any Student's t-probabilities.

Image Credits

Figure 7.1: Phillip Glickman (2019). Public domain. Retrieved from <https://unsplash.com/photos/4wnZbnW9Bv0>

Figure 7.2: Kindred Grey via Virginia Tech (2021). “Comparing the Standard Normal Distribution and Student’s T Distribution.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Comparing_the_Standard_Normal_Distribution_and_Student%27s_T_Distribution.png

Figure 7.3: Kindred Grey via Virginia Tech (2021). “T Distribution with Different Degrees of Freedom.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:T_Distribution_with_Different_Degrees_of_Freedom.png

7.2 Inference for the Mean in Practice

We have discussed the sampling distribution of the sample mean follows a normal distribution when the population standard deviation, σ , is known and the t distribution when it is not. In practice, we rarely know the population standard deviation. For larger samples we can typically get away with using Z according to the CLT. In summary, the majority of the time we opt to use t is when we do not know σ and we have a small sample ($n < 30$)

Confidence Intervals for the Mean (σ Unknown)

The general format of a **confidence interval** is:

$$(PE - MoE, PE + MoE)$$

The population **parameter** is μ . The **point estimate (PE)** for μ is \bar{x} , the sample mean.

If the population standard deviation is not known, the **margin of error (MoE)** for a population mean is:

- $MoE = \left(t_{\frac{\alpha}{2}}\right) \left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

Example

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.¹

The FEC has reported financial information for 556 Leadership PACs that operated during the

1. "Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).

2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs. (In dollars)

Figure 7.4: PAC Receipt Data

\$46,500.00	\$0	\$40,966.50	\$105,887.20	\$5,175.00
\$29,050.00	\$19,500.00	\$181,557.20	\$31,500.00	\$149,970.80
\$2,555,363.20	\$12,025.00	\$409,000.00	\$60,521.70	\$18,000.00
\$61,810.20	\$76,530.80	\$119,459.20	\$0	\$63,520.00
\$6,500.00	\$502,578.00	\$705,061.10	\$708,258.90	\$135,810.00
\$2,000.00	\$2,000.00	\$0	\$1,287,933.80	\$219,148.30

$$\bar{x} = \$251,854.23$$

$$s = \$521,130.41$$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

Note that we are not given the population standard deviation, only the standard deviation of the sample.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=234#h5p-156>

Your turn!

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in the figure below. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

Figure 7.5: Student TV Data

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Hypothesis Tests for the Mean (σ Unknown)

Remember, we will use the **t-distribution** when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.

If you are testing a single population mean, and we decide to use t, the steps say the same, but our test statistic will change slightly.

$$t = \frac{\bar{x} - \mu_o}{\left(\frac{s}{\sqrt{n}}\right)}$$

You should have no problem using technology to find p-values associated with a t test statistic. However, if you want to use your t table you'll find it is somewhat limited in finding exact p-values. Despite that you can still estimate a range of values for your p-val and then compare it to your significance level.

Examples

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores below:

65, 65, 70, 67, 66, 63, 63, 68, 72, 71

Perform the hypothesis test using a 5% level of significance to test the instructor's claim.



An interactive H5P element has been excluded from this version of the text. You can view it

online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=234#h5p-157>

Examples

It is believed that a stock price for a particular company will grow at a rate of \$5 per week. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows:

\$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2

Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p -value, state your conclusion, and identify the Type I and Type II errors.

Summary of Assumptions

When you perform inference on a single population mean μ using a Student's t -distribution (often called a t -test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t -test will work even if the population is not approximately normally distributed).

When you perform a hypothesis test of a single population mean μ using a normal distribution (often called a z -test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

7.3 The Sampling Distribution of the Sample Proportion

We have now talked at length about the basics of inference on the mean of quantitative data. What if the variable we are interested in is categorical? We cannot calculate means, variances, and the like for categorical data. However, we can count the number of individuals that have a characteristic we are interested in and divide by the total number in our population to get the **population proportion (p)**.

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a point estimate of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the parameter of interest. When the parameter is a proportion, it is often denoted by p , and we often refer to the **sample proportion** as \hat{p} (pronounced “p-hat”). Unless we collect responses from every individual in the population, p remains unknown, and we use \hat{p} as our estimate of p . The difference we observe from the poll versus the parameter is called the error in the estimate.

Understanding the Variability of a Proportion

Suppose we know the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%? We want to understand, how does the sample proportion, \hat{p} , behave when the true population proportion is 0.88. We can simulate responses we would get from a simple random sample of 1000 American adults, which is only possible because we know the actual support for expanding solar energy is 0.88. Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write “support” on 88% of them and “not” on the other 12%.
2. Mix up the pieces of paper and pull out 1000 pieces to represent our sample of 1000 American adults.
3. Compute the fraction of the sample that say “support”.

Any volunteers to conduct this simulation? Probably not. Running this simulation with 250 million pieces of paper would be time-consuming and very costly, but we can simulate it using technology. In this simulation, one sample gave a point estimate of $\hat{p}_1 = 0.894$. We know the population proportion for the simulation was $p = 0.88$, so we know the estimate had an error of $0.894 - 0.88 = +0.014$. One simulation isn't enough to get a great sense of the distribution of estimates we might expect in the simulation, so we should run more simulations. In a second simulation, we get $\hat{p}_2 = 0.885$, which has an error of $+0.005$. In another, $\hat{p}_3 = 0.878$ for an error of -0.002 . And in another, an estimate of $\hat{p}_4 = 0.859$ with an error of -0.021 . With the help of a computer, we've run the simulation 10,000 times and created a histogram of the results from all 10,000 simulations in the following figure:

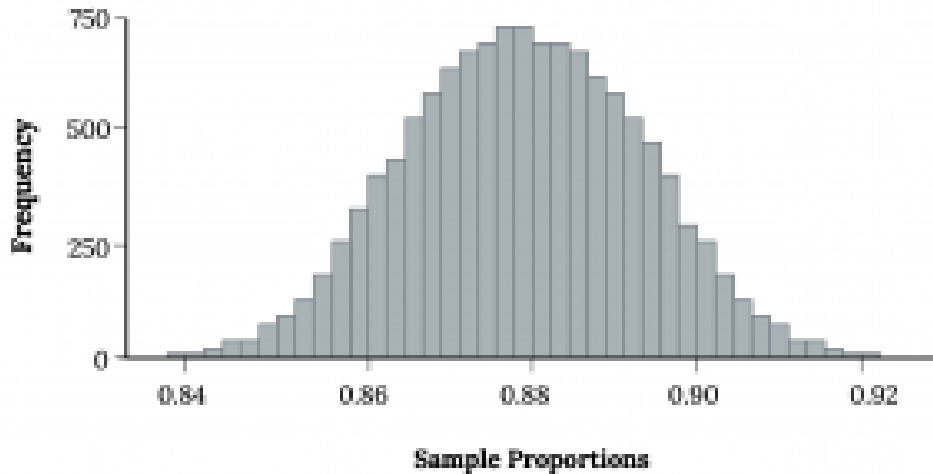


Figure 7.6: Histogram From Simulation

This simulates the sampling distribution of the sample proportion. We can characterize this sampling distribution as follows:

- **Center:** The center of the distribution is $\bar{x}_{\hat{p}} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.
- **Spread:** The standard deviation of the distribution is $s_{\hat{p}} = 0.010$. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term standard error rather than standard deviation, and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.
- **Shape:** The distribution is symmetric and bell-shaped, and it resembles a normal distribution.

When the population proportion is $p = 0.88$ and the sample size is $n = 1000$, the sample proportion \hat{p} looks to give an unbiased estimate of the population proportion and resembles a normal distribution. It looks as if we can apply the **central limit theorem** here too under the following conditions.

Conditions for the CLT for p

When observations are independent and the sample size is sufficiently large, the sample proportion \hat{p} will tend to follow a normal distribution with parameters:

- $\mu_{\hat{p}} = p$

$$\bullet \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$ *. Hopefully you see some similarity here to the normal approximation to the binomial which is one of the underlying theories here.

*Note: Some resources may use 5 here but 10 is safer.

What if we do not meet these conditions? Consider the following distributions and see if any patterns emerge:

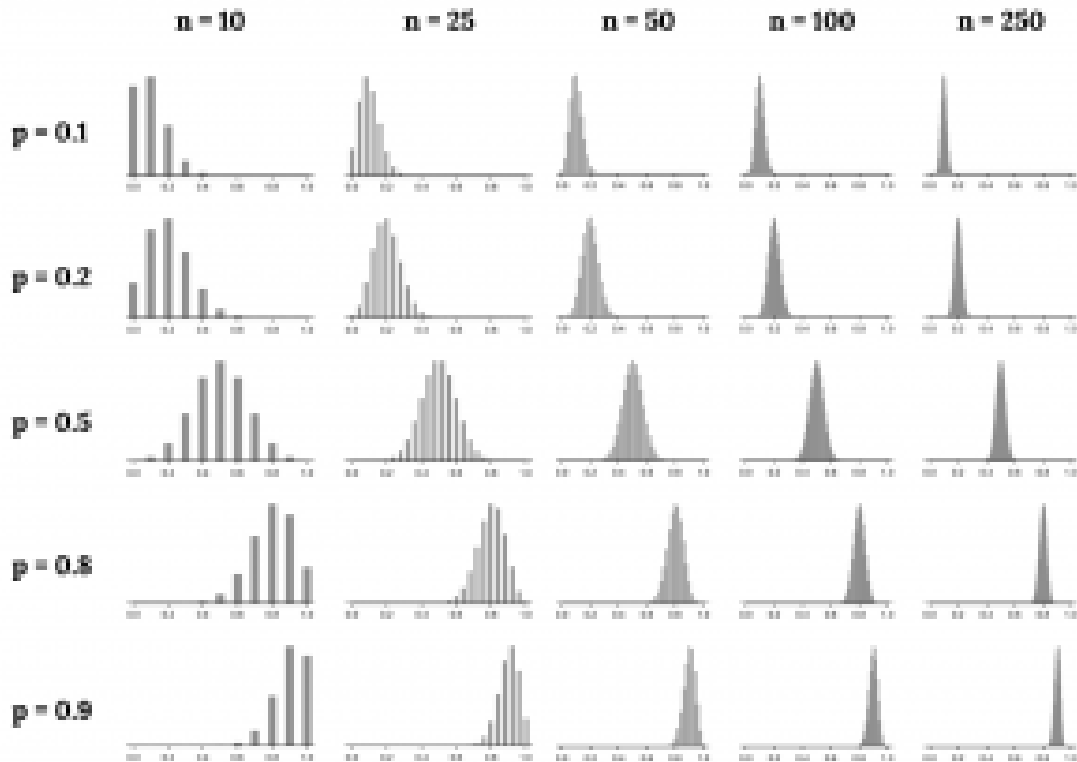


Figure 7.7: Same Size Conditions

From these distributions we can see some patterns:

1. When either np or $n(1-p)$ is small, the distribution is more discrete, i.e. not continuous.
2. When np or $n(1-p)$ is smaller than 10, the skew in the distribution is more noteworthy.
3. The larger both np and $n(1-p)$, the more normal the distribution. This may be a little harder to see for the larger sample size in these plots as the variability also becomes much smaller.
4. When np and $n(1-p)$ are both very large, the distribution's discreteness is hardly evident, and the distribution looks much more like a normal distribution.

In regards to how the mean and standard error of the distributions change:

1. The centers of the distribution are always at the population proportion, p , that was used to generate the simulation. Because the sampling distribution of \hat{p} is always centered at the population parameter p , it means the sample proportion \hat{p} is unbiased when the data are independent and drawn from such a population.
2. For a particular population proportion p , the variability in the sampling distribution decreases as the sample size n becomes larger. This will likely align with your intuition: an estimate based on a larger sample size will tend to be more accurate.
3. For a particular sample size, the variability will be largest when $p = 0.5$. The differences may be a little subtle, so take a close look. This reflects the role of the proportion p in the standard error formula. The standard error is largest when $p = 0.5$.

At no point will the distribution of \hat{p} look perfectly normal, since \hat{p} will always be take discrete values (x/n). It is always a matter of degree, and we will use the standard success-failure condition with minimums of 10 for np and $n(1 - p)$ as our guideline within this book.

Image References

Figure 7.6: Kindred Grey (2020). “Figure 7.4.” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_7.4.png

Figure 7.7: Kindred Grey via Virginia Tech (2020). “Figure 7.5” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_7.5.png . Adaptation of Figures 5.4 and 5.5 from OpenIntro Introductory Statistics (2019) (CC BY-SA 3.0). Retrieved from <https://www.openintro.org/book/os/>

7.4 Inference for a Proportion

If we are working with **categorical data** our **parameter** of interest is often the **population proportion**, p . The **point estimate** for p is $\hat{p} = \frac{x}{n}$ where x is the number of successes and n is the sample size. It is also sometimes denoted as p' . We saw previously that if we meet conditions, $np \geq 10$ and $n(1 - p) \geq 10$, we can apply the **central limit theorem** and assume:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

How do you know you are dealing with a proportion problem? First, the underlying distribution is a **binomial distribution**. This will be categorical data with no mention of a mean or average. If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success.

Hypothesis Tests for p

When you perform a **hypothesis test** of a single population proportion p , the steps are exactly the same as what we have seen before, however we will calculate our Test Statistic differently. When conducting a test for p , our hypotheses will look as follows:

- $H_0: p = p_0$
- $H_a: p (<, >, \neq) p_0$

Recall, the general form of a **test statistic** is:

$$Z = \frac{\text{point estimate} - \text{null value}}{\text{SE}}$$

For the normal distribution of proportions, the z-score formula is as follows:

If $\hat{p} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ then the z-score formula is:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Intuitively, you might think we use this as our test statistic but remember two things:

1. We do not actually know p
2. In a hypothesis test we begin by assuming the null is true

Sure to these facts, we substitute in p_0 for p in the standard error which gives us:

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

We then can find a p-value and make our decision as normal

Example

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=240#h5p-158>

You Try It

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p -value, state your conclusion, and identify the Type I and Type II errors.

Confidence Intervals for p

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: $(0.40 - 0.03, 0.40 + 0.03)$.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that

own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

Constructing Confidence Intervals for p

The structure of, and procedure to find the **confidence interval** for a proportion is similar to that for the population mean, but the formulas are different.

The general format of a confidence interval is:

$$(PE - MoE, PE + MoE)$$

The population parameter is p . The point estimate for p , is \hat{p} , the sample proportion.

The **margin of error** bound for a proportion is:

$$MoE = \left(z_{\frac{\alpha}{2}}\right) \left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \text{ where } \hat{q} = 1 - \hat{p}$$

This formula is similar to the error bound formula for a mean, except that the “appropriate standard error” is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ as the standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the sample proportions \hat{p} and \hat{q} are estimates of the unknown population proportions p and q . The estimated proportions \hat{p} and \hat{q} are used because p and q are not known. The sample proportions \hat{p} and \hat{q} are calculated from the data: \hat{p} is the estimated proportion of successes, and \hat{q} is the estimated proportion of failures.

Example

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.





An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=240#h5p-159>

Your turn!

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

7.5 Behavior of Confidence Intervals for a Proportion

Confidence intervals for p behave similarly to intervals for μ , however there are a few subtleties.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

Recall the **margin of error** for a population proportion is:

- $MoE = \left(z_{\frac{\alpha}{2}}\right) \left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$
- Solving for n gives you an equation for the sample size.
- $n = \left(\frac{z_{\frac{\alpha}{2}}}{MoE}\right)^2 p(1-p)$

Recall the objective of a CI. If we are looking to estimate p , then we do not know what is, however it appears in this formula. So what do we plug in for p ? We have a few options:

- If you have prior information such as a previous sample and can calculate a point estimate (\hat{p}) plug it in!
- You can use your best guess at p
- You can use a “conservative” estimate of p , 0.5.*

***Note:** Remember that $\hat{q} = (1 - \hat{p})$. But, we do not know \hat{p} yet. Since we multiply \hat{p} and \hat{q} together, we make them both equal to 0.5 because $\hat{p}\hat{q} = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n . This gives us a large enough sample so that we can be CL% confident that we are within three percentage points of the true population proportion.

Example

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=242#h5p-160>

Your turn!

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

“Plus Four” Confidence Interval for p .

This is an alternative **optional** method for constructing a CI for p , stemming from the continuity correction of the binomial approximation .

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example

A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Six students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=242#h5p-161>

Your turn!

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus-four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Chapter 7 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://ecampusontario.pressbooks.pub/significantstats/?p=252#h5p-162>

Section Reviews

7.1 Sampling Distribution of the Sample Mean

In many cases, the researcher does not know the population standard deviation, σ , of the measure being studied. In these cases, it is common to use the sample standard deviation, s , as an estimate of σ . The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t -distribution is much better. Define a t -score using the following formula:

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

When using inference techniques for a single population mean the following distributions should be used under certain circumstances:

1. A Student's t -test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal (Z) test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

We can construct confidence Interval with the corresponding critical value or perform Hypothesis tests with the correct Tests statistic

7.2 Inference for the Mean in Practice

The t-score follows the Student's t-distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right) \frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α .

s = the standard deviation of sample values.

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population

mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T , has a Student's t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t-score in the Student's t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

= (point estimate - EBM, point estimate + EBM)

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}} \right)$$

7.3 Sampling Distribution of the Sample Proportion

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Some statistical measures, like many survey questions, measure categorical rather than quantitative data. In this case, the population parameter being estimated is a proportion.

The variable $\hat{p} = x / n$, where x represents the number of successes and n represents the sample size, is the sample proportion and serves as the point estimate for the true population proportion.

The variable \hat{p} has a binomial distribution that can be approximated with the normal distribution shown below given you meet the criteria.

$$\hat{p} \sim N \left(p, \sqrt{\frac{p(1-p)}{n}} \right)$$

7.4 Inference for a Proportion

It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let \hat{p} represent the sample proportion, x/n , where x represents the number of successes and n represents the sample size. Let $q = 1 - \hat{p}$.

The general form of a CI is:

(lower bound, upper bound)

Then the confidence interval for a population proportion is given by the following formula:

$$(\hat{p} - MoE, \hat{p} + MoE)$$

The Margin of Error (MoE) for a proportion is:

$$z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Putting that together:

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

7.5 Behavior of Confidence Intervals for a Proportion

A CI for p behaves similarly to what we have seen for Z CIs for μ as far as Confidence levels and sample sizes.

provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error MoE.

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{MoE} \right)^2 p(1-p)$$

Options to plug in for p :

- If you have prior information such as a previous sample and can calculate a point estimate (\hat{p}) plug it in!
- You can use your best guess at p
- You can use a “conservative” estimate of p , 0.5.*

The “plus four” method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Use:

$$\hat{p} = \frac{x+2}{n+4}$$

Then find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

7.1 Sampling Distribution for the Sample Mean

- Sampling distribution
- Student's t-distribution
- Degrees of freedom
- P-value

7.2 Inference for the Mean in Practice

- Confidence interval
- Point estimate
- Margin of error (MoE)
- T-distribution

7.3 Sampling Distribution of the Sample Proportion

- Population proportion (p)
- Sample proportion (\hat{p})
- Central limit theorem (CLT)

7.4 Inference for a Proportion

- Categorical data
- Population proportion

- **Point estimate**
- **Binomial distribution**
- **Test statistic**
- **Confidence interval**
- **Margin of error (MoE)**

7.5 Behavior of Confidence Intervals for a Proportion

- **Confidence interval**
- **Margin of error (MoE)**

Extra Practice

7.1 Sampling Distribution for the Sample Mean

Which two distributions can you use for inference on a mean?

- Solution:
 1. The normal distribution
 2. Student's t -distribution

2. Which distribution do you use when you are testing a population mean and the population standard deviation is known and/or $n \geq 30$?

- Solution: The normal distribution

3. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

- Solution: Use a Student's t -distribution
-

4. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

5. A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

- Solution: a normal distribution for a single population mean
-

7. You are performing a hypothesis test of a single population mean using a Student's t -distribution. What must you assume about the distribution of the data?

- Solution: It must be approximately normally distributed.
-

7.2 Inference for the Mean in Practice

1. The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the “In utero/newborn” group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. The figure below shows how many of the targeted chemicals were found in each infant's cord blood.¹

1. “Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at <http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn> (accessed July 2, 2013)

Figure 7.8

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

Solution:

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so $\alpha = 1 - CL = 1 - 0.90 = 0.10$
 $\frac{\alpha}{2} = 0.05$, $t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 1.729 \left(\frac{25.965}{\sqrt{20}} \right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

2. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

a. Identify the following:

- \bar{x} = _____
- s_x = _____
- n = _____
- $n - 1$ = _____

b. Define the random variables X and \bar{X} in words.

- Solution: X is the number of hours a patient waits in the emergency room before being called back to be examined. \bar{X} is the mean wait time of 70 patients in the emergency room.

c. Which distribution should you use for this problem?

d. Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

- Solution: CI: (1.3808, 1.6192)

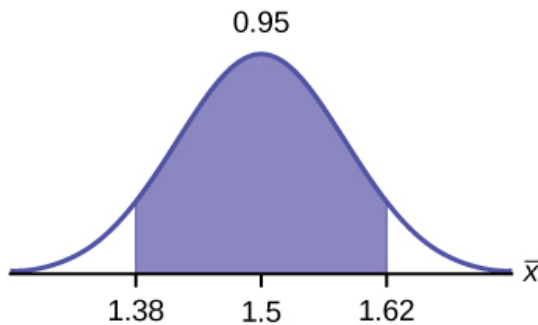


Figure 7.9

- Solution: EBM = 0.12

e. Explain in complete sentences what the confidence interval means.

3. One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

a. Identify the following:

- \bar{x} = _____
- s_x = _____
- n = _____
- $n - 1$ = _____

Solutions:

- \bar{x} = 151
- s_x = 32
- n = 108
- $n - 1$ = 107

b. Define the random variable X in words.

c. Define the random variable \bar{x} in words.

- Solution: \bar{x} is the mean number of hours spent watching television per month from a sample of 108 Americans.

- d. Which distribution should you use for this problem?
- e. Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

- Solution: CI: (142.92, 159.08)

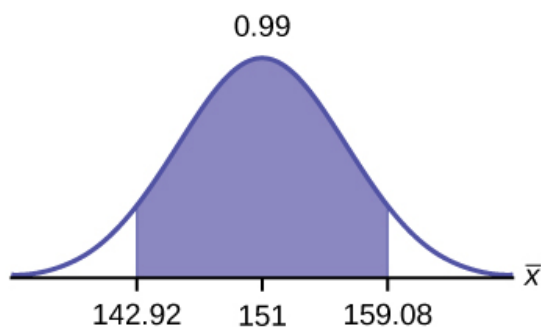


Figure 7.10

- Solution: EBM = 8.08

- f. Why would the error bound change if the confidence level were lowered to 95%?

4. The data in the table below are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

Figure 7.11

X	Freq.
1	1
2	7
3	18
4	7
5	6

- a. Calculate the following:

- \bar{x} = _____
- s_x = _____
- n = _____

Solutions:

- 3.26
- 1.02
- 39

b. Define the random variable \bar{x} in words.

c. What is \bar{x} estimating?

- Solution: μ

d. Is σ_x known?

e. As a result of your answer to the questions above, state the exact distribution to use when calculating the confidence interval.

- Solution: t_{38}

f. Construct a 95% confidence interval for the true mean number of colors on national flags. How much area is in both tails (combined)?

g. How much area is in each tail?

- Solution: 0.025

h. Calculate the following:

- lower limit
- upper limit
- error bound

i. The 95% confidence interval is_____.

- Solution: (2.93, 3.59)

j. Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

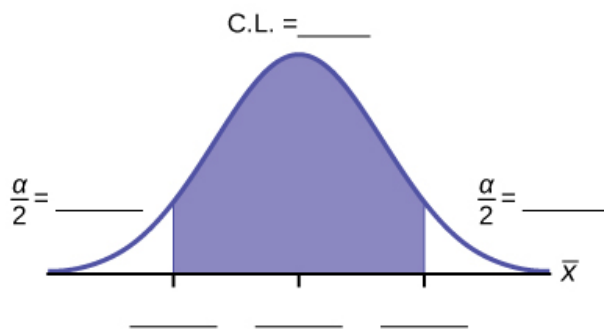


Figure 7.12

k. In one complete sentence, explain what the interval means.

- Solution: We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

l. Using the same \bar{x} , s_x , and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

- Solution: The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

m. Using the same \bar{x} , s_x , and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

6. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

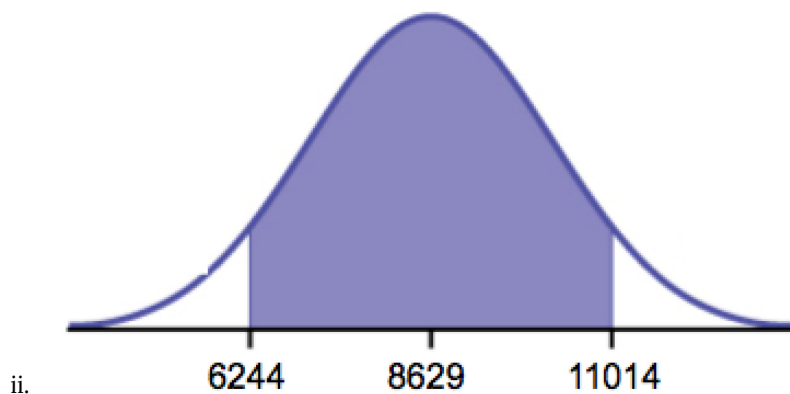
- $\bar{x} = \text{_____}$
 - $s_x = \text{_____}$
 - $n = \text{_____}$
 - $n - 1 = \text{_____}$
- Define the random variables X and \bar{x} in words.
 - Which distribution should you use for this problem? Explain your choice.
 - Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
 - State the confidence interval.
 - Sketch the graph.

- iii. Calculate the error bound.
- e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solutions:

- i. 8629
- ii. 6944
- iii. 35
- iv. 34
- b. t_{34}
 - i. CI: (6244, 11,014)

Figure 7.13



- iii. EB = 2385
- c. It will become smaller

7. Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

- i. \bar{x} = _____
- ii. s_x = _____
- iii. n = _____
- iv. $n - 1$ = _____
- b. Define the random variables X and \bar{x} in words.
- c. Which distribution should you use for this problem? Explain your choice.
- d. Construct a 95% confidence interval for the population mean time wasted.
 - i. State the confidence interval.

- ii. Sketch the graph.
- iii. Calculate the error bound.
- e. Explain in a complete sentence what the confidence interval means.

8. A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7, 2.8, 3.0, 2.3, 2.3, 2.2, 2.8, 2.1, and 2.4.

- i. \bar{x} = _____
- ii. s_x = _____
- iii. n = _____
- iv. $n - 1$ = _____
- b. Define the random variable X in words.
- c. Define the random variable \bar{x} in words.
- d. Which distribution should you use for this problem? Explain your choice.
- e. Construct a 95% confidence interval for the population mean length of time.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- f. What does it mean to be “95% confident” in this problem?

Solutions:

- i. $\bar{x} = 2.51$
- ii. $s_x = 0.318$
- iii. $n = 9$
- iv. $n - 1 = 8$
- b. the effective length of time for a tranquilizer
- c. the mean effective length of time of tranquilizers from a sample of nine patients
- d. We need to use a Student's-t distribution, because we do not know the population standard deviation.
 - i. CI: (2.27, 2.76)
 - ii. Check student's solution.
 - iii. EBM: 0.25
- e. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

9. Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long

they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

- i. \bar{x} = _____
 - ii. s_x = _____
 - iii. n = _____
 - iv. $n - 1$ = _____
 - b. Define the random variable X in words.
 - c. Define the random variable \bar{x} in words.
 - d. Which distribution should you use for this problem? Explain your choice.
 - e. Construct a 99% confidence interval for the population mean length of time using training wheels.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
 - f. Why would the error bound change if the confidence level were lowered to 90%?
-

10. *Forbes* magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The figure below shows the ages of the corporate CEOs for a random sample of these firms.²

Figure 7.14

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's t -distribution.

11. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are

2. "America's Best Small Companies." *Forbes*, 2013. Available online at <http://www.forbes.com/best-small-companies/list/> (accessed July 2, 2013).

randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

- i. \bar{x} = _____
 - ii. s_x = _____
 - iii. n = _____
 - iv. $n-1$ = _____
- b. Define the random variables X and \bar{x} in words.
 - c. Which distribution should you use for this problem? Explain your choice.
 - d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.

Solutions:

- i. $\bar{x} = 11.6$
 - ii. $s_x = 4.1$
 - iii. $n = 225$
 - iv. $n - 1 = 224$
- b. X is the number of unoccupied seats on a single flight. \bar{x} is the mean number of unoccupied seats from a sample of 225 flights.
 - c. We will use a Student's-t distribution, because we do not know the population standard deviation.
 - i. CI: (11.12 , 12.08)
 - ii. Check student's solution.
 - iii. EBM: 0.48

12. In a recent sample of 84 used car sales costs, the sample mean was \$6,425 with a standard deviation of \$3,156. Assume the underlying distribution is approximately normal.

- a. Which distribution should you use for this problem? Explain your choice.
- b. Define the random variable \bar{x} in words.
- c. Construct a 95% confidence interval for the population mean cost of a used car.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
- d. Explain what a "95% confidence interval" means for this study.

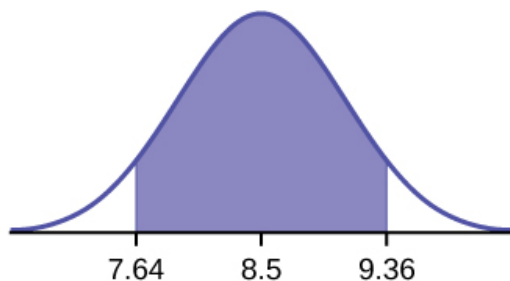
13. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8, 8, 10, 7, 9, 9. Assume the underlying distribution is approximately normal.

- Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
 - State the confidence interval.
 - Sketch the graph.
 - Calculate the error bound.
- If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
- Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
- Calculate the mean.
- Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

Solutions:

- CI: (7.64 , 9.36)

Figure 7.15



ii.

- EBM: 0.86

- The sample should have been increased.
- Answers will vary.
- Answers will vary.
- Answers will vary.

14. A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢, 75¢, 50¢, 65¢, 30¢, 55¢, 40¢, 40¢, 30¢, 55¢, \$1.50, 40¢, 65¢, 40¢.³ Assume the underlying distribution is approximately normal.

- i. $\bar{x} =$ _____
 - ii. $s_x =$ _____
 - iii. $n =$ _____
 - iv. $n-1 =$ _____
- b. Define the random variables X and \bar{x} in words.
 - c. Which distribution should you use for this problem? Explain your choice.
 - d. Construct a 95% confidence interval for the population mean worth of coupons.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
 - e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.
-

15. A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

a. Find the 95% Confidence Interval for the true population mean for the amount of soda served.

- a. (12.42, 14.18)
- b. (12.32, 14.29)
- c. (12.50, 14.10)
- d. Impossible to determine

- Solution: b

b. What is the error bound?

- a. 0.87
 - b. 1.98
 - c. 0.99
 - d. 1.74
-

12. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ _____

- a. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$
 - b. $N(7.24, 1.93)$
 - c. t_{22}
 - d. t_{21}
-

7.3 Sampling Distribution of the Sample Proportion

9. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq ?

- Solution: They must both be greater than five.
-

10. You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

11. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

- Solution: binomial distribution
-

7.4 Inference for a Proportion

5. In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

- a. Define the random variables X and \hat{p} in words.
- b. Which distribution should you use for this problem? Explain your choice
- c. Calculate \hat{p} .
- d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
 - i. State the confidence interval.

- ii. Sketch the graph.
- iii. Calculate the error bound.
- e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

_____?

1. For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution:

$$x = 300 \text{ and } n = 500$$

$$\hat{p} = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = p' = 1 - 0.600 = 0.400$$

$$\text{Since } CL = 0.90, \text{ then } \alpha = 1 - CL = 1 - 0.90 = 0.10 \left(\frac{\alpha}{2} \right) = 0.05$$

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

$$ME = \left(z_{\frac{\alpha}{2}} \right) \sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.645) \sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - ME = 0.60 - 0.036 = 0.564$$

$$p' + ME = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - MoE, p' + MoE) = (0.564, 0.636)$.

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level: Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

2. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

3. Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

b. If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

- Solution: It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.
-

4. Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

a. Identify the following:

a. $x =$ _____

b. $n =$ _____

c. $\hat{p} =$ _____

b. Define the random variables X and \hat{p} in words.

- Solution: X is the number of “successes” where the woman makes the majority of the purchasing decisions for the household.
- \hat{p} is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

c. Which distribution should you use for this problem?

- You can use the Normal distribution here

d. Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

- Solution: CI: (0.5321, 0.6679)

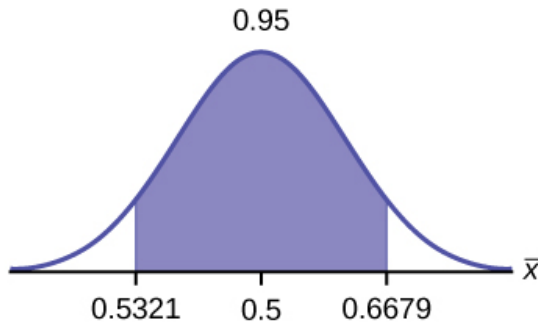


Figure 7.16

- Solution: EBM: 0.0679

e. List two difficulties the company might have in obtaining random results, if this survey were done by email.

5. Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

a. We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and \hat{p} in words.

- Solution: X is the number of “successes” where an executive prefers a truck. \hat{p} is the percentage of executives sampled who prefer a truck.

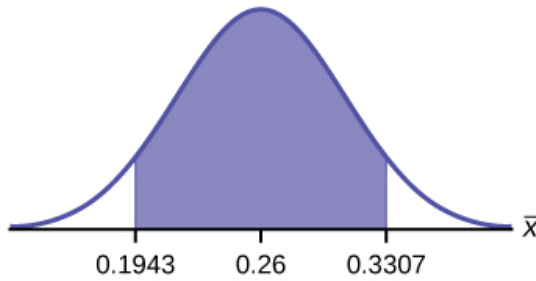
b. Which distribution should you use for this problem?

- You can use the Normal distribution here

c. Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

- Solution: CI: (0.19432, 0.33068)

Figure 7.17



- Solution: EBM: 0.0707

d. Suppose we want to lower the sampling error. What is one way to accomplish that?

e. The sampling error given in the survey is $\pm 2\%$. Explain what the $\pm 2\%$ means.

- Solution: The sampling error means that the true mean can be 2% above or below the sample mean.

6. A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

a. Define the random variable X in words.

b. Define the random variable \hat{p} in words.

- Solution: \hat{p} is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

c. Which distribution should you use for this problem?

d. Construct a 90% confidence interval, and state the confidence interval and the error bound.

- CI: (0.62735, 0.67265)
- EBM: 0.02265

e. What would happen to the confidence interval if the level of confidence were 95%?

7. The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64

girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

a. What is being counted?

- The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

b. In words, define the random variable X .

c. Calculate the following:

- $x = \underline{\hspace{2cm}}$
- $n = \underline{\hspace{2cm}}$
- $p \approx \underline{\hspace{2cm}}$
 - $x = 64$
 - $n = 80$
 - $p \approx 0.8$

d. State the estimated distribution of X . $X \sim \underline{\hspace{2cm}}$

e. Define a new random variable \hat{p} . What is p estimating?

- p

f. In words, define the random variable \hat{p} .

g. State the estimated distribution of \hat{p} . Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

- $\hat{p} \sim N \left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}} \right) = (0.72171, 0.87829).$

h. How much area is in both tails (combined)?

i. How much area is in each tail?

- 0.04

j. Calculate the following:

- lower limit
- upper limit
- error bound

k. The 92% confidence interval is _____.

- (0.72; 0.88)

l. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

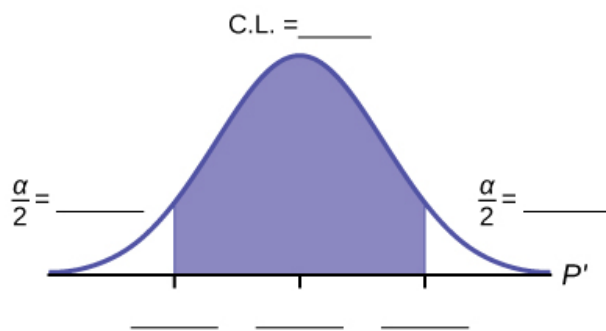


Figure 7.18

m. In one complete sentence, explain what the interval means.

- With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

n. Using the same \hat{p} and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

o. Using the same \hat{p} and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

- The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

p. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

8. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

- When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?

- b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Solutions:

- a. 1,068
b. The sample size would need to be increased since the critical value increases as the confidence level increases.

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

- i. $x =$ _____
ii. $n =$ _____
iii. $\hat{p} =$ _____
b. Define the random variables X and \hat{p} , in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.
-

9. According to a survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

- a. Define the random variables X and \hat{p} in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

Solutions:

- a. X = the number of people who feel that the president is doing an acceptable job;
 P' = the proportion of people in a sample who feel that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

- i. CI: (0.59, 0.63)
- ii. Check student's solution
- iii. EBM: 0.02

10. An article regarding interracial dating and marriage appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.⁴

- a. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and \hat{p} , in words.
- b. Which distribution should you use for this problem? Explain your choice.
- c. Construct a 95% confidence interval.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.

11. Refer to the information in Number 10.

- a. Construct three 95% confidence intervals.
 - i. percent of all Asians who would welcome a white person into their families.
 - ii. percent of all Asians who would welcome a Latino into their families.
 - iii. percent of all Asians who would welcome a black person into their families.
- b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
- c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
- d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Solutions:

4. Fears, Darryl., and Deane, Claudia. "Biracial Couples Report Tolerance." Washington Post, July 5, 2001. Available online at <https://www.washingtonpost.com/archive/politics/2001/07/05/biracial-couples-report-tolerance/c1ce88c8-ba7c-44f5-a348-b86776df9112>. (accessed January 26, 2021).

- i. (0.72, 0.82)
 - ii. (0.65, 0.76)
 - iii. (0.60, 0.72)
 - b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
 - c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
 - d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.
-

12. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

- a. Define the random variables X and \hat{p} in words.
 - b. Which distribution should you use for this problem? Explain your choice.
 - c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight-year period.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
 - d. Explain what a “97% confidence interval” means for this study.
-

13. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.”⁵ We are interested in the population proportion of adult Americans who feel that crime is the main problem.

- a. Define the random variables X and \hat{p} in words.
- b. Which distribution should you use for this problem? Explain your choice.
- c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.

- d. Suppose we want to lower the sampling error. What is one way to accomplish that?
- e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is $\pm 3\%$. In one to three complete sentences, explain what the $\pm 3\%$ represents.

Solutions:

- a. X = the number of adult Americans who feel that crime is the main problem; \hat{p} = the proportion of adult Americans who feel that crime is the main problem
- b. Since we are estimating a proportion, given $\hat{p} = 0.2$ and $n = 1000$, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.
 - i. CI: (0.18, 0.22)
 - ii. Check student's solution.
 - iii. EBM: 0.02
- c. One way to lower the sampling error is to increase the sample size.
- d. The stated " $\pm 3\%$ " represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

14. Refer to the information above. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

- a. Define the random variables X and \hat{p} in words.
 - b. Which distribution should you use for this problem? Explain your choice.
 - c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
 - i. State the confidence interval.
 - ii. Sketch the graph.
 - iii. Calculate the error bound.
 - d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is $\pm 3\%$. In one to three complete sentences, explain what the $\pm 3\%$ represents.
-

15. According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California.⁶ We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

6. The Field Poll. Available online at <http://field.com/fieldpollonline/subscribers/> (accessed July 2, 2013).

a. A point estimate for the true population proportion is:

- a. 0.90
- b. 1.27
- c. 0.79
- d. 400

- Solution: c

b. A 90% confidence interval for the population proportion is _____.

- a. (0.761, 0.820)
- b. (0.125, 0.188)
- c. (0.755, 0.826)
- d. (0.130, 0.183)

c. The error bound is approximately _____.

- a. 1.581
- b. 0.791
- c. 0.059
- d. 0.030

- Solution: d

16. Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

a. Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

- a. (0.2975, 0.3796)
- b. (0.6270, 0.6959)
- c. (0.3041, 0.3730)
- d. (0.6204, 0.7025)

b. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

- a. 0.6614
- b. 0.3386

- c. 173
- d. 338

- Solution: a

17. On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a $\pm 3\%$ margin of error.⁷

- a. Determine the estimated proportion from the sample.
- b. Determine the sample size.
- c. Identify CL and α .
- d. Calculate the error bound based on the information provided.
- e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
- f. Create a confidence interval for the results of this study.
- g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

18. A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.⁸

- a. Find the point estimate and the error bound for this confidence interval.
- b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
- c. Use the point estimate from part a and $n = 1,000$ to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
- d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

Solutions:

7. Saad, Lydia. "Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at <http://www.gallup.com/poll/162758/three-fourworkers-plan-work-past-retirement-age.aspx> (accessed July 2, 2013).
8. "52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

- a. $\hat{p} = \left(\frac{0.55+0.49}{2} \right) = 0.52$; MoE = $0.55 - 0.52 = 0.03$
- b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
- c. CL = 0.75, so $\alpha = 1 - 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125$ $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1 - 0.125 = 0.875$.)

$$MoE = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.150 \sqrt{\frac{0.52(0.48)}{1,000}} \approx 0.018$$

$$(p^- - MoE, p^+ + MoE) = (0.52 - 0.018, 0.52 + 0.018) = (0.502, 0.538)$$
- d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

19. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.⁹

- a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
- b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
- c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

20. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

Solution:

CL = 0.95 $\alpha = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$ $z_{\frac{\alpha}{2}} = 1.96$. Use the conservative estimate of $\hat{p} = \hat{q} = 0.5$.

$$n = \frac{z_{\frac{\alpha}{2}}^2 \hat{p} \hat{q}}{MoE^2} = \frac{1.96^2 (0.5)(0.5)}{0.05^2} = 384.16$$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

21. In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their

9. Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at <http://www.publicpolicypolling.com/Day2MusicPoll.pdf> (accessed July 2, 2013).

community as “likely” or “very likely.”¹⁰ Use the “plus four” method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

7.5 Behavior of Confidence Intervals for a Proportion

1. The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users.¹¹ In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the “plus four” method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

highlight - change to \hat{p}

Solution:

Using “plus-four,” we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$$\hat{p} = \frac{15}{54} \approx 0.278$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.278 = 0.722$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$.

$$z_{0.05} = 1.645$$

$$MoE = \left(z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{\hat{p}\hat{q}}{n}} \right) = (1.645) \left(\sqrt{\frac{(0.278)(0.722)}{54}} \right) \approx 0.100$$

$$\hat{p} - MoE = 0.278 - 0.100 = 0.178$$

$$\hat{p} + MoE = 0.278 + 0.100 = 0.378$$

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook

10. Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at <http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll> (accessed July 2, 2013).
11. Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~media/Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

References

Image References

Figure 7.9: Figure 8.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.10: Figure 8.14 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.12: Figure from Lumen Learning Introduction to Statistics (CC BY 4.0). Retrieved from <https://courses.lumenlearning.com/introstats1/chapter/section-exercises-7/>

Figure 7.13: Figure 8.20 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.15: Figure 8.21 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.16: Figure 8.15 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.17: Figure 8.16 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/8-solutions#eip-589-solution>

Figure 7.18: Figure from Lumen Learning Introduction to Statistics (CC BY 4.0). Retrieved from <https://courses.lumenlearning.com/introstats1/chapter/section-exercises-7/>

Text

“America’s Best Small Companies.” Forbes, 2013. Available online at <http://www.forbes.com/best-small-companies/list/> (accessed July 2, 2013).

Data from Microsoft Bookshelf.

Data from <http://www.businessweek.com/>.

Data from <http://www.forbes.com/>.

“Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).

“Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at <http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn> (accessed July 2, 2013).

“Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at <http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml> (accessed July 2, 2013).

Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at <http://www.publicpolicypolling.com/Day2MusicPoll.pdf> (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at <http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx> (accessed July 2, 2013).

Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~media/Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. “Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at <http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx> (accessed July 2, 2013).

The Field Poll. Available online at <http://field.com/fieldpollonline/subscribers/> (accessed July 2, 2013).

Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at <http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll> (accessed July 2, 2013).

“52% Say Big-Time College Athletics Corrupt Education Process.” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

CHAPTER 8: INFERENCE FOR TWO SAMPLES

8.1 Inference for Two Dependent Samples (Matched Pairs)

Learning Objectives

By the end of this chapter, the student should be able to:

- Classify hypothesis tests by type
- Conduct and interpret hypothesis tests for two population means, population standard deviations known
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown
- Conduct and interpret hypothesis tests for matched or paired samples
- Conduct and interpret hypothesis tests for two population proportions



Figure 8.1: Types of Breakfasts. If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) we will use a two sample analysis.

Studies often compare two groups. For example, maybe researchers are interested in the effect aspirin has in preventing heart attacks. One group is given aspirin and the other a **placebo**, and the heart attack rate is studied over several years. Other studies may compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct **inference** on single **means** and single **proportions**. We know that the first step is deciding what type of data we are working with. For **quantitative data** we are focused on means, while for **categorical** we are focused on proportions. In this chapter we will compare two means or two proportions to each other. The general procedure is still the same, just expanded. With two sample analysis it is good to know what the formulas look like and where they come from, however you will probably lean heavily on technology in performing the calculations.

To compare two means we are obviously working with two groups, but first we need to think about the relationship between them. The groups are classified either as **independent** or dependent. Independent samples consist of two samples that have no relationship, that is, sample values selected from one population

are not related in any way to sample values selected from the other population. Dependent samples consist of two groups that have some sort of identifiable relationship.

Two Dependent Samples (Matched Pairs)

Two samples that are dependent typically come from a **matched pairs** experimental design. The parameter tested using matched pairs is the **population mean difference**. When using inference techniques for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of (or two extremely similar) individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for analysis.

To perform statistical inference techniques we first need to know about the **sampling distribution** of our parameter of interest. Remember although we start with two samples, the differences are the data we are interested in and our parameter of interest is μ_d , the mean difference. Our **point estimate** is \bar{x}_d . In a perfect world we could assume that both samples come from a normal distribution, therefore the difference in those normal distributions are also normal. However in order to use Z, we must know the population standard deviation which is near impossible for a difference distribution. Also it is very hard to find large numbers of matched pairs so the sampling distribution we typically use for \bar{x}_d is a t distribution with $n - 1$ degrees of freedom, where n is the number of differences.

Confidence intervals may be calculated on their own for two samples but often, especially in the case of matched pairs, we first want to formally check to see if a difference exists with a hypothesis test. If we do find a statistically significant difference then we may estimate it with a CI after the fact.

Hypothesis Tests for the Mean difference

In a **hypothesis test** for matched or paired samples, subjects are matched in pairs and differences are calculated, and the population mean difference, μ_d , is our parameter of interest. Although it is possible to test for a certain magnitude of effect, we are most often just looking for a general effect. Our hypothesis would then look like:

$$H_0: \mu_d = 0$$

$$H_a: \mu_d (<, >, \neq) 0$$

The steps are the same as we are familiar with, but it is tested using a Student's-t test for a single population mean with $n - 1$ degrees of freedom, with the test statistic:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the figure below. A lower score indicates less pain. The “before” value is matched to an “after” value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Figure 8.2: Reported Pain Data

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-163>

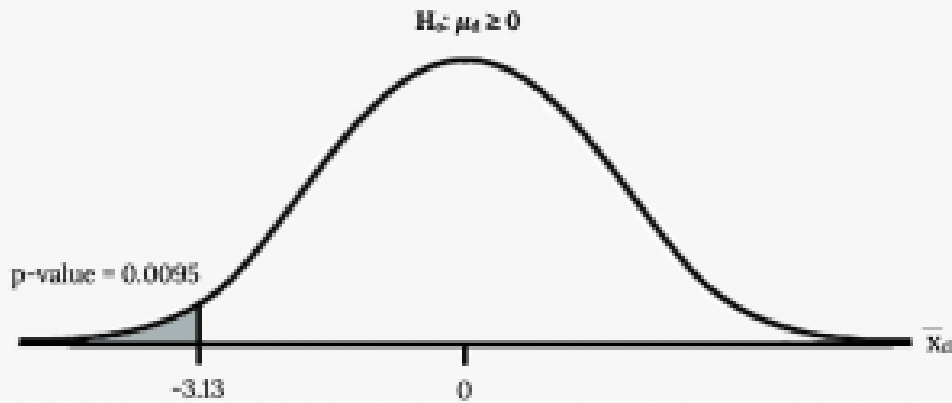


Figure 8.3: Reported Pain P-Value

Your turn!

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Figure 8.4: Cholesterol Levels

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Confidence Intervals for the Mean difference

The general format of a confidence interval is:

$$(PE - MoE, PE + MoE)$$

The population parameter of interest is μ_d , the mean difference. Our point estimate is \bar{x}_d . If we are using the t distribution, the error bound for the population mean difference is:

- $MoE = \left(t_{\frac{\alpha}{2}}\right) \left(\frac{s_d}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, where n is the number of pairs
- s_d = standard deviation of the differences.

Example

A college football coach was interested in whether the college’s strength development class increased his players’ maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Figure 8.5: Weight Lifted

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-164>

Using the differences data, calculate the sample mean and the sample standard deviation.





An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-165>

Using the difference data, this becomes a test of a single _____ (fill in the blank).

Define the random variable: \bar{X}_d mean difference in the maximum lift per player.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-166>

Graph:

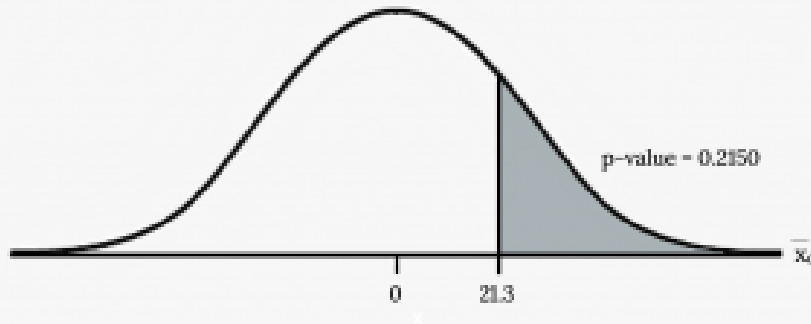


Figure 8.6: Weight Lifted P-Value

Calculate the p-value:



An interactive H5P element has been excluded from this version of the text. You can view it

— online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-167>

Decision:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-168>

What is the conclusion?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=258#h5p-169>

Your turn!

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data recorded in the figure below. Are the scores, on average, higher after the class? Test at a 5% level.

Figure 8.7: SAT Scores

SAT Scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Image Credits

Figure 8.1: Ali Inay (2015). “Brunching with Friends.” Public domain. Retrieved from <https://unsplash.com/photos/y3aP9oo9Pjc>

Figure 8.3: Kindred Grey via Virginia Tech (2020). “Figure 8.3” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_8.3.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Figure 8.6: Kindred Grey via Virginia Tech (2020). “Figure 8.6” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_8.6.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

8.2 Inference for Two Independent Sample Means

Suppose we have two **independent** samples of **quantitative data**. If there is no apparent relationship between the means, our **parameter** of interest is the **difference in means**, $\mu_1 - \mu_2$ with a **point estimate** of $\bar{X}_1 - \bar{X}_2$.

The comparison of two population means is very common. A difference between the two samples depends on both the means and their respective standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, and divide by the **standard error** in order to standardize the difference. We know that when conducting an **inference** for means, the sampling distribution we use (Z or t) depends on our knowledge of the population standard deviation.

Both Population Standard Deviations Known (Z)

Even though this situation is not likely since the population standard deviations are rarely known, we will begin demonstrating these ideas under the ideal circumstances. If we know both mean's **sampling distributions** are normal, the sampling distribution for the difference between the means is normal and both populations must be normal. We can combine the standard errors of each sampling distribution to get a standard error of:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$$

So the sampling distribution of $\bar{X}_1 - \bar{X}_2$ assuming we know both standard deviations is approximately:

$$N\left(\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right)$$

Therefore the **Z test statistic** would be:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Our **confidence interval** would be of the form:

$$(PE - MoE, PE + MoE)$$

Where our point estimate is:

$$\bar{X}_1 - \bar{X}_2$$

And the Margin of error is made up of:

- $MoE = \left(z_{\frac{\alpha}{2}}\right)(SE)$,
- $z_{\frac{\alpha}{2}}$ is the z critical value with area to the right equal to $\frac{\alpha}{2}$

- and SE is $\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$

Since we rarely know one population's standard deviation, much less two, the only situation where we might consider using this in practice is for two very large samples

Both Population Standard Deviations UnKnown (t)

Most likely we will not know the population standard deviations, but we can estimate them using the two sample standard deviations from our independent samples. In this case we will use a t sampling distribution with standard error:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

Assumptions for the Difference in Two Independent Sample Means

Recall we need to be able to assume an underlying normal distribution and no **outliers** or skewness in order to use the t distribution. We can relax these assumptions as our sample sizes get bigger and can typically just use the Z for very large sample sizes.

The remaining question is what do we do for **degrees of freedom** when comparing two groups? One method requires a somewhat complicated calculation but if you have access to a computer or calculator this isn't an issue. We can find a precise *df* for two independent samples as follows:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2} \right)^2}{\left(\frac{1}{n_1-1} \right) \left(\frac{(s_1)^2}{n_1} \right)^2 + \left(\frac{1}{n_2-1} \right) \left(\frac{(s_2)^2}{n_2} \right)^2}$$

NOTES: The *df* are not always a whole number, you usually want to round down. It is not necessary to compute this by hand. Find a reliable technology to do this.

If you are working on your own without access to technology, the above formula could be daunting. Another method is to use a conservative estimate of the *df*:

$\min\{n_1-1, \text{ and } n_2-1\}$

Hypothesis Tests for the Difference in Two Independent Sample Means

Recall the steps to a **hypothesis test** never change. When our parameter of interest is $\mu_1 - \mu_2$ we are often

interested in an effect between the two groups. In order to show an effect, we will have to first assume there is no difference by stating it in the Null Hypothesis as:

$$H_0: \mu_1 - \mu_2 = 0 \text{ OR } H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 - \mu_2 (<, >, \neq) 0 \text{ OR } H_a: \mu_1 (<, >, \neq) \mu_2$$

The **t test statistic** is calculated as follows:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where:

- s_1 and s_2 , the sample standard deviations, are estimates of σ_1 and σ_2 , respectively.
- \bar{x}_1 and \bar{x}_2 are the sample means. μ_1 and μ_2 are the population means. (Note: that in the null we are typically assuming $\mu_1 - \mu_2 = 0$)

Confidence Intervals for the Difference in Two Independent Sample Means

Once we have identified we have a difference in a hypothesis test, we may want to estimate it. Our Confidence Interval would be of the form:

$$(PE - MoE, PE + MoE)$$

Where our point estimate is:

$$\bar{X}_1 - \bar{X}_2$$

And the MoE is made up of:

- $MoE = \left(t_{\frac{\alpha}{2}}\right) (SE),$
- $t_{\frac{\alpha}{2}}$ is the t critical value with area to the right equal to $\frac{\alpha}{2}$
- and SE is $\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

8.3 Inference for Two Sample Proportions

Comparing two proportions, like comparing two means, is also very common when we are working with **categorical data**. If our parameter of inference is $p_1 - p_2$, then we can estimate it with $\hat{p}_1 - \hat{p}_2$

When conducting **inference** on two **independent population proportions**, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Sampling Distribution of the Difference in Two Proportions

We can build a **sampling distribution** for $\hat{p}_1 - \hat{p}_2$ similar to how we did for the difference in two independent sample means. The difference of two proportions follows an approximate normal distribution. We will wait to show the **standard error** and sampling distribution because we calculate it slightly differently for hypothesis tests and confidence intervals

Hypothesis Test for the Difference in Two Proportions

If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A **hypothesis test** can help determine if a difference in the estimated proportions reflects a difference in the population proportions. A confidence Interval can then

Generally, the null hypothesis states that the two proportions are the same, that is, $H_0: p_1 = p_2$. Since we are assuming there is no difference in the null, we can use both samples to estimate the pooled proportion, p_p , calculated as follows:

$$p_p = \frac{x_1 + x_2}{n_1 + n_2}$$

We can use this **pooled proportion** in the calculation of our Z test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_p(1 - p_p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Example

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=263#h5p-170>

Graph:

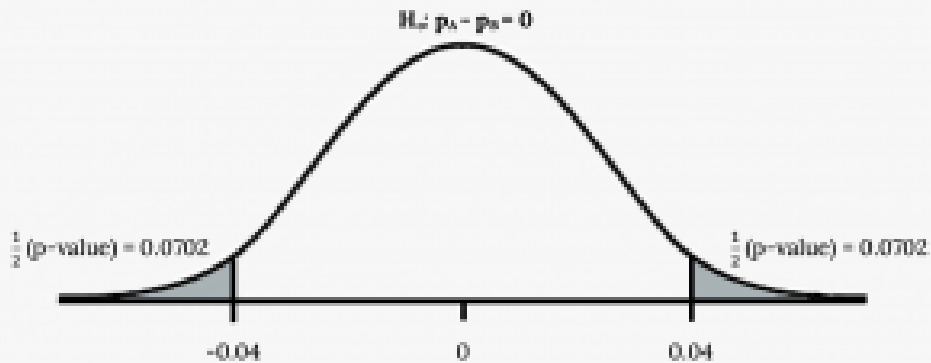


Figure 8.8: Medication A and B

Your turn!

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve A cracked under 4,500 psi. Six out of a random sample of 100 of Valve B cracked under 4,500 psi. Test at a 5% level of significance.

Confidence Intervals for the Difference in Two Proportions

Once we have identified we have a difference in a two sample test, we may want to estimate it. Our **confidence interval** would be of the form:

$$(\{PE - MoE, \{PE + MoE\})$$

Where our **point estimate** is $\hat{p}_1 - \hat{p}_2$

And the MoE is made up of:

- $MoE = \left(z_{\frac{\alpha}{2}}\right) (SE),$
- $z_{\frac{\alpha}{2}}$ is the z critical value with area to the right equal to $\frac{\alpha}{2}$
- And $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- In the SE will we estimate p_1 with \hat{p}_1 and p_2 with \hat{p}_2 if we do not know them.

Putting that all together our formula for a CI to estimate the difference in two proportions will be:

$$\hat{p}_1 - \hat{p}_2 \pm \left(z_{\frac{\alpha}{2}}\right) \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Image References

Figure 8.8: Kindred Grey via Virginia Tech (2020). “Figure 8.8” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_8.8.png . Adaptation of Figure 5.39 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/5-practice>

Chapter 8 Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=281#h5p-171>

Section Reviews

8.1 Inference for 2 Dependent Samples (Matched Pairs)

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: \bar{x}_d = mean of the differences
- Distribution: Student's-t distribution with $n - 1$ degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

Test Statistic (t-score): $t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$ where: \bar{x}_d is the mean of the sample differences. μ_d is the mean of the population differences. s_d is the sample standard deviation of the differences. n is the sample size.

8.2 Inference for 2 Independent Sample Means

A hypothesis test of two population means from independent samples where the population standard deviations are known (typically approximated with the sample standard deviations), will have these characteristics:

- Random variable: $\bar{X}_1 - \bar{X}_2$ = the difference of the means
- Distribution: normal distribution

Normal Distribution:

$$\bar{X}_1 - \bar{X}_2 \sim N \left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \right].$$

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

And Null hypothesis Generally $H_0: \mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$

where:

σ_1 and σ_2 are the known population standard deviations. n_1 and n_2 are the sample sizes. \bar{x}_1 and \bar{x}_2 are the sample means. μ_1 and μ_2 are the population means.

However, most likely we do not know σ_1 and σ_2 so we usually use the t distribution with Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

8.3 Inference for 2 Sample Proportions

Test of two population proportions from independent samples

- Random variable: $p_1 - p_2$, the difference between the two estimated proportions
- Distribution: normal distribution

We often use the “Pooled” Proportion:

$$p_p = \frac{x_1 + x_2}{n_1 + n_2}$$

Distribution for the differences:

$$\hat{p}_1 - \hat{p}_2 \sim N \left[0, \sqrt{p_p (1 - p_p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]$$

With Test Statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{p_p (1 - p_p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

And Null hypothesis $H_0: p_A = p_B$ or $H_0: p_A - p_B = 0$.

where:

\hat{p}_1 and \hat{p}_2 are the sample proportions, p_1 and p_2 are the population proportions,

p_p is the pooled proportion, and n_1 and n_2 are the sample sizes.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

8.1 Inference for 2 Dependent Samples (Matched Pairs)

- **Placebo**
- **Inference**
- **Means**
- **Proportions**
- **Quantitative data**
- **Categorical data**
- **Independent**
- **Matched pairs**
- **Population mean difference**
- **Sampling distribution**
- **Point estimate**

8.2 Inference for 2 Independent Sample Means

- **Independent**
- **Quantitative data**
- **Parameter**
- **Difference in means**
- **Point estimate**
- **Standard error**
- **Inference**

- Sampling distribution
- Test statistic
- Confidence interval
- Outliers
- Degrees of freedom

8.3 Inference for 2 Sample Proportions

- Categorical data
- Inference
- Independent
- Population proportion
- Sampling distribution
- Standard error
- Pooled proportion
- Confidence interval
- Point estimate

Extra Practice

8.1 Inference for 2 Dependent Samples (Matched Pairs)

1. Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded below.

Figure 8.9

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Conduct a hypothesis test to determine whether the mean difference in distances between the children’s dominant versus weaker hands is significant.

Record the **differences** data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: {2, 12, 7, -1, 2, 0, 4}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

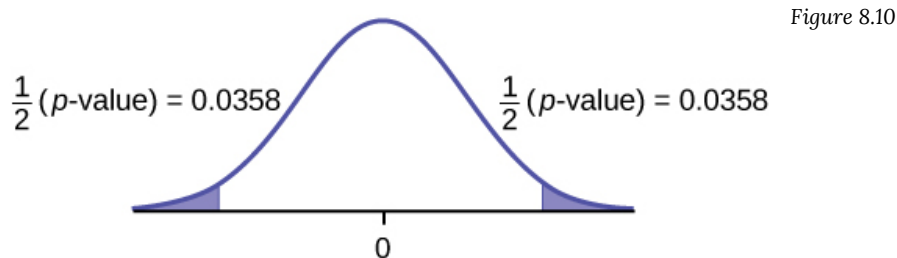
- $\bar{x}_d = 3.71, s_d = 4.5$.

Random variable: \bar{X}_d = mean difference in the distances between the hands.

Distribution for the hypothesis test: t_6

$H_0: \mu_d = 0$ $H_a: \mu_d \neq 0$

Graph:



Calculate the p-value: The p-value is 0.0716 (using the data directly).

- (test statistic = 2.18. p-value = 0.0719 using $(\bar{x}_d = 3.71, s_d = 4.5)$)

Decision: Assume $\alpha = 0.05$. Since $\alpha < p$ -value, Do not reject H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot-put.

2. Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded below. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

Figure 8.11

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant Hand	120	111	135	140	125
Off-hand	105	109	98	111	99

3. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown below. The “before” value is matched to an “after” value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Figure 8.12

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

a. What is the random variable?

- the mean difference of the system failures

b. State the null and alternative hypotheses.

c. What is the p -value?

- 0.0067

d. Draw the graph of the p -value.

e. What conclusion can you draw about the software patch?

- With a p -value 0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

4. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Figure 8.13

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

a. State the null and alternative hypotheses.

b. What is the p -value?

- 0.0021

- c. What is the sample mean difference?
- d. Draw the graph of the p -value.

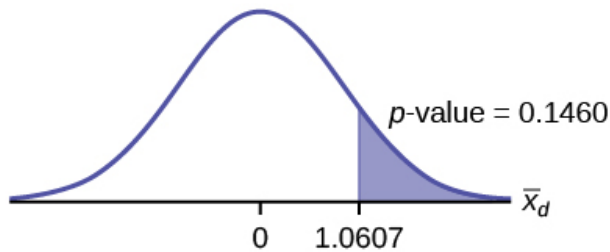


Figure 8.14

- e. What conclusion can you draw about the juggling class?

5. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Figure 8.15

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

- a. State the null and alternative hypotheses.

- $H_0: \mu_d \geq 0$
- $H_a: \mu_d < 0$

- b. What is the test statistic?

- c. What is the p -value?

- 0.0699

- d. What is the sample mean difference?

- e. What is the conclusion?

- We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [Appendix E](#). Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note: If you are using a Student's *t*-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

6. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded below. Do you think that their cholesterol levels were significantly lowered?

Figure 8.16

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

- $p\text{-value} = 0.1494$
- At the 5% significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

7. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t = treated patient and ut = untreated patient.

The appropriate hypotheses are:

- $H_0: p_t < p_{ut}$ and $H_a: p_t \geq p_{ut}$
- $H_0: p_t \leq p_{ut}$ and $H_a: p_t > p_{ut}$
- $H_0: p_t = p_{ut}$ and $H_a: p_t \neq p_{ut}$
- $H_0: p_t = p_{ut}$ and $H_a: p_t < p_{ut}$

If the p -value is 0.0062 what is the conclusion (use $\alpha = 0.05$)?

- a. The method has no effect.
 - b. There is sufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.
 - c. There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients who develop AIDS after four years.
 - d. There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.
- Solution: b
-

8. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a “biofeedback exercise program.” Six subjects were randomly selected and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after – before) producing the following results: $\bar{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

The distribution for the test is:

- a. t_5
- b. t_6
- c. $N(-10.2, 8.4)$
- d. $N(-10.2, \frac{8.4}{\sqrt{6}})$

If $\alpha = 0.05$, the p -value and the conclusion are

- a. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
 - b. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
 - c. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
 - d. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.
- Solution: c
-

9. A golf instructor is interested in determining if her new technique for improving players’ golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

Figure 8.17

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

The correct decision is:

- a. Reject H_0 .
 - b. Do not reject the H_0 .
-

10. A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are shown below.

Figure 8.18

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Test: two matched pairs or paired samples (t-test)

Random variable: \bar{X}_d

Distribution: t_{12}

$H_0: \mu_d = 0$ $H_a: \mu_d > 0$

The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.

Graph: right-tailed

p-value: 0.0004

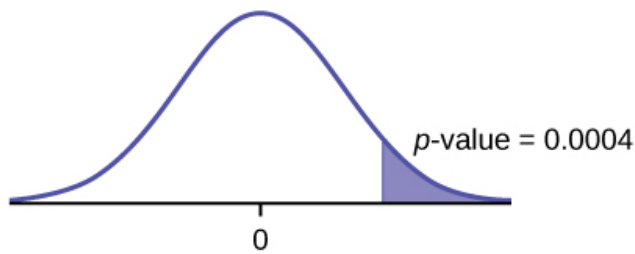


Figure 8.19

Decision: Reject H_0

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

11. A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is shown below. Test at the 1% level of significance.

Figure 8.20

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

12. A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are shown below.

Figure 8.21

Northeastern States	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Test: matched or paired samples (t-test)

Difference data: $\{-0.9, -3.7, -3.2, -0.5, 0.6, -1.9, -0.5, 0.2, 0.6, 0.4, 1.7, -2.4, 1.8\}$

Random Variable: \bar{X}_d

Distribution: $H_0: \mu_d = 0$ $H_a: \mu_d < 0$

The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.

Graph: left-tailed.

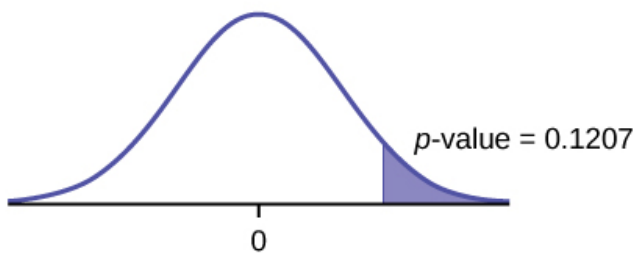


Figure 8.22

p-value: 0.1207

Decision: Do not reject H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

13-22. Indicate which of the following choices best identifies the hypothesis test.

- a. independent group means, population standard deviations and/or variances known
- b. independent group means, population standard deviations and/or variances unknown
- c. matched or paired samples
- d. single mean
- e. two proportions
- f. single proportion

13. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

14. Use the list from 13 to choose the best hypothesis test. A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

- e
-

15. Use the list from 13 to choose the best hypothesis test. The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from nine males and 16 females.

16. Use the list from 13 to choose the best hypothesis test. A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

- d
-

17. Use the list from 13 to choose the best hypothesis test. A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and one year, respectively.

18. Use the list from 13 to choose the best hypothesis test. According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

- f
-

19. Use the list from 13 to choose the best hypothesis test. According to a recent study, U.S. companies have a mean maternity-leave of six weeks.

20. Use the list from 13 to choose the best hypothesis test. A recent drug survey showed an increase in use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.

- e
-

21. Use the list from 13 to choose the best hypothesis test. A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Figure 8.23

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

22. Use the list from 13 to choose the best hypothesis test. University of Michigan researchers reported in the JOURNAL OF THE NATIONAL CANCER INSTITUTE that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.¹

- f

23. Lesley E. Tan investigated the relationship between left-handedness vs. right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown below. Determine the appropriate test and best distribution to use for that test.

Figure 8.24

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

- Two independent means, normal distribution
- Two independent means, Student's-t distribution
- Matched or paired samples, Student's-t distribution
- Two population proportions, normal distribution

24. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are:

Figure 8.25

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

This is:

1. Data from the American Cancer Society. Available online at <http://www.cancer.org/index> (accessed June 17, 2013).

- a test of two independent means.
- a test of two proportions.
- a test of a single mean.
- a test of a single proportion.

- Solution: a

8.2 Inference for 2 Independent Sample Means

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. **Twenty floors** are randomly assigned **to test each wax**. Both populations have a normal distributions. The data are recorded in [\(Figure\)](#).

Wax	Sample Mean Number of Months Floor Wax Lasts	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X}_1 - \bar{X}_2$ = difference in the mean number of months the competing floor waxes last.

$$H_0: \mu_1 \leq \mu_2$$

$$H_a: \mu_1 > \mu_2$$

The words “**is more effective**” says that **wax 1 lasts longer than wax 2**, on average. “Longer” is a “>” symbol and goes into H_a . Therefore, this is a right-tailed test.

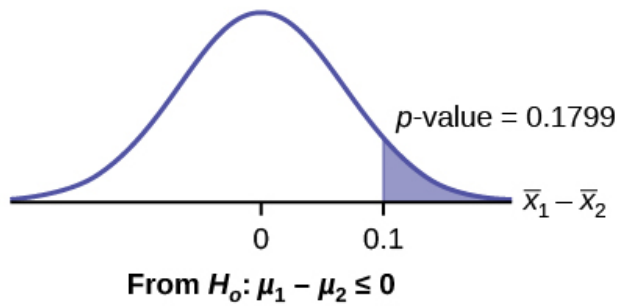
Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula, the distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N \left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}} \right)$$

Since $\mu_1 \leq \mu_2$ then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculate the p-value using the normal distribution: $p\text{-value} = 0.1799$

Graph:



$$\bar{X}_1 - \bar{X}_2 = 3 - 2.9 = 0.1$$

Compare α and the p-value: $\alpha = 0.05$ and $p\text{-value} = 0.1799$. Therefore, $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

Try It

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. (Figure) shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

Engine	Sample Mean Number of RPM	Population Standard Deviation
1	1,500	50
2	1,600	60

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. (Figure) shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

What is the random variable?

The difference in mean speeds of the fastball pitches of the two pitchers

State the null and alternative hypotheses.

What is the test statistic?

-2.46

What is the p -value?

At the 1% significance level, what is your conclusion?

At the 1% significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez's fastball is faster than Wesley's.

Use the following information to answer the next five exercises. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

Is the population standard deviation known or unknown?

State the null and alternative hypotheses.

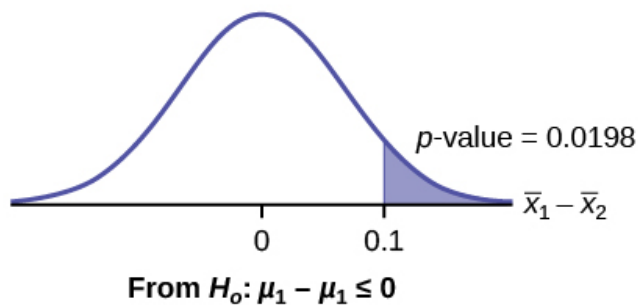
Subscripts: 1 = Food, 2 = No Food

$H_0: \mu_1 \leq \mu_2$

$H_a: \mu_1 > \mu_2$

What is the p -value?

Draw the graph of the p -value.



At the 1% significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. 15 pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample Mean Melting Temperatures (°F)	Population Standard Deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

State the null and alternative hypotheses.

Subscripts: 1 = Gamma, 2 = Zeta

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

Is this a right-, left-, or two-tailed test?

What is the p -value?

0.0062

Draw the graph of the p -value.

At the 1% significance level, what is your conclusion?

There is sufficient evidence to reject the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was \$679. For 23 teenage girls, it was \$559. From past years, it is known that the population standard deviation for each group is \$180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

Subscripts: 1 = boys, 2 = girls

- $H_0: \mu_1 \leq \mu_2$
- $H_a: \mu_1 > \mu_2$
- The random variable is the difference in the mean auto insurance costs for boys and girls.
- normal
- test statistic: $z = 2.50$
- p -value: 0.0062
- Check student's solution.
 - Alpha: 0.05
 - Decision: Reject the null hypothesis.
 - Reason for Decision: p -value < alpha
 - Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were \$947 and \$1,011, respectively. The population standard deviations are known to be \$254 and \$87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that the population standard deviations are known to be six and three, respectively. Conduct a hypothesis test to evaluate the manufacturers claim.

Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

- a. $H_0: \mu_1 \geq \mu_2$
- b. $H_a: \mu_1 < \mu_2$
- c. The random variable is the difference in the mean miles per gallon of non-hybrid sedans and hybrid sedans.
- d. normal
- e. test statistic: 6.36
- f. p-value: 0
- g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for decision: $p\text{-value} < \alpha$
 - iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). (Figure) contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

- $H_0: \mu_d = 0$
- $H_a: \mu_d < 0$
- The random variable X_d is the average difference between husband's and wife's satisfaction level.
- t_9
- test statistic: $t = -1.86$
- p -value: 0.0479
- Check student's solution
 - Alpha: 0.05
 - Decision: Reject the null hypothesis, but run another test.
 - Reason for Decision: p -value < alpha
 - Conclusion: This is a weak test because alpha and the p -value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

Independent groups

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in [\(Figure\)](#). Each population has a normal distribution.

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\bar{X}_g - \bar{X}_b$ = difference in the sample mean amount of time girls and boys play sports each day.

$$H_0: \mu_g = \mu_b \quad H_0: \mu_g - \mu_b = 0$$

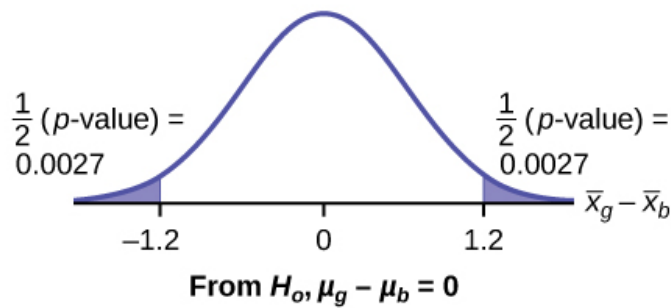
$$H_a: \mu_g \neq \mu_b \quad H_a: \mu_g - \mu_b \neq 0$$

The words **“the same”** tell you H_0 has an “=”. Since there are no other words to indicate H_a , assume it says **“is different.”** This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the p -value using a Student's t -distribution: p -value = 0.0054

Graph:



$$s_g = 0.866$$

$$s_b = 1$$

$$\text{So, } \bar{x}_g - \bar{x}_b = 2 - 3.2 = -1.2$$

Half the p -value is below -1.2 and half is above 1.2 .

Make a decision: Since $\alpha > p$ -value, reject H_0 . This means you reject $\mu_g = \mu_b$. The means are different.

Press STAT. Arrow over to TESTS and press 4:2-SampTTest. Arrow over to Stats and press ENTER. Arrow down and enter 2 for the first sample mean, 0.866 for $Sx1$, 9 for $n1$, 3.2 for the second sample mean, 1 for $Sx2$, and 16 for $n2$. Arrow down to $\mu1$: and arrow to does not equal $\mu2$. Press ENTER. Arrow down to Pooled: and No. Press ENTER. Arrow down to Calculate and press ENTER. The p -value is $p = 0.0054$, the dfs are approximately 18.8462, and the test statistic is -3.14 . Do the procedure again but instead of Calculate do Draw.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different (mean number of hours boys aged seven to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged seven to 11 play sports per day is greater than the mean number of hours played by boys).

Try It

Two samples are shown in (Figure). Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

NOTE

When the sum of the sample sizes is larger than 30 ($n_1 + n_2 > 30$) you can use the normal distribution to approximate the Student's t .

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes

with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes**, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?

a. two means

b. Are the populations standard deviations known or unknown?

b. unknown

c. Which distribution do you use to perform the test?

c. Student's t

d. What is the random variable?

d. $\overline{X}_A - \overline{X}_B$

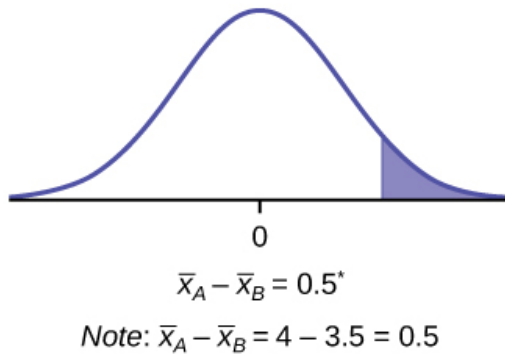
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.

e.

- $H_o : \mu_A \leq \mu_B$
- $H_a : \mu_A > \mu_B$

f. Is this test right-, left-, or two-tailed?

f.



right

g. What is the p -value?

g. 0.1928

h. Do you reject or not reject the null hypothesis?

h. Do not reject.

i. **Conclusion:**

i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

Try It

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

- a. Are the population standard deviations known?
- b. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower

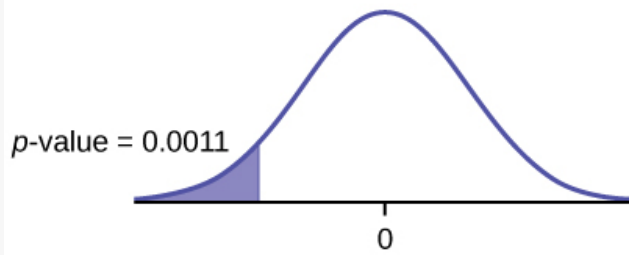
than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in (Figure) and (Figure).

Online Class									
67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Face-to-face Class									
77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

- Is this a test of two means or two proportions?
 - Are the population standard deviations known or unknown?
 - Which distribution do you use to perform the test?
 - What is the random variable?
 - What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
 - Is this test right, left, or two tailed?
 - What is the p -value?
 - Do you reject or not reject the null hypothesis?
 - At the ____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
- two means
 - unknown
 - Student's t
 - $\overline{X}_1 - \overline{X}_2$
 - $H_0: \mu_1 = \mu_2$ Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
 - $H_a: \mu_1 < \mu_2$ Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
 - left-tailed
 - p -value = 0.0011



- g. Reject the null hypothesis
- h. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.

At the 5% level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

Cohen's Standards for Small, Medium, and Large Effect SizesCohen's d is a measure of effect size based on the differences between two means. Cohen's d , named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Cohen's Standard Effect Sizes	
Size of effect	d
Small	0.2
medium	0.5
Large	0.8

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \text{ where } s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Calculate Cohen's d for (Figure). Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

$$\begin{aligned} \mu_1 &= 4 \quad s_1 = 1.5 \quad n_1 = 11 \\ \mu_2 &= 3.5 \quad s_2 = 1 \quad n_2 = 9 \\ d &= 0.384 \end{aligned}$$

The effect is small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small indicating that there is not a significant difference between them.

Calculate Cohen's d for (Figure). Is the size of the effect small, medium or large? Explain what the size of the effect means for this problem.

$d = 0.834$; Large, because 0.834 is greater than Cohen's 0.8 for a large effect size. The size of the differences between the means of the Final Exam scores of online students and students in a face-to-face class is large indicating a significant difference.

Try It

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013 are listed in (Figure) and (Figure), respectively.

Northeast										
94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6	
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5	
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0	

West										
126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6	
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6	
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5	

Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

- Is this a test of two means or two proportions?
- Are the population standard deviations known or unknown?
- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
- Is this test right, left, or two tailed?
- What is the p -value?
- Do you reject or not reject the null hypothesis?
- At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
- Calculate Cohen's d and interpret it.

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

- independent group means, population standard deviations, and/or variances known

- b. independent group means, population standard deviations, and/or variances unknown
- c. matched or paired samples
- d. single mean
- e. two proportions
- f. single proportion

It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

two proportions

A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

matched or paired samples

The known standard deviation in salary for all mid-level professionals in the financial industry is \$11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is \$80,000. The sample mean salary for mid-level professionals in Company B is \$96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

The average worker in Germany gets eight weeks of paid vacation.

single mean

According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

independent group means, population standard deviations and/or variances unknown

The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random

sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

two proportions

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.

It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

independent group means, population standard deviations and/or variances unknown

Varsity athletes practice five times a week, on average.

A sample of 12 in-state graduate school programs at school A has a mean tuition of \$64,000 with a standard deviation of \$8,000. At school B, a sample of 16 in-state graduate programs has a mean of \$80,000 with a standard deviation of \$6,000. On average, are the mean tuitions different?

independent group means, population standard deviations and/or variances unknown

A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

two proportions

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

Are standard deviations known or unknown?

What is the random variable?

The random variable is the difference between the mean amounts of sugar in the two soft drinks.

Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

Is this a test of means or proportions?

means

State the null and alternative hypotheses.

- a. H_0 : _____
- b. H_a : _____

Is this a right-tailed, left-tailed, or two-tailed test?

two-tailed

In symbols, what is the random variable of interest for this test?

In words, define the random variable of interest for this test.

the difference between the mean life spans of whites and nonwhites

Which distribution (normal or Student's t) would you use for this hypothesis test?

Explain why you chose the distribution you did for [\(Figure\)](#).

This is a comparison of two population means with unknown population standard deviations.

Calculate the test statistic and p -value.

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p -value.

Check student's solution.

Find the p -value.

At a pre-conceived $\alpha = 0.05$, what is your:

- a. Decision:
- b. Reason for the decision:
- c. Conclusion (write out in a complete sentence):

- a. Reject the null hypothesis
- b. $p\text{-value} < 0.05$
- c. There is not enough evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

Does it appear that the means are the same? Why or why not?

Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [Appendix E](#). Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's t -distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

A student at a four-year college claims that mean enrollment at four-year colleges is higher than at two-year colleges in the United States. Two surveys are conducted. Of the 35 two-year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

Subscripts: 1: two-year colleges; 2: four-year colleges

- a. $H_0: \mu_1 \geq \mu_2$
- b. $H_a: \mu_1 < \mu_2$
- c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.
- d. Student's- t
- e. test statistic: -0.2480
- f. $p\text{-value}$: 0.4019
- g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Do not reject
 - iii. Reason for Decision: $p\text{-value} > \alpha$
 - iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean

enrollment at four-year colleges is higher than at two-year colleges.

At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were \$46,100 and \$46,700, respectively. Their standard deviations were \$3,450 and \$4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

Subscripts: 1: mechanical engineering; 2: electrical engineering

- $H_0: \mu_1 \geq \mu_2$
- $H_a: \mu_1 < \mu_2$
- $\bar{X}_1 - \bar{X}_2$ is the difference between the mean entry level salaries of mechanical engineers and electrical engineers.
- t_{108}
- test statistic: $t = -0.82$
- p -value: 0.2061
- Check student's solution.
 - Alpha: 0.05
 - Decision: Do not reject the null hypothesis.
 - Reason for Decision: p -value > alpha
 - Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each)

with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from [\(Figure\)](#) to answer the next four exercises.

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

- a. $H_0: \mu_1 = \mu_2$
- b. $H_a: \mu_1 \neq \mu_2$
- c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean times for completing a lap in races and in practices.
- d. $t_{20.32}$
- e. test statistic: -4.70
- f. p -value: 0.0001
- g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for Decision: p -value < alpha
 - iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Repeat the test in [Exercise 10.83](#), but use Lap 5 data this time.

Repeat the test in [Exercise 10.83](#), but this time combine the data from Laps 1 and 5.

- a. $H_0: \mu_1 = \mu_2$
- b. $H_a: \mu_1 \neq \mu_2$
- c. is the difference between the mean times for completing a lap in races and in practices.
- d. $t_{40.94}$
- e. test statistic: -5.08
- f. p -value: zero
- g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for Decision: p -value < alpha
 - iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question. "Does Terri Vogel drive faster in races than she does in practices?"

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer

conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to answer the next two exercises.

The **exact** distribution for the hypothesis test is:

- the normal distribution
- the Student's t -distribution
- the uniform distribution
- the exponential distribution

If the level of significance is 0.05, the conclusion is:

- There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
- There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
- There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
- Unable to determine

c

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. A concluding statement is:

- There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
- There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.

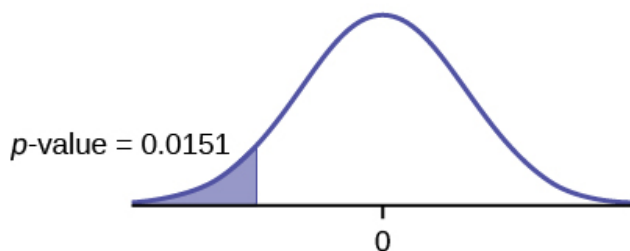
- c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
- d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

Test: two independent sample means, population standard deviations unknown.

Random variable: $\bar{X}_1 - \bar{X}_2$

Distribution: $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 < \mu_2$ The mean age of entering prostitution in Canada is lower than the mean age in the United States.



Graph: left-tailed

p -value : 0.0151

Decision: Do not reject H_0 .

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

- a. $\mu_{\text{day}} > \mu_{\text{night}}$
- b. $\mu_{\text{day}} < \mu_{\text{night}}$
- c. $\mu_{\text{day}} = \mu_{\text{night}}$

d. $\mu_{\text{day}} \neq \mu_{\text{night}}$

8.3 Inference for 2 Sample Proportion

1. A research study was conducted about gender differences in “sexting.” The researcher believed that the proportion of girls involved in “sexting” is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized below. Is the proportion of girls sending sexts less than the proportion of boys “sexting?” Test at a 1% level of significance.²

Figure 8.26

	Males	Females
Sent “sexts”	183	156
Total number surveyed	2231	2169

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then p_M and p_F are the desired population proportions.

Random variable: $p_{\hat{F}} - p_{\hat{M}}$ = difference in the proportions of males and females who sent “sexts.”

$H_0: p_F = p_M$ $H_0: p_F - p_M = 0$

$H_a: p_F < p_M$ $H_a: p_F - p_M < 0$

The words “**less than**” tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_p = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077$$

$$1 - p_p = 0.923$$

Therefore,

$$\hat{p}_F - \hat{p}_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right)$$

$p_{\hat{F}} - p_{\hat{M}}$ follows an approximate normal distribution.

Calculate the p -value using the normal distribution:

p -value = 0.1045

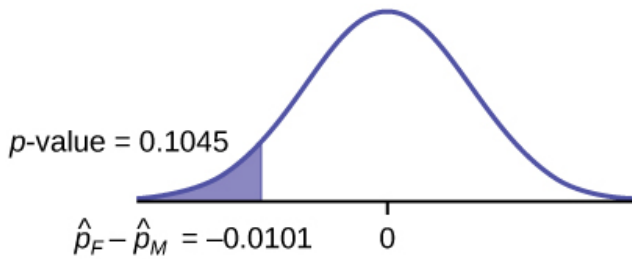
Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

2. Hinduja, Sameer. “Sexting Research and Gender Differences.” Cyberbullying Research Center, 2013. Available online at <http://cyberbullying.us/blog/sexting-research-and-gender-differences/> (accessed June 17, 2013).

Figure 8.27



Decision: Since $\alpha < p\text{-value}$, Do not reject H_0

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending “sexts” is less than the proportion of boys sending “sexts.”

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press ENTER. Arrow down to Calculate and press ENTER. The p-value is $P = 0.1045$ and the test statistic is $z = -1.256$.

2. Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?³

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then p_W and p_A are the desired population proportions.

Random variable: $\hat{p}_W - \hat{p}_A$ = difference in the proportions of Android and iPhone users.

$$H_0: p_W = p_A \quad H_0: p_W - p_A = 0$$

$$H_a: p_W > p_A \quad H_a: p_W - p_A > 0$$

The words “more popular” indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_p = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927$$

$$1 - p_p = 0.9073$$

Therefore,

$$\hat{p}_W - \hat{p}_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right)$$

$\hat{p}_W - \hat{p}_A$ follows an approximate normal distribution.

3. “Smart Phone Users, By the Numbers.” Visually, 2013. Available online at <http://visual.ly/smart-phone-users-numbers> (accessed June 17, 2013).

Calculate the p -value using the normal distribution:

p -value = 0.0077

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

<!-- LALALA 🎵🎵🎵 CONTINUE INSERTING NEW EXAMPLE 3 HERE -->

Graph:

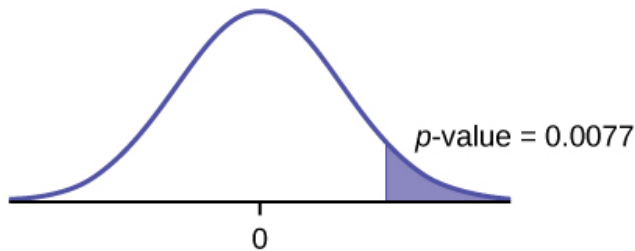


Figure 8.28

Decision: Since $\alpha > p$ -value, reject the H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

3. An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.⁴

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is $30 + 30 = 60$, which is greater than 30, so we can use the normal approximation to the Student's-t distribution. Subscripts: 1: Democratic senators 2: Republican senators

Random variable: $\bar{X}_1 - \bar{X}_2$ = difference in the mean age of Democratic and Republican U.S. senators.

$H_0: \mu_1 \leq \mu_2$ $H_0: \mu_1 - \mu_2 \leq 0$

$H_a: \mu_1 > \mu_2$ $H_a: \mu_1 - \mu_2 > 0$

The words "older than" translates as a ">" symbol and goes into H_a . Therefore, this is a right-tailed test.

4. \ "List of current United States Senators by Age." Wikipedia. Available online at http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age (accessed June 17, 2013).

Distribution for the test: The distribution is the normal approximation to the Student's t for means, independent groups. Using the formula, the distribution is: $\bar{X}_1 - \bar{X}_2 \sim N \left[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}} \right]$

Since $\mu_1 \leq \mu_2$, $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

(Calculating the p -value using the normal distribution gives p -value = 0.4955)

Graph:

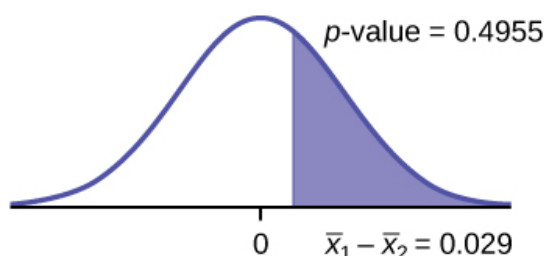


Figure 8.29

Compare α and the p -value: $\alpha = 0.05$ and p -value = 0.4955. Therefore, $\alpha < p$ -value.

Make a decision: Since $\alpha < p$ -value, do not reject H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

4. A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category.⁵ Test at a 5% significance level. Answer the following questions:

- Is this a test of two means or two proportions?
- Which distribution do you use to perform the test?
- What is the random variable?
- What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.
- Is this test right-, left-, or two-tailed?
- What is the p -value?

5. "Texas Crime Rates 1960–1012." FBI, Uniform Crime Reports, 2013. Available online at: <http://www.disastercenter.com/crime/txcrime.htm> (accessed June 17, 2013).

g. Do you reject or not reject the null hypothesis?

h. At the ____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

5. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS₁ had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS₂ had system failures within the first eight hours of operation. OS₂ is believed to be more stable (have fewer crashes) than OS₁.

a. Is this a test of means or proportions?

b. What is the random variable?

- $\hat{p}_{OS1} - \hat{p}_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS₁ and OS₂.

c. State the null and alternative hypotheses.

d. What is the p-value?

- 0.1018

e. What can you conclude about the two operating systems?

6. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races.⁶ Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

a. Is this a test of means or proportions?

- proportions

6. "State of the States." Gallup, 2013. Available online at <http://www.gallup.com/poll/125066/StateStates.aspx?ref=interactive> (accessed June 17, 2013).

b. State the null and alternative hypotheses.

c. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

- right-tailed

d. What is the random variable of interest for this test?

- In words, define the random variable for this test.

e. The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

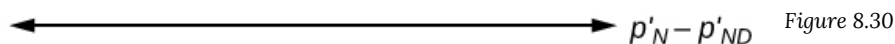
f. Which distribution (normal or Student's t) would you use for this hypothesis test?

g. Explain why you chose the distribution you did.

- Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

h. Calculate the test statistic.

i. Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p -value.



j. Find the p -value.

k. At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:

b. Reason for the decision:

c. Conclusion (write out in a complete sentence):

- Reject the null hypothesis.
- p -value < alpha
- At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

l. Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

7. DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in (Figure). Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note: If you are using a Student's t -distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

a. A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

b. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

8. We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the whites and the blacks races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24.⁷ We will let female suicide victims be our population.

- a. $H_0: P_W = P_B$
 - b. $H_a: P_W \neq P_B$
 - c. The random variable is the difference in the proportions of white and black suicide victims, aged 15 to 24.
 - d. normal for two proportions
 - e. test statistic: -0.1944
 - f. p -value: 0.8458
 - g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for decision: p -value > alpha
 - iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of white and black female suicide victims, aged 15 to 24, are different.
-

9. Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{\text{larger} + \text{smaller dimension}}{\text{larger dimension}}\right)$ was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064.⁸ Do you think that there is a significant difference in the Golden Ratio calculation?

10. A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students.⁹ In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

- a. $H_0: p_1 = p_2$
 - b. $H_a: p_1 \neq p_2$
 - c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
 - d. normal for two proportions
 - e. test statistic: 4.29
 - f. p -value: 0.00002
 - g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for decision: p -value < alpha
 - iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.
-

11. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011.¹⁰ Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010

8. Data from Whitney Exhibit on loan to San Jose Museum of Art

9. Data from the Chancellor's Office, California Community Colleges, November 1994.

10. "West Nile Virus." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/ncidod/dvbid/westnile/index.htm> (accessed June 17, 2013).

proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

- “2011” subscript: 2011 group.
- “2010” subscript: 2010 group

a. This is:

- a test of two proportions
- a test of two independent means
- a test of a single mean
- a test of matched pairs.

b. An appropriate null hypothesis is:

- $p_{2011} \leq p_{2010}$
- $p_{2011} \geq p_{2010}$
- $\mu_{2011} \leq \mu_{2010}$
- $p_{2011} > p_{2010}$

- Solution: a

c. The p -value is 0.0022. At a 1% level of significance, the appropriate conclusion is

- There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
- There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
- There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
- There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.

12. Researchers conducted a study to find out if there is a difference in the use of eReaders by different age

groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders.¹¹

Test: two independent sample proportions.

Random variable: $p_1 - p_2$

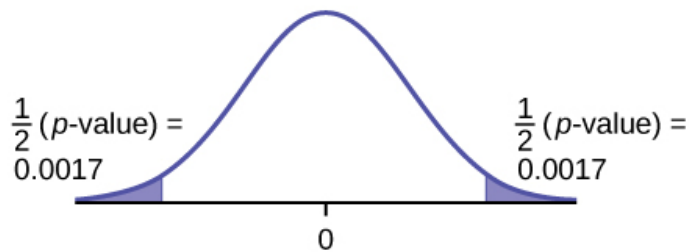
Distribution:

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$

The proportion of eReader users is different for the 16- to 29-year-old users from that of the 30 and older users.

Graph: two-tailed



p -value : 0.0033

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of eReader users 16 to 29 years old is different from the proportion of eReader users 30 and older.

13. Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown below.¹² Test at the 1% level of significance.

11. Data from Educational Resources, December catalog.

12. "State-Specific Prevalence of Obesity Among Adults—United States, 2007." MMWR, CDC. Available online at <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm> (accessed June 17, 2013).

Figure 8.32

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

14. Two computer users were discussing tablet computers. At one point in time a higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. The figure below details the number of tablet owners for each age group. Test at the 1% level of significance.

Figure 8.33

	16–29 year olds	30 years old and older
Own a Tablet	69	231
Sample Size	628	2,309

Test: two independent sample proportions

Random variable: $p_1 - p_2$

Distribution:

$H_0: p_1 = p_2$

$H_a: p_1 > p_2$

A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Graph: right-tailed

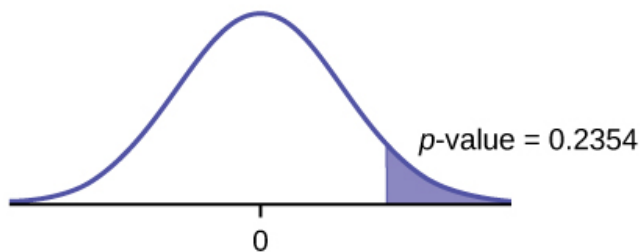


Figure 8.34

p -value: 0.2354

Decision: Do not reject the H_0 .

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

15. A group of friends debated whether more men use smartphones than women. They consulted a research

study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones.¹³ Test at the 5% level of significance.

16. While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Subscripts: 1: men; 2: women

- a. $H_0: p_1 \leq p_2$
 - b. $H_a: p_1 > p_2$
 - c. $p_1 - p_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
 - d. normal for two proportions
 - e. test statistic: 0.22
 - f. p -value: 0.4133
 - g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Do not reject the null hypothesis.
 - iii. Reason for Decision: p -value > alpha
 - iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.
-

17. We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was \$31.14 with a standard deviation of \$4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was \$33.86 with a standard deviation of \$10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

18. Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as

13. "Smart Phone Users, By the Numbers." Visually, 2013. Available online at <http://visual.ly/smart-phone-users-numbers> (accessed June 17, 2013).

high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

- a. $H_0: p_1 = p_2$
 - b. $H_a: p_1 \neq p_2$
 - c. $p_1 - p_2$ is the difference between the proportions of men and women that have at least one pierced ear.
 - d. normal for two proportions
 - e. test statistic: -4.82
 - f. p -value: zero
 - g. Check student's solution.
 - i. Alpha: 0.05
 - ii. Decision: Reject the null hypothesis.
 - iii. Reason for Decision: p -value < alpha
 - iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.
-

19. "To Breakfast or Not to Breakfast?" by Richard Ayore: In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, ..., birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data below solve our problem.

Figure 8.35

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

- $H_0: \mu_d = 0$
- $H_a: \mu_d > 0$
- The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
- t_9
- test statistic: 4.8963
- p -value: 0.0004
- Check student's solution.
 - Alpha: 0.05
 - Decision: Reject the null hypothesis.
 - Reason for Decision: p -value < alpha
 - Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

References

Image References

Figure 8.10: Figure 10.5.3 from LibreTexts Introductory Statistics (2020) (CC BY 4.0). Retrieved from [https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_\(OpenStax\)/10%3A_Hypothesis_Testing_with_Two_Samples/10.05%3A_Matched_or_Paired_Samples](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(OpenStax)/10%3A_Hypothesis_Testing_with_Two_Samples/10.05%3A_Matched_or_Paired_Samples)

Figure 8.14: Figure 10.19 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-solutions#fs-idm5499280-solution>

Figure 8.19: Figure 10.23 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-solutions#fs-idm5499280-solution>

Figure 8.22: Figure 10.24 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-solutions#fs-idm5499280-solution>

Figure 8.27: Figure 10.8 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/10-3-comparing-two-independent-population-proportions>

Figure 8.28: Figure 10.9 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/10-3-comparing-two-independent-population-proportions>

Figure 8.29: Figure 10.6 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/10-2-two-population-means-with-known-standard-deviations>

Figure 8.30: Figure 10.17 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-practice#element-814>

Figure 8.31: Figure 10.21 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-solutions>

Figure 8.34: Figure 10.22 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/statistics/pages/10-solutions>

Text

Data from Educational Resources, December catalog.

Data from Hilton Hotels. Available online at <http://www.hilton.com> (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at <http://hyatt.com> (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at <http://www.cancer.org/index> (accessed June 17, 2013).

Data from the Chancellor's Office, California Community Colleges, November 1994.

"State of the States." Gallup, 2013. Available online at <http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive> (accessed June 17, 2013).

"West Nile Virus." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/ncidod/dvbid/westnile/index.htm> (accessed June 17, 2013).

CHAPTER 9: SIMPLE LINEAR REGRESSION

9.1 Introduction to Bivariate Data and Scatterplots

Learning Objectives

By the end of this chapter, the student should be able to:

- Display and describe relationships in bivariate data
- Describe bivariate data numerically
- Understand basic ideas of linear regression
- Predict future value using your regression line
- Understand the impact of influential points and outliers in the context of linear regression
- Apply ideas of inference to linear regression



Figure 9.1: Auto Mechanic Salaries. Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience.

Professionals often want to know how two (or more) numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in these examples is bivariate data — “bi” for two variables. In this chapter, you will be studying the “simple linear regression”. Note that this does not imply that these ideas are “simple” but just that we are working with one independent variable (x) and a linear relationship. This involves data that fits a line in two dimensions.

Bivariate Data

When we are looking at **bivariate data** we first need to decide, if possible, does changing one variable seems to lead to a change in the other. A **response variable** (also called y , dependent variable, predicted variable) measures or records an outcome of a study. An **explanatory variable** (also called x , independent variable, predictor variable) explains changes in the response variable.

When considering the relationship between two quantitative variables:

1. Start with a graph (scatterplot)
2. Look for an overall pattern and deviations from the pattern
3. Use numerical descriptions of the data and overall pattern (correlation, coefficient of determination)
4. Consider a mathematical model (regression)

Scatterplots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y . The most common and easiest way is a scatter plot. A scatter plot shows a lot about the relationship between the variables. When you look at a scatterplot, you want to notice the overall pattern and any potential deviations from the pattern. You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are together. When looking at a scatterplot you always want to note:

1. Shape
2. Trend
3. Strength

The following scatterplot examples illustrate these concepts.

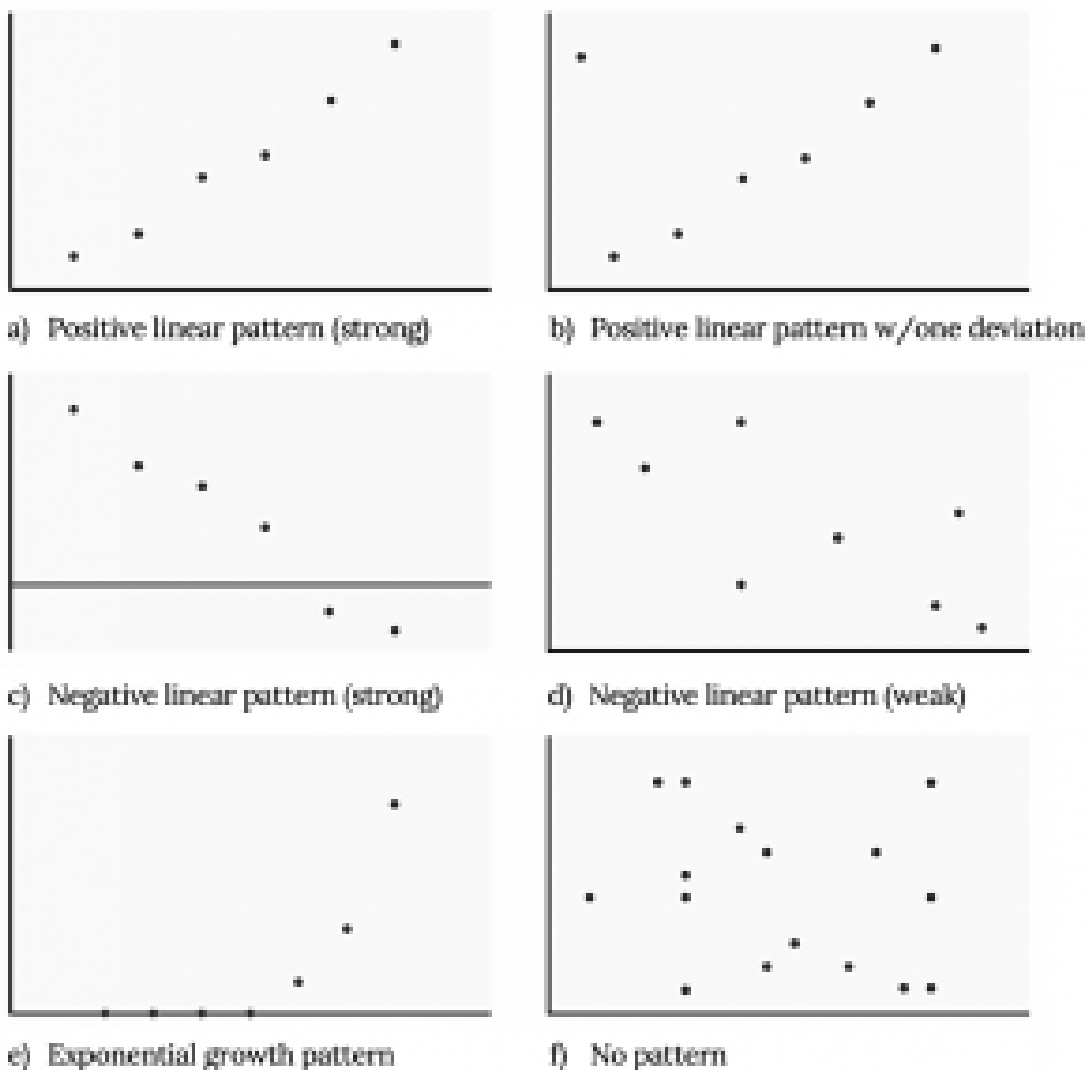


Figure 9.2: Scatterplot Configurations

Shape

Although we may see other shapes in a scatter plot, at this point we are only interested in applying these ideas when we see a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can will later be

calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable.

Trend

If we do see a linear pattern, what sort of relationship is there? A positive trend is seen when increasing x also increases y . On the other hand a negative (inverse) trend is seen when increasing x appears to cause y to decrease. In other words:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

Strength

At this point we can think about the strength of a relationship as how tightly do the points on a scatterplot fit the linear pattern. A stronger relationship has points clustered together closely while in a weaker one, points are more spread out. The strength of a relationship is not always apparent in a scatterplot but we will see numerical measures of this in the future.

Example

1. Does the scatter plot appear linear? Strong or weak? Positive or negative?

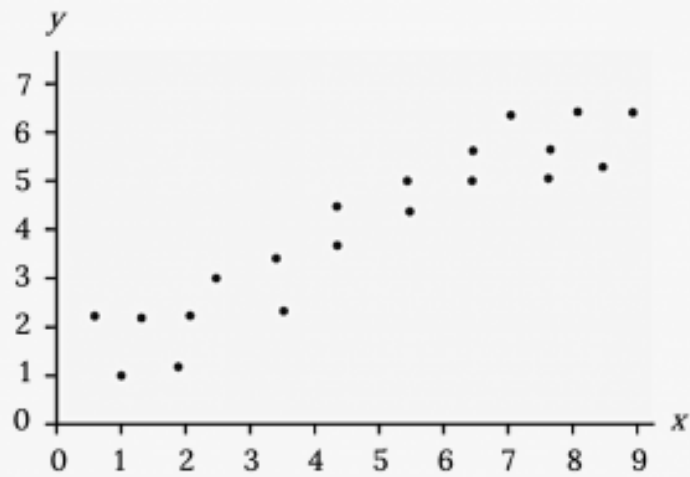


Figure 9.3: Scatterplot 1



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=289#h5p-172>

2. Does the scatter plot appear linear? Strong or weak? Positive or negative?

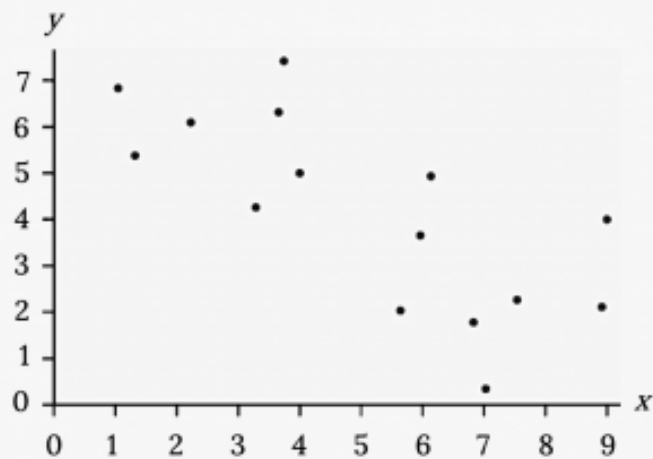


Figure 9.4: Scatterplot 2



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=289#h5p-173>

3. Does the scatter plot appear linear? Strong or weak? Positive or negative?

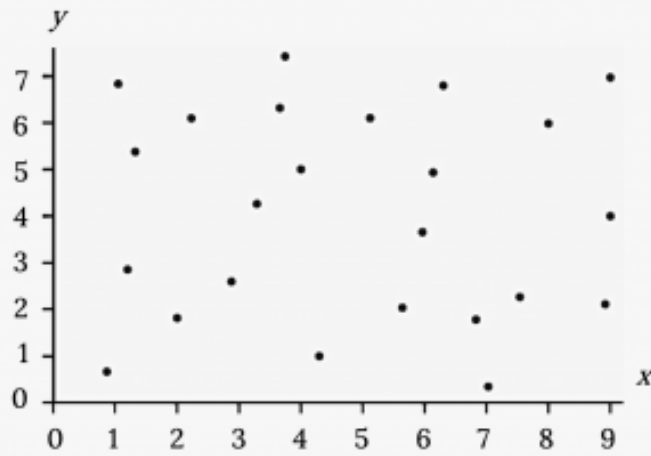


Figure 9.5: Scatterplot 3



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=289#h5p-174>

Your turn!

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

Figure 9.6: Amelia's Points

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Image References

Figure 9.1: Aaron Huber (2018). "Artist at Work." Public domain. Retrieved from <https://unsplash.com/photos/KxeFuXta4SE>

Figure 9.2: Kindred Grey via Virginia Tech (2020). "Figure 9.2" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.2.png . Adaptation of Figures 12.6, 12.7, and 12.8 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-2-scatter-plots>

Figure 9.3: Kindred Grey via Virginia Tech (2020). "Figure 9.3" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.3.png . Adaptation of Figure 12.26 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-practice>

Figure 9.4: Kindred Grey via Virginia Tech (2020). "Figure 9.4" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.4.png . Adaptation of Figure 12.27 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-practice>

Figure 9.5: Kindred Grey via Virginia Tech (2020). "Figure 9.5" CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.5.png . Adaptation of Figure 12.28 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-practice>

9.2 Measures of Association

Besides looking at the scatter plot and seeing that a linear relationship seems reasonable, and identifying a positive or negative trend, how can you tell more about this relationship? While it is always good practice to first examine things visually, you may find that deciphering a scatterplot, especially the strength of a relationship can be tricky. The next step is then to then calculate numerical measures of this association.

The Correlation Coefficient, r

The **correlation coefficient**, r , developed by Karl Pearson in the early 1900s, is a numerical measure that provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y .

The correlation coefficient can be calculated using the formula:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

where n = the number of data points.

The formula for r is formidable, so I would not recommend doing this by hand, however technology can make quick work of the calculation.

If you suspect a linear relationship between x and y , then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of r is always between -1 and $+1$: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y . Values of r close to -1 or to $+1$ indicate a stronger linear relationship between x and y .
- If $r = 0$ there is likely no linear correlation. It is important to view the scatterplot, however, because data that exhibit a curved or horizontal pattern may have a correlation of 0.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (positive correlation).
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (negative correlation).
- The sign of r is the same as the sign of the slope, b , of the best-fit line.

Note: Strong correlation does not suggest that x causes y or y causes x . We say “correlation does not imply causation.”

Example

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200.

Figure 9.7: Third and Final Exam Scores Data

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

A scatter plot showing the scores on the final exam based on scores from the third exam is as follows.

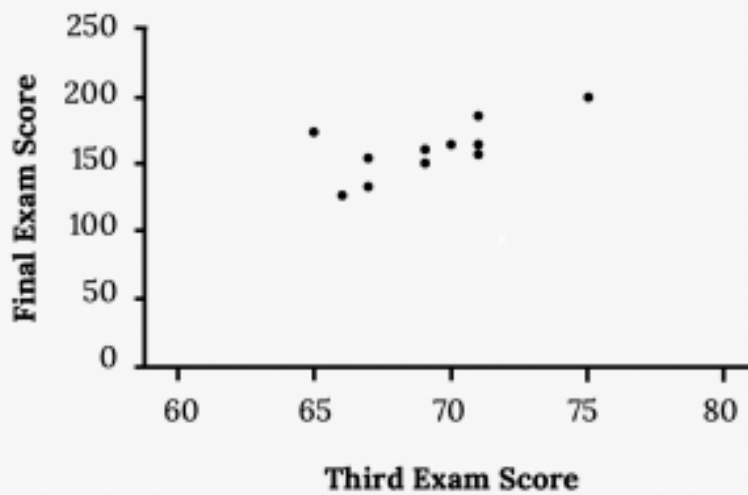


Figure 9.8: Third and Final Exam Scores Scatterplot

Find the correlation coefficient:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=293#h5p-175>

Your turn!

Match the following scatter plots with their description of correlation coefficient



Figure 9.9: Matching Scatterplots to Correlation Coefficients

1. $-1 < r < 0$
2. $r = 0$
3. $0 < r < 1$



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=293#h5p-176>

The Coefficient of Determination, r^2

The coefficient of determination, r^2 , is (obviously) the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2 , when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Example

Recall our previous example using a student's third exam scores to predict their final exam scores:

We found the correlation coefficient is $r = 0.6631$.

Find the coefficient of determination:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=293#h5p-177>

Interpret of r^2 in the context of this example:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=293#h5p-178>

Image References

Figure 9.8: Kindred Grey via Virginia Tech (2020). “Figure 9.8” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.8.png . Adaptation of Figure 12.9 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation>

Figure 9.9: Kindred Grey via Virginia Tech (2020). “Figure 9.9” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.9.png . Adaptation of Figure 12.13 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation>

9.3 Modeling Linear Relationships

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Imagine collecting data on this and constructing a scatterplot of the points on graph paper. Then draw a line that appears to “fit” the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of “best fit.” According to your equation, what is the predicted height for a pinky length of 2.5 inches? You have just started the process of linear regression.

Linear Regression

Data rarely perfectly fit a straight line, but we can be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to “fit” a straight line. This is called a Line of Best Fit or Least-Squares Line. This process of fitting the best-fit line is called **linear regression**.

The equation of the regression line is $\hat{y} = a + bx$

The \hat{y} is read “y hat” and is the estimated value of y. It is the value of y obtained using the regression line. It may or may not be equal to values of y observed from the data.

The sample means of the x values and the y values are \bar{x} and \bar{y} , respectively. The best fit line always passes through the point (\bar{x}, \bar{y}) .

The **slope**, b can be written as $b = r \left(\frac{s_y}{s_x} \right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

The **y-intercept**, a , can then be calculated by using the slope, and means of x and y.

Example

Recall our example:

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Figure 9.7 (repeat): Third and Final Exam Scores Data

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=296#h5p-179>

Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least-squares linear regression has the form (x, \hat{y}) .

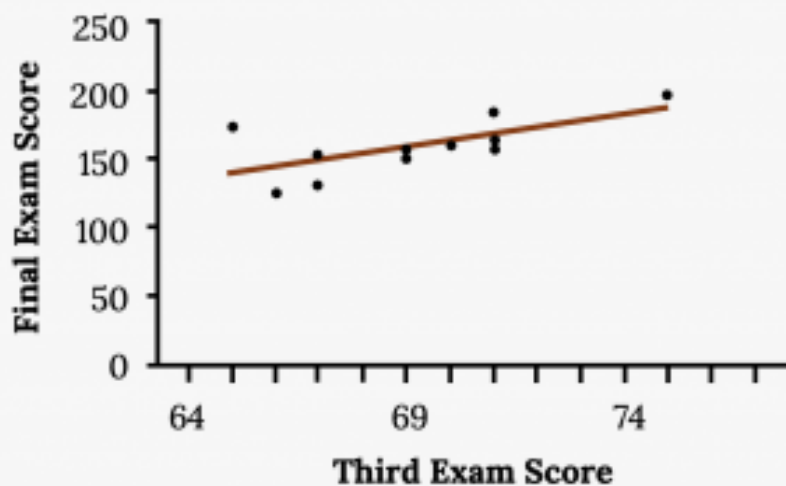


Figure 9.10: Line of Best Fit



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=296#h5p-180>

Your turn!

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the figure below show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

Figure 9.11: SCUBA Diver Stats

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Understanding Slope

The **slope** of the line, b , describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

Example

[Previous Example Continued]

The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Understanding the Y-Intercept

The **y-intercept** of the line, a , can tell us what we would predict the value of y to be when x is 0. This may make

sense in some cases, but in many it may not make sense for x to be equal to 0, therefore the y intercept may not be useful.

Example

[Previous Example Continued]

The y -intercept of the line is -173.51

Interpretation: In this context it does not really make sense for x to be 0 (unless a student did not take the exam or try at all). Therefore our y intercept does not make sense.

Prediction

The next, and most useful step in regression is to actually use that equation to predict future values of y .

Recall in our example we have examined the scatterplot and found the **correlation coefficient** and **coefficient of determination**. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Example

[Previous Example Continued]

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores (x -values) range from 65 to 75. Since 73 is between the x -values 65 and 75, substitute $x = 73$ into the equation. Then:

$$y = -173.51 + 4.83(73) = 179.08$$



An interactive H5P element has been excluded from this version of the text. You can view it



online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=296#h5p-181>

What would you predict the final exam score to be for a student who scored a 66 on the third exam?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=296#h5p-182>

What would you predict the final exam score to be for a student who scored a 90 on the third exam?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=296#h5p-183>

Your turn!

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Image References

Figure 9.10: Kindred Grey via Virginia Tech (2020). “Figure 9.10” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.10.png . Adaptation of Figure 12.11 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation>

9.4 Cautions about Regression

While regression is a very useful and powerful tool, it is also commonly misused. The main things we need to keep in mind when interpreting our results are:

1. Linearity assumption
2. Association And/Or correlation do not mean Causation
3. Extrapolation
4. Outliers and influential points

Linearity

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use the methods we are discussing.

Correlation Does Not Imply Causation

Even when we do have an apparent linear relationship and find a reasonable value of r , there can always be confounding or lurking variables at work. Be wary of spurious correlations and make sure the connection you are making makes sense!

There are also often situations where it may not be clear which variable is causing which. Does lack of sleep lead to higher stress levels or does high stress levels lead to lack of sleep? Which came first, the chicken or the egg? Sometimes these may not be answerable, but at least we are able to show an association there.

Extrapolation

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x -values in the sample data, but not necessarily for x -values outside that domain. The process of predicting inside of the observed x values observed in the data is called interpolation. The process of predicting outside of the observed x values observed in the data is called **extrapolation**.

Recall our example from the previous section. You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score

for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x -values in the sample data, which are between 65 and 75.

To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution $x = 90$ into the equation.

$$y = -173.51 + 4.83(90) = 261.19$$

The final-exam score is predicted to be 261.19. The largest a final-exam score could be is 100.

Outliers and Influential Points

In some data sets, there are values (observed data points) that may appear to be outliers x or y . **Outliers** are points that seem to stick out from the rest of the group in a single variable. Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that do not follow the trend of the rest of the data. These points may have a big effect on the calculation of the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

How do we handle these unusual points? Sometimes they should not be included in the analysis of the data. It is possible that an outlier or influential point is a result of erroneous data. Other times it may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier and/or influential point.

Identifying Outliers and/or Influential Points

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

We know how to find outliers in a single variable using fence rules and boxplots. However, we would like some guideline as to how far away a point needs to be in order to be considered an influential point. They also have large “errors”, where the “error” or residual is the vertical distance from the line to the point. As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. The graphical procedure is shown in the example below, followed by the numerical calculations in the next example. You would generally need to use only one of these methods.

Example

Continuing with the example from the previous section, you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the SSE should be smaller and the correlation coefficient ought to be closer to 1 or -1 .

Here it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to $2s$ or more, then we would consider the data point to be “too far” from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

- $\hat{y} = -173.5 + 4.83x$ is the line of best fit.
- Let Y2 = $-173.5 + 4.83x - 2(16.4)$
- Let Y3 = $-173.5 + 4.83x + 2(16.4)$

Notice Y2 and Y3 have the same slope as the line of best fit.

If we graph the scatterplot with the best fit line in equation Y1, and the two extra lines as Y2 and Y3, you will find that the only data point that is not between lines Y2 and Y3 is the point $x = 65$, $y = 175$. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than two standard deviations away from the best-fit line.

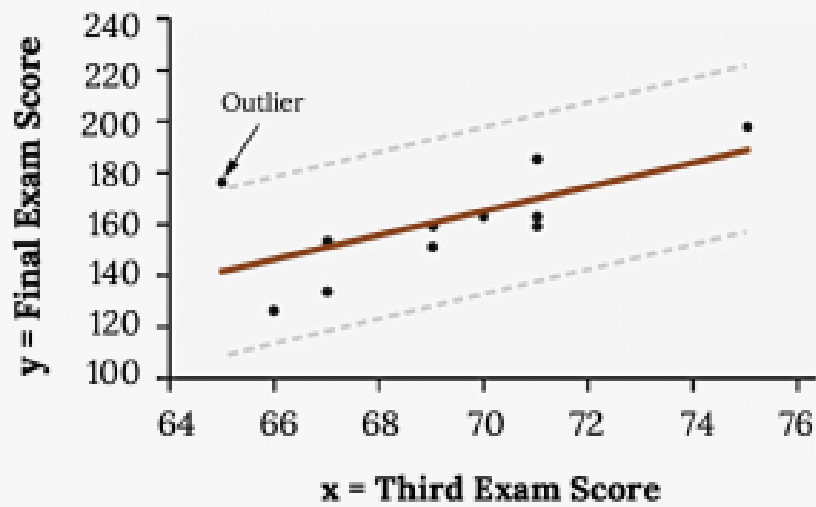


Figure 9.12: One Method of Identifying Outliers in Scatterplots

Your turn!

Identify the potential outlier in the scatter plot by drawing two separate lines. Suppose the standard deviation of the residuals or errors (s) is approximately $s=8.6$.

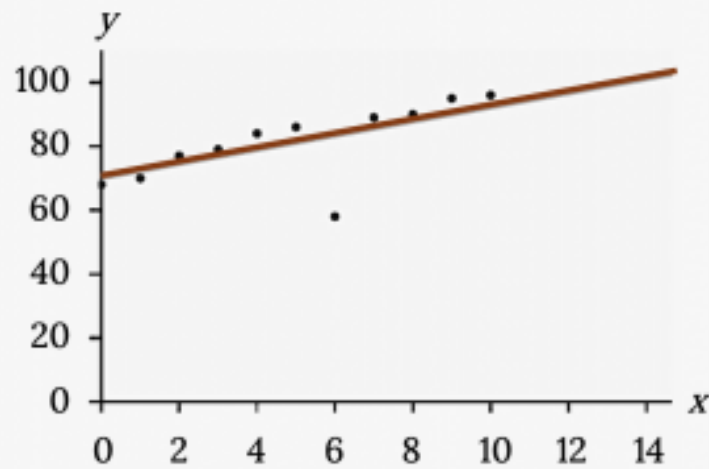


Figure 9.13: Identify the Outlier

Residuals

In the process of numerically identifying outliers and influential points, one of the most important tools we have is called the **residual**. It is found by $y_0 - \hat{y}_0 = \varepsilon_0$ (ε = the Greek letter epsilon) and is called the “error”. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y . In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y . If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y .

In the diagram below, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

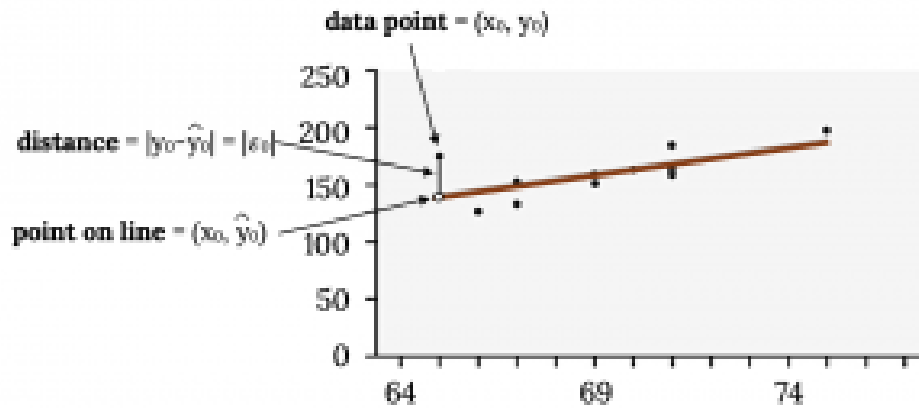


Figure 9.14: Residuals Diagram

Points that fall far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line then we call it an influential point. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line. Let's see how to do this mathematically:

Example

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \epsilon_i$ for $i = 1, 2, 3, \dots, 11$. Each $|\epsilon_i|$ is a vertical distance. In the following table, the first two columns are the third-exam and final-exam data. The third column shows the predicted \hat{y} values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, have been calculated in the fourth column of the table: observed y value–predicted y value = $y - \hat{y}$.

Figure 9.15: Calculating Residuals

x	y	\hat{y}	$y - \hat{y}$
65	175	140	$175 - 140 = 35$
67	133	150	$133 - 150 = -17$
71	185	169	$185 - 169 = 16$
71	163	169	$163 - 169 = -6$
66	126	145	$126 - 145 = -19$
75	198	189	$198 - 189 = 9$
67	153	150	$153 - 150 = 3$
70	163	164	$163 - 164 = -1$
71	159	169	$159 - 169 = -10$
69	151	160	$151 - 160 = -9$
69	159	160	$159 - 160 = -1$

For this example, there are 11 ϵ values. If you square each ϵ and add, you get:

$$(\epsilon_1)^2 + (\epsilon_2)^2 + \dots + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon_i^2$$

This is called the Sum of Squared Errors (SSE).

For our example the calculation is as follows:

First, **square each $|y - \hat{y}|$**

The squares are $35^2 17^2 16^2 6^2 19^2 9^2 3^2 1^2 10^2 9^2 1^2$

Then, add (sum) all the $|y - \hat{y}|$ squared terms using the formula

$$\sum_{i=1}^{11} \left(|y_i - \hat{y}_i| \right)^2 = \sum_{i=1}^{11} \epsilon_i^2 \text{ (Recall that } y_i - \hat{y}_i = \epsilon_i \text{.)}$$

$$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$$

$= 2440 = \text{SSE}$. The result, **SSE** is the Sum of Squared Errors.

s is the standard deviation of all the $y - \hat{y} = \epsilon$ values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{SSE}{n-2}}$$

Note: We divide by $(n - 2)$ as our df because the regression model involves two estimates.

For our example:

$$s = \sqrt{\frac{2440}{11-2}} = 16.47.$$

Note: Rather than calculate these ourselves, we can find s using the computer or calculator.

More on Influential Points

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $2s$, then we would consider the data point to be “too far” from the line of best fit. We call that point a potential influential point.

Back to our example, multiply s by 2:

$$(2)(16.47) = 32.94$$

32.94 is 2 standard deviations away from the mean of the $y - \hat{y}$ values.

So for this example, if any of the $|y - \hat{y}|$ values are **at least** 32.94, the corresponding (x, y) data point is a potential outlier.

We are looking for all data points for which the residual is greater than $2s = 2(16.4) = 32.8$ or less than -32.8 . Compare these values to the residuals in column four of the table. It appears all the $|y - \hat{y}|$'s are less than 31.29 except for the first one which is 35.

$$35 > 31.29 \text{ That is, } |y - \hat{y}| \geq (2)(s)$$

The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line? Numerically and graphically, we have identified the point (65, 175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.

The next step is to compute a new best-fit line using the ten remaining points. The new line of best fit and the correlation coefficient are:

$$\hat{y} = -355.19 + 7.39x \text{ and } r = 0.9121$$

The new line with $r = 0.9121$ is a stronger correlation than the original ($r = 0.6631$) because $r = 0.9121$ is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score. The point we deleted appeared to be an influential point

It is often tempting to remove outliers and influential points. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm

ignored the largest market swings – the “outliers” – they would soon go bankrupt by making poorly thought-out investments.

When outliers are deleted, the researcher should either record that data was deleted, and why, or the researcher should provide results both with and without the deleted data. If data is erroneous and the correct values are known (e.g., student one actually scored a 70 instead of a 65), then this correction can be made to the data.

Using this new line of best fit (based on the remaining ten data points in the “exam example” used in previous sections, what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

- Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.
- The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Your turn!

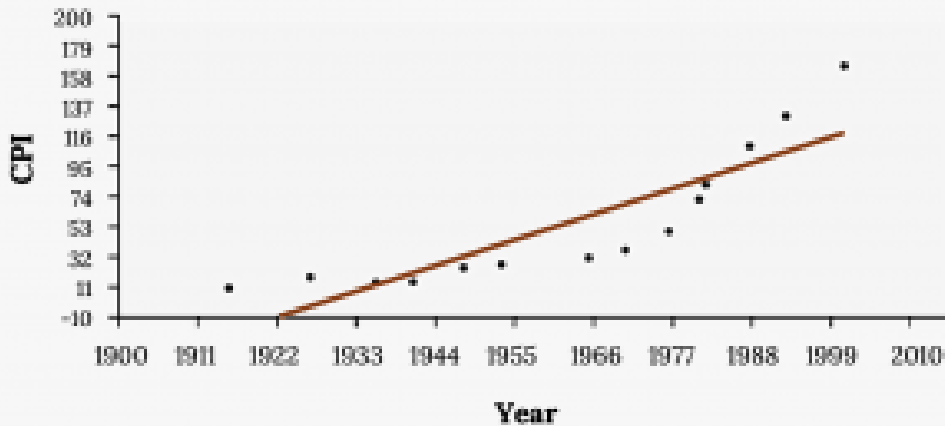
The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation’s economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI’s trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Figure 9.16: CPI Values

x	y	x	y
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

- a. Draw a scatterplot of the data.

- b. Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
- c. Draw the line on the scatterplot.
- d. Find the correlation coefficient.



- e. What is the average CPI for the year 1990?
- f. Comment on the appropriateness of this linear model. Do there appear to be any outliers or influential points?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://ecampusontario.pressbooks.pub/significantstats/?p=302#h5p-184>

Image References

Figure 9.12: Kindred Grey via Virginia Tech (2020). “Figure 9.12” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.12.png . Adaptation of Figure 12.18 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-6-outliers>

Figure 9.13: Kindred Grey via Virginia Tech (2020). “Figure 9.13” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.13.png . Adaptation of Figure 12.19 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-6-outliers>

Figure 9.14: Kindred Grey via Virginia Tech (2020). “Figure 9.14” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.14.png . Adaptation of Figure 12.10 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation>

Figure 9.17: Kindred Grey via Virginia Tech (2020). “Figure 9.17” CC BY-SA 4.0. Retrieved from https://commons.wikimedia.org/wiki/File:Figure_9.17.png . Adaptation of Figure 12.20 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-6-outliers>

9.5 Inference for Regression

The previous sections in this chapter have focused on linear regression as a tool for summarizing trends in data and making predictions. These numerical summaries are analogous to the methods discussed in Chapter 2 for displaying and summarizing data. Regression is also used to make inferences about a population. The same ideas covered in Chapters 6–8 about using data from a sample to draw inferences about population parameters can also apply to regression. Previously, the goal was to draw inference about a population parameter such as μ or p . In regression, the population parameter of interest is typically the slope parameter β . Inference about the intercept term is rare, but can be done where the vertical intercept is meaningful.

Each of the elements we have calculated in your regression equation are actually point estimates of corresponding population parameters and have their own sampling distributions

- r is an unbiased point estimate of the population correlation, ρ
- a is an unbiased estimate for the Y-intercept, β_0
- b is an unbiased estimate for slope, β_1
- \hat{y} is an unbiased estimate for mean response, μ_y

So a set of ordered pairs (x, y) used when fitting a least squares regression line are assumed to have been sampled from a population in which the relationship between the explanatory and response variables with the following equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where ε , the error, is assumed to have a normal distribution with mean 0 and standard deviation σ ($\varepsilon \sim N(0, \sigma)$).

Inference for regression Assumptions:

Like any inferential technique, we need to meet an underlying set of assumptions in order to appropriately use them.

1. Like always, the data are collected from a well-designed, random sample or randomized experiment.
2. The relationship is Linear
3. The standard deviation of y is the same for all values of x .
4. The response y varies Normally around its mean
5. The residual errors are independent (no pattern).

You can check most of these using a residual plot.

Regression Standard Errors

Statistical software is typically used to obtain t-statistics and p-values for inference with regression, since using the formulas for calculating standard error can be cumbersome. However, the formulas are displayed below

The standard error of the residuals is:

$$s = \sqrt{\frac{\sum \text{residual}^2}{n - 2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

Notice **n-2** degrees of freedom!

Using that to calculate the standard error of the sampling distribution of the slope, b:

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

We can use this to do perform inference on the slope.

Inference on the slope

We now know the sampling distribution of the slope, b, is:

$$b_1 \sim t(\beta_1, \frac{s}{\sqrt{\sum (x - \bar{x})^2}}) \text{ w/DF} = n - 2$$

Hypothesis tests and confidence intervals for regression parameters have the same basic form as tests and intervals about population means.

Hypothesis tests on the slope

We often times want to check for the significance of the slope of a regression equation. In other words is the slope actually different from 0? In this case your Null Hypothesis would be:

$$H_0 : \beta_1 = 0$$

We can use the test statistic:

$$t = \frac{b_1}{SE_{b_1}}$$

NOTES:

- Since our slope, b , is calculated directly from the correlation coefficient, r , testing for a significant slope is essentially the same as testing for significance of the correlation coefficient
- It is possible to test for values other than 0 in the Null

Confidence Intervals for the slope

Finally, we may also be interested in estimating the true value of the slope with a confidence interval. We know the sampling distribution of the slope and can follow the same format we are familiar with for a Confidence Interval for the slope, b :

$$b_1 \pm t^* SE_{b_1} \text{ w/ } DF = n - 2$$

We interpret this interval similar to how we have before. For instance you could be 95% confident the interval you've created captures the true value of the slope.

Chapter Wrap Up

Concept Check



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://ecampusontario.pressbooks.pub/significantstats/?p=312#h5p-185>

Section Reviews

9.1 Introduction to Bi-variate Data and Scatterplots

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

9.2 Measures of Association

The correlation coefficient r measures the strength of the linear association between x and y . The variable r has to be between -1 and $+1$. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2 , is equal to the square of the correlation coefficient. When expressed as a

percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

9.3 Modeling Linear Relationships

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y . The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

9.4 Cautions about Regression

To determine if a point is an outlier, do one of the following:

1. Must have a linear relationship to use these methods!
2. Correlation is not causation
3. Be careful with Extrapolation
4. Beware of Influential Points

9.5 Inference for Regression

We can apply our inference techniques to regression, especially for the slope.

Key Terms

Try to define the terms below on your own. Scroll over any term to check your response!

9.1 Introduction to Bi-variate Data and Scatterplots

- **Bivariate data**
- **Response variable**
- **Explanatory variable**

9.2 Measures of Association

- **Correlation coefficient (r)**

9.3 Modeling Linear Relationships

- **Linear regression**
- **Slope**
- **Y-intercept**
- **Correlation coefficient**
- **Coefficient of determination**

9.4 Cautions about Regression

- **Extrapolation**
- **Outlier**
- **Influential point**
- **Residual**

Extra Practice

9.1 Introduction to Bi-variate Data and Scatterplots

1. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The figure below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Figure 9.19

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

2. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data.

Figure 9.20

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

3. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on

mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data. Note that tuition is the independent variable and salary is the dependent variable.

Figure 9.21

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

9.2 Measures of Association

1. Can a coefficient of determination be negative? Why or why not?
2. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The figure below shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Figure 9.19

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Find:

- r
- r^{2x}

3. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data.

Figure 9.20

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Find:

- r
- r^2

4. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data. Note that tuition is the independent variable and salary is the dependent variable.

Figure 9.21

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Find:

- r

- r^2

9.3 Modeling Linear Relationships

1. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

Figure 9.22

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

- a. Draw a scatter plot of the data.
- b. Use regression to find the equation for the line of best fit.
 - $\hat{y} = 2.23 + 1.99x$
- c. Draw the line of best fit on the scatter plot.
- d. What is the slope of the line of best fit? What does it represent?
 - The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another \$1.99 million in pay each year.
- e. What is the y-intercept of the line of best fit? What does it represent?
- f. What does an r value of zero mean?
 - It means that there is no correlation between the data sets.
- g. When $n = 2$ and $r = 1$, are the data significant? Explain.
- h. When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.
 - Yes, there are enough data points and the value of r is strong enough to show that there is a strong negative correlation between the data sets.

2. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

9.4 Cautions about Regression

1. The following table shows economic development measured in per capita income PCINC.

Figure 9.23

Year	PCINC	Year	PCINC
1870	340	1920	1050
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

- What are the independent and dependent variables?
 - Draw a scatter plot.
 - Use regression to find the line of best fit and the correlation coefficient.
 - Interpret the significance of the correlation coefficient.
 - Is there a linear relationship between the variables?
 - Find the coefficient of determination and interpret it.
 - What is the slope of the regression equation? What does it mean?
 - Use the line of best fit to estimate PCINC for 1900, for 2000.
 - Determine if there are any outliers.
-

2. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

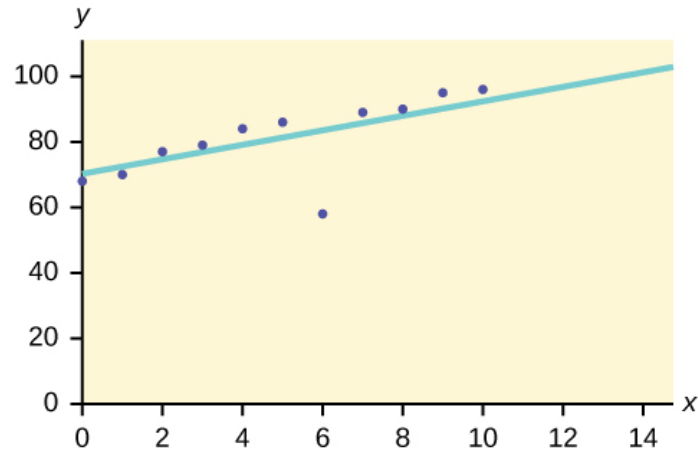


Figure 9.24

a. Do there appear to be any outliers?

- Yes, there appears to be an outlier at (6, 58).

b. A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

c. What effect did the potential outlier have on the line of best fit?

- The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

d. Are you more or less confident in the predictive ability of the new line of best fit?

e. The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?

- $s = 1.75$

f. The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

3. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Figure 9.25

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

- Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables?
- Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
- Find the correlation coefficient. Is it significant?
- Find the estimated heights for 32 stories and for 94 stories.
- Based on the data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
- Are there any outliers in the data? If so, which point(s)?
- What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
- Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
- What is the slope of the least squares (best-fit) line? Interpret the slope.

4. Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return: 74, 66, 81, 52, 73, 62, 52, 45, 62, 46, 60, 46, 38

Percent new: 5, 6, 8, 11, 12, 15, 16, 17, 18, 18, 19, 20, 20

- Enter the data into your calculator and make a scatter plot.
- Use your calculator’s regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
- Explain in words what the slope and y-intercept of the regression line tell us.
- How well does the regression line fit the data? Explain your response.

- e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
- f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?
- Solution to a and b: Check student's solution.
 - Solution to c: The slope of the regression line is -0.3031 with a y-intercept of 31.93 . In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 32% new sparrow hawks, which doesn't make sense since if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 30.3%.
 - Solution to d: If we examine r^2 , we see that only 57.52% of the variation in the percent of new birds is explained by the model and the correlation coefficient, $r = -0.7584$ only indicates a somewhat strong correlation between returning and new percentages.
 - Solution to e: The ordered pair $(66, 6)$ generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.98%. If we remove this data pair, we see only an adjusted slope of -0.2789 and an adjusted intercept of 30.9816 . In other words, even though this data generates the largest residual, it is not an outlier, nor is the data pair an influential point.
 - Solution to f: If there are 70% returning birds, we would expect to see $y = -0.2789(70) + 30.9816 = 0.114$ or 11.4% new birds in the colony.

5. The following table shows data on average per capita coffee consumption and heart disease rate in a random sample of 10 countries.

Figure 9.26

Yearly coffee consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

- a. Enter the data into your calculator and make a scatter plot.
- b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
- c. Explain in words what the slope and y-intercept of the regression line tell us.
- d. How well does the regression line fit the data? Explain your response.
- e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
- f. Do the data provide convincing evidence that there is a linear relationship between the amount of coffee consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

6. The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Figure 9.27

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

- Enter the data into your calculator and make a scatter plot.
- Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
- Explain in words what the slope and y-intercept of the regression line tell us.
- How well does the regression line fit the data? Explain your response.
- Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

- Check student's solution.
- Check student's solution.
- We have a slope of -1.4946 with a y-intercept of 193.88 . The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the y-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim $2,000$ meters, the less effort the heart puts out), the y-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
- Since only 1.5% of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
- The point $(34.72, 124)$ generates the largest residual of -11.82 . This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes -2.953 with the y-intercept changing to 247.1616 . While the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the y-intercept becomes more meaningful.

7. A researcher is investigating whether population impacts homicide rate. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are white males.

Figure 9.28

Population Size	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

- Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
 - Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
 - Discuss what the following mean in context.
 - The slope of the regression equation
 - The y-intercept of the regression equation
 - The correlation r
 - The coefficient of determination r^2 .
 - Do the data provide convincing evidence that there is a linear relationship between population size and homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.
-

8. Use the table below to answer (a) and (b).

Figure 9.29

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

- a. Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.
- b. If we remove the two service academies (the tuition is \$0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.
-

9. The average number of people in a family that attended college for various years is given below.

Figure 9.30

Year	Number of Family Members Attending College
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

- a. Using “year” as the independent variable and “Number of Family Members Attending College” as the dependent variable, draw a scatter plot of the data.
- b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- c. Find the correlation coefficient. Is it significant?
- d. Pick two years between 1969 and 1991 and find the estimated number of family members attending

college.

- e. Based on the data, is there a linear relationship between the year and the average number of family members attending college?
 - f. Using the least-squares line, estimate the number of family members attending college for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
 - g. Are there any outliers in the data?
 - h. What is the estimated average number of family members attending college for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
 - i. What is the slope of the least squares (best-fit) line? Interpret the slope.
-

10. The percent of female wage and salary workers who are paid hourly rates is given in below for the years 1979 to 1992.

Figure 9.31

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- a. Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
- b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
- c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- d. Find the correlation coefficient. Is it significant?
- e. Find the estimated percents for 1991 and 1988.
- f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
- g. Are there any outliers in the data?
- h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
- i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

- Check student's solution.
- yes
- $\hat{y} = -266.8863 + 0.1656x$
- 0.9448; Yes
- 62.8233; 62.3265
- yes
- no; (1987, 62.7)
- 72.5937; no
- slope = 0.1656.

As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

11. The cost of a leading liquid laundry detergent in different sizes is given below.

Figure 9.32

Size (ounces)	Cost (\$)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

- Complete the table for the cost per ounce of the different sizes.
- Using “size” as the independent variable and “cost per ounce” as the dependent variable, draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- Find the correlation coefficient. Is it significant?
- If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
- If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
- Does it appear that a line is the best way to fit the data? Why or why not?
- Are there any outliers in the the data?
- Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost per ounce? Why or why not?
- What is the slope of the least-squares (best-fit) line? Interpret the slope.

- **Figure 9.33**

Size (ounces)	Cost (\$)	cents/oz
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

- Check student's solution.
- There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
- $\hat{y} = 20.2368 - 0.0819x$
- $r = -0.8086$
- 40-oz: 16.96 cents/oz
- 90-oz: 12.87 cents/oz
- The relationship is not linear; the least squares line is not appropriate.
- no outliers
- No, you would be extrapolating. The 300-oz size is outside the range of x .
- slope = -0.08194 ; for each additional ounce in size, the cost per ounce decreases by 0.082 cents.

12. According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Figure 9.34

Net Taxable Estate (\$)	Approximate Probate Fees and Taxes (\$)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
- Find the correlation coefficient. Is it significant?
- Find the estimated total cost for a next taxable estate of \$1,000,000. Find the cost for \$2,500,000.
- Does it appear that a line is the best way to fit the data? Why or why not?

- h. Are there any outliers in the data?
 - i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
 - j. What is the slope of the least-squares (best-fit) line? Interpret the slope.
-

13. The following are advertised sale prices of color televisions at Anderson's.

Figure 9.35

Size (inches)	Sale Price (\$)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

- a. Decide which variable should be the independent variable and which should be the dependent variable.
 - b. Draw a scatter plot of the data.
 - c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
 - d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
 - e. Find the correlation coefficient. Is it significant?
 - f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
 - g. Does it appear that a line is the best way to fit the data? Why or why not?
 - h. Are there any outliers in the data?
 - i. What is the slope of the least-squares (best-fit) line? Interpret the slope.
- Size is x , the independent variable, price is y , the dependent variable.
 - Check student's solution.
 - The relationship does not appear to be linear.
 - $\hat{y} = -745.252 + 54.75569x$
 - $r = 0.8944$, yes it is significant
 - 32-inch: \$1006.93, 50-inch: \$1992.53
 - No, the relationship does not appear to be linear. However, r is significant.
 - no, the 60-inch TV
 - For each additional inch, the price increases by \$54.76
-

14. The figure below shows the average heights for American boys in 1990.

Figure 9.36

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

- Decide which variable should be the independent variable and which should be the dependent variable.
 - Draw a scatter plot of the data.
 - Does it appear from inspection that there is a relationship between the variables? Why or why not?
 - Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
 - Find the correlation coefficient. Is it significant?
 - Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
 - Does it appear that a line is the best way to fit the data? Why or why not?
 - Are there any outliers in the data?
 - Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?
 - What is the slope of the least-squares (best-fit) line? Interpret the slope.
-

15. Use the table below to answer (a)–(n).

Figure 9.37

State	# letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

- What are the independent and dependent variables?
 - What do you think the scatter plot will look like? Make a scatter plot of the data.
 - Does it appear from inspection that there is a relationship between the variables? Why or why not?
 - Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
 - Find the correlation coefficient. What does it imply about the significance of the relationship?
 - Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
 - Use the two points in part f to plot the least-squares line on your graph from part b.
 - Does it appear that a line is the best way to fit the data? Why or why not?
 - Are there any outliers?
 - Use the least squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
 - Delete “Hawaii” and substitute “Alaska” for it. Alaska is the forty-ninth, state with an area of 656,424 square miles.
 - Calculate the new least-squares line.
 - Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that’s the case?
 - Do you think that, in general, newer states are larger than the original states?
- Let rank be the independent variable and area be the dependent variable.
 - Check student’s solution.
 - There appears to be a linear relationship, with one outlier.
 - $\hat{y}(\text{area}) = 24177.06 + 1010.478x$
 - $r = 0.50047$, r is not significant so there is no relationship between the variables.
 - Alabama: 46407.576 Colorado: 62575.224
 - Alabama estimate is closer than Colorado estimate.

- If the outlier is removed, there is a linear relationship.
- There is one outlier (Hawaii).
- rank 51: 75711.4; no

- **Figure 9.38**

Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

- $\hat{y} = -87065.3 + 7828.532x$
- Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.
- yes, with the exception of Hawaii

References

Image References

Figure 9.24: Figure 12.30 from OpenStax Introductory Statistics (2013) (CC BY 4.0). Retrieved from <https://openstax.org/books/introductory-statistics/pages/12-practice>

Text

Data from the House Ways and Means Committee, the Health and Human Services Department.

Data from Microsoft Bookshelf.

Data from the United States Department of Labor, the Bureau of Labor Statistics.

Data from the Physician's Handbook, 1990.

CLASS GROUP ACTIVITIES

Normal Distribution (Lap Times)

Normal Distribution (Lap Times)

Class Time:

Names:

Student Learning Outcome

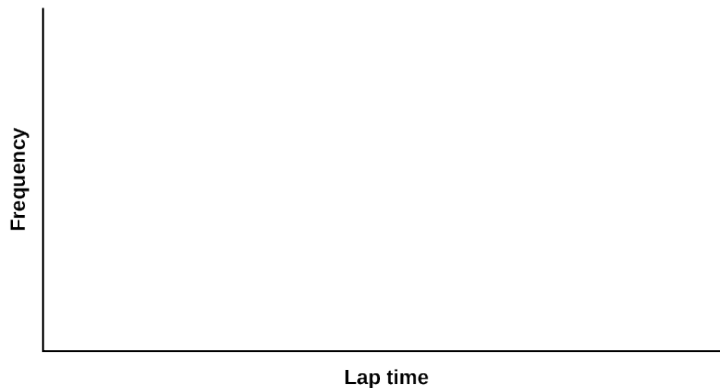
- The student will compare and contrast empirical data and a theoretical distribution to determine if Terry Vogel's lap times fit a continuous distribution.

Directions Round the relative frequencies and probabilities to four decimal places. Carry all other decimal answers to two places.

Collect the Data

- Use the data from [Appendix C](#). Use a stratified sampling method by lap (races 1 to 20) and a random number generator to pick six lap times from each stratum. Record the lap times below for laps two to seven.

- Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



- Calculate the following:

a. $\bar{x} = \underline{\hspace{2cm}}$

b. $s =$ _____

4. Draw a smooth curve through the tops of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution Using your sample mean, sample standard deviation, and histogram to help, what is the approximate theoretical distribution of the data?

- $X \sim$ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data Use the data you collected to complete the following statements.

- The IQR goes from _____ to _____.
- $IQR =$ _____. ($IQR = Q_3 - Q_1$)
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The empirical probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution Using the theoretical distribution, complete the following statements. You should use a normal approximation based on your sample data.

- The IQR goes from _____ to _____.
- $IQR =$ _____.
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this distribution.

Discussion Questions Do the data from the section titled [Collect the Data](#) give a close approximation to the theoretical distribution in the section titled [Analyze the Distribution](#)? In complete sentences and comparing the result in the sections titled [Describe the Data](#) and [Theoretical Distribution](#), explain why or why not.

Normal Distribution (Pinkie Length)

Normal Distribution (Pinkie Length)

Class Time:

Names:

Student Learning Outcomes

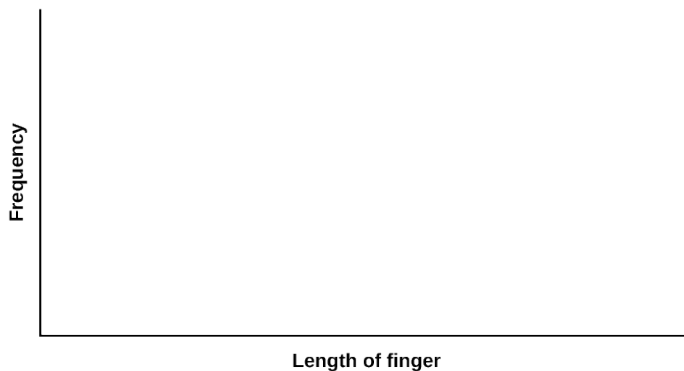
- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

Collect the Data Measure the length of your pinky finger (in centimeters).

1. Randomly survey 30 adults for their pinky finger lengths. Round the lengths to the nearest 0.5 cm.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the following.

a. \bar{x} = _____

b. s = _____

4. Draw a smooth curve through the top of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution Using your sample mean, sample standard deviation, and histogram, what was the approximate theoretical distribution of the data you collected?

- $X \sim \text{_____}(\text{_____,} \text{_____})$
- How does the histogram help you arrive at the approximate distribution?

Describe the Data Using the data you collected complete the following statements. (Hint: order the data)
Remember

$(IQR = Q_3 - Q_1)$

- $IQR = \text{_____}$
- The 15th percentile is _____.
- The 85th percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution Using the theoretical distribution, complete the following statements. Use a normal approximation based on the sample mean and standard deviation.

- $IQR = \text{_____}$
- The 15th percentile is _____.
- The 85th percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Discussion Questions Do the data you collected give a close approximation to the theoretical distribution? In complete sentences and comparing the results in the sections titled [Describe the Data](#) and [Theoretical Distribution](#), explain why or why not.

Central Limit Theorem (Pocket Change)

Central Limit Theorem (Pocket Change)

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate and compare properties of the central limit theorem.

Note

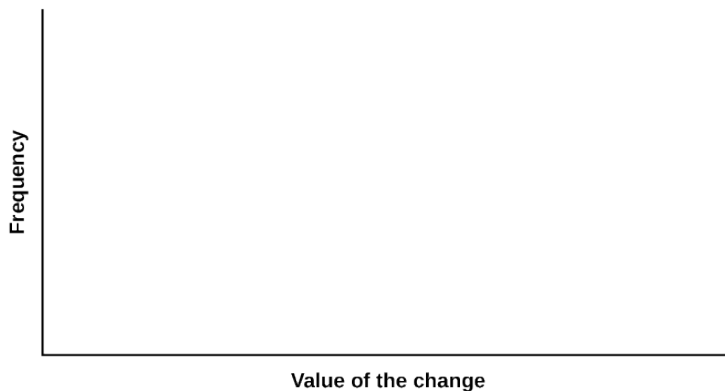
This lab works best when sampling from several classes and combining data.

Collect the Data

- Count the change in your pocket. (Do not include bills.)
- Randomly survey 30 classmates. Record the values of the change in [\(Figure\)](#).

-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----

- Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



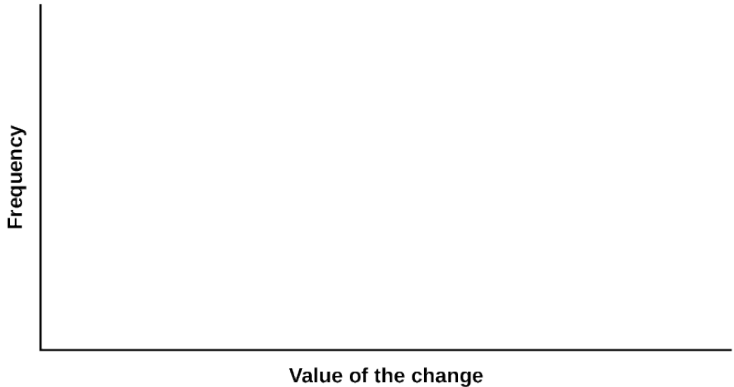
- Calculate the following ($n = 1$; surveying one person at a time):
 - $\bar{x} = \text{-----}$
 - $s = \text{-----}$
- Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences

to describe the general shape of the curve.

Collecting Averages of Pairs Repeat steps one through five of the section [Collect the Data](#), with one exception. Instead of recording the change of 30 classmates, record the average change of 30 pairs.

1. Randomly survey 30 **pairs** of classmates.
2. Record the values of the average of their change in [\(Figure\)](#).

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled [Collect the Data](#). Sketch the graph using a ruler and a pencil.

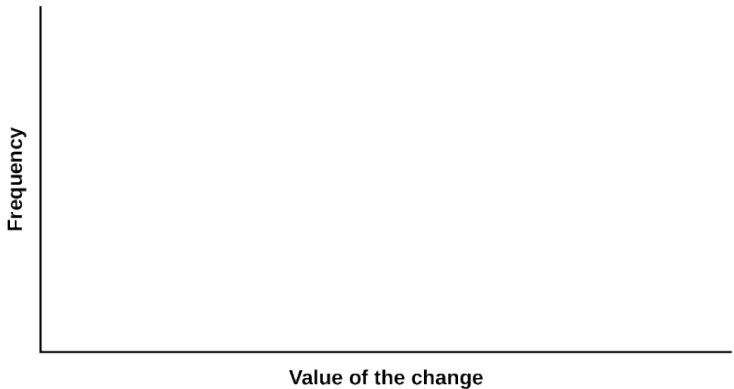


4. Calculate the following ($n = 2$; surveying two people at a time):
 - a. $\bar{x} =$ _____
 - b. $s =$ _____
5. Draw a smooth curve through tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Collecting Averages of Groups of Five Repeat steps one through five (of the section titled [Collect the Data](#)) with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of five.

1. Randomly survey 30 **groups of five** classmates.
2. Record the values of the average of their change.

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled [Collect the Data](#). Sketch the graph using a ruler and a pencil.



4. Calculate the following ($n = 5$; surveying five people at a time):
- $\bar{x} =$ _____
 - $s =$ _____
5. Draw a smooth curve through tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

- Why did the shape of the distribution of the data change, as n changed? Use one to two complete sentences to explain what happened.
- In the section titled [Collect the Data](#), what was the approximate distribution of the data? $X \sim$ _____(_____,_____)
- In the section titled [Collecting Averages of Groups of Five](#), what was the approximate distribution of the averages? $\bar{X} \sim$ _____(_____,_____)
- In one to two complete sentences, explain any differences in your answers to the previous two questions.

Central Limit Theorem (Cookie Recipes)

Central Limit Theorem (Cookie Recipes)

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate and compare properties of the central limit theorem.

Given X = length of time (in days) that a cookie recipe lasted at the Olmstead Homestead. (Assume that each of the different recipes makes the same quantity of cookies.)

Recipe #	X	Recipe #	X	Recipe #	X	Recipe #	X
1	1	16	2	31	3	46	2
2	5	17	2	32	4	47	2
3	2	18	4	33	5	48	11
4	5	19	6	34	6	49	5
5	6	20	1	35	6	50	5
6	1	21	6	36	1	51	4
7	2	22	5	37	1	52	6
8	6	23	2	38	2	53	5
9	5	24	5	39	1	54	1
10	2	25	1	40	6	55	1
11	5	26	6	41	1	56	2
12	1	27	4	42	6	57	4
13	1	28	1	43	2	58	3
14	3	29	6	44	6	59	6
15	2	30	2	45	2	60	5

Calculate the following:

a. $\mu_X =$ _____

b. $\sigma_X =$ _____

Collect the Data Use a random number generator to randomly select four samples of size $n = 5$ from the given population. Record your samples in [\(Figure\)](#). Then, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

- Complete the table:

Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means: \bar{x} = _____	\bar{x} = _____	\bar{x} = _____	\bar{x} = _____	

- Calculate the following:

- \bar{x} = _____
- $s\bar{x}$ = _____

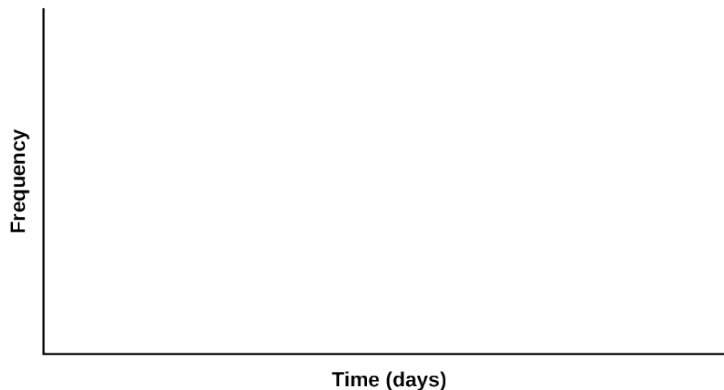
- Again, use a random number generator to randomly select four samples from the population. This time, make the samples of size $n = 10$. Record the samples in [\(Figure\)](#). As before, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups
Means: \bar{x} = _____	\bar{x} = _____	\bar{x} = _____	\bar{x} = _____	

- Calculate the following:

- \bar{x} = _____
- $s\bar{x}$ = _____

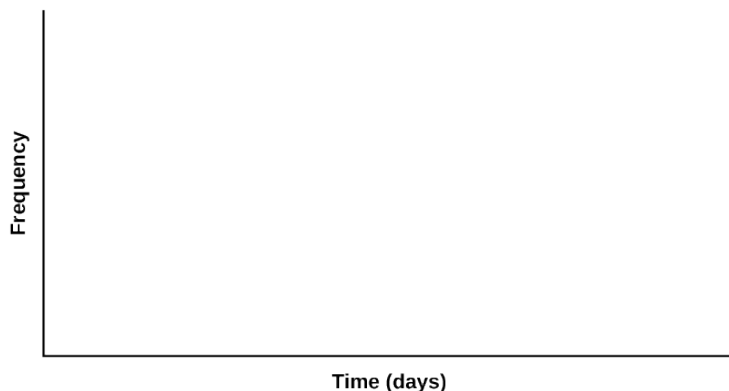
- For the original population, construct a histogram. Make intervals with a bar width of one day. Sketch the graph using a ruler and pencil. Scale the axes.



- Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n = 5$

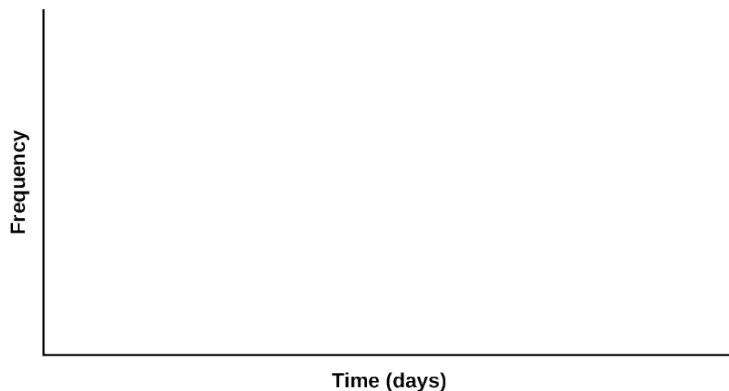
1. For the sample of $n = 5$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ a day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n = 10$

1. For the sample of $n = 10$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ a day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.
2. State the theoretical (according to the clt) distributions for the sample means.

a. $n = 5: \bar{X} \sim \text{-----}(\text{-----}, \text{-----})$

b. $n = 10: \bar{X} \sim \text{-----}(\text{-----}, \text{-----})$

3. Are the sample means for $n = 5$ and $n = 10$ “close” to the theoretical mean, μ_x ? Explain why or why not.
4. Which of the two distributions of sample means has the smaller standard deviation? Why?
5. As n changed, why did the shape of the distribution of the data change? Use one to two complete sentences to explain what happened.

Confidence Interval (Home Costs)

Confidence Interval (Home Costs)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate the 90% confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

Note

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

1. Complete the table:

-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----

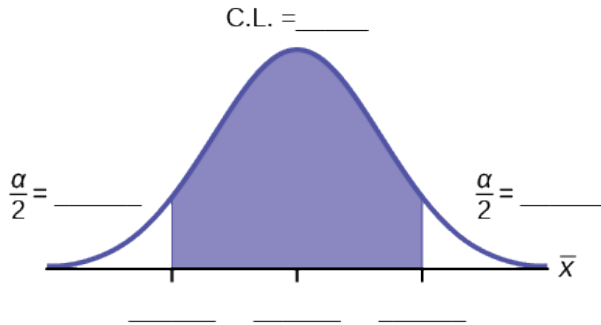
Describe the Data

1. Compute the following:
 - a. \bar{x} = -----
 - b. s_x = -----
 - c. n = -----
2. In words, define the random variable \bar{X} .
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.

- Confidence Interval: _____
 - Error Bound: _____
- How much area is in both tails (combined)? $\alpha =$ _____
 - How much area is in each tail? $\frac{\alpha}{2} =$ _____
 - Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.



- Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

- In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
- In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

- Using the given information, construct a confidence interval for each confidence level given.

Confidence level	EBM/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

- What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

Confidence Interval (Place of Birth)

Confidence Interval (Place of Birth)

Class Time:

Names:

Student Learning Outcomes

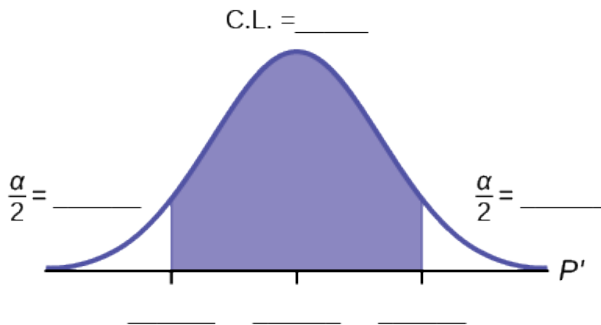
- The student will calculate the 90% confidence interval the proportion of students in this school who were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them if they were born in this state. Let X = the number that were born in this state.
 - a. n = _____
 - b. x = _____
2. In words, define the random variable P' .
3. State the estimated distribution to use.

Find the Confidence Interval and Error Bound

1. Calculate the confidence interval and the error bound.
 - a. Confidence Interval: _____
 - b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.



Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as though you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.
3. Construct a confidence interval for each confidence level given.

Confidence level	EBP/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

4. What happens to the EBP as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

Confidence Interval (Women's Heights)

Confidence Interval (Women's Heights)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate a 90% confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percentage of constructed intervals that contain the population mean.

Given:

Heights of 100 Women (in Inches)									
59.4	71.6	69.3	65.0	62.9	66.5	61.7	55.2		
67.5	67.2	63.8	62.9	63.0	63.9	68.7	65.5		
61.9	69.6	58.7	63.4	61.8	60.6	69.8	60.0		
64.9	66.1	66.8	60.6	65.6	63.8	61.3	59.2		
64.1	59.3	64.9	62.4	63.5	60.9	63.3	66.3		
61.5	64.3	62.9	60.6	63.8	58.8	64.9	65.7		
62.5	70.9	62.9	63.1	62.2	58.7	64.7	66.0		
60.5	64.7	65.4	60.2	65.0	64.1	61.1	65.3		
64.6	59.2	61.4	62.0	63.5	61.4	65.5	62.3		
65.5	64.7	58.8	66.1	64.9	66.9	57.9	69.8		
58.5	63.4	69.2	65.9	62.2	60.0	58.1	62.5		
62.4	59.1	66.4	61.2	60.4	58.7	66.7	67.5		
63.2	56.6	67.7	62.5						

1. [\(Figure\)](#) lists the heights of 100 women. Use a random number generator to select ten data values randomly.
2. Calculate the sample mean and the sample standard deviation. Assume that the population standard deviation is known to be 3.3 inches. With these values, construct a 90% confidence interval for your sample of ten values. Write the confidence interval you obtained in the first space of [\(Figure\)](#).
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into [\(Figure\)](#).

90% Confidence Intervals

-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----

Discussion Questions

1. The actual population mean for the 100 heights given (Figure) is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percent of confidence intervals that contains the mean μ . Write this percent here: _____.
3. Is the percent of confidence intervals that contain the population mean μ close to 90%?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percent of confidence intervals that contained the population mean?
5. When we construct a 90% confidence interval, we say that we are **90% confident that the true population mean lies within the confidence interval**. Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated based on that data. How many of the 100 data values lie within your confidence interval? What percent is this? Is this percent close to 90%?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of ten women and calculated a confidence interval from this information. Without knowing the population mean μ , would you have any way of knowing **for certain** if your interval actually contained the value of μ ? Explain.

Continuous Distribution

Continuous Distribution

Class Time:

Names:

Student Learning Outcomes

- The student will compare and contrast empirical data from a random number generator with the uniform distribution.

Collect the Data Use a random number generator to generate 50 values between zero and one (inclusive). List them in [\(Figure\)](#). Round the numbers to four decimal places or set the calculator MODE to four places.

1. Complete the table.

-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----

2. Calculate the following:

- a. \bar{x} = -----
- b. s = -----
- c. first quartile = -----
- d. third quartile = -----
- e. median = -----

Organize the Data

1. Construct a histogram of the empirical data. Make eight bars.



2. Construct a histogram of the empirical data. Make five bars.



Describe the Data

1. In two to three complete sentences, describe the shape of each graph. (Keep it simple. Does the graph go straight across, does it have a V shape, does it have a hump in the middle or at either end, and so on. One way to help you determine a shape is to draw a smooth curve roughly through the top of the bars.)
2. Describe how changing the number of bars might change the shape.

Theoretical Distribution

1. In words, $X =$ _____.
2. The theoretical distribution of X is $X \sim U(0,1)$.
3. In theory, based upon the distribution $X \sim U(0,1)$, complete the following.
 - a. $\mu =$ _____
 - b. $\sigma =$ _____
 - c. first quartile = _____
 - d. third quartile = _____
 - e. median = _____
4. Are the empirical values (the data) in the section titled [Collect the Data](#) close to the corresponding theoretical values? Why or why not?

Plot the Data

1. Construct a box plot of the data. Be sure to use a ruler to scale accurately and draw straight edges.
2. Do you notice any potential outliers? If so, which values are they? Either way, justify your answer numerically. (Recall that any DATA that are less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ are potential outliers. IQR means interquartile range.)

Compare the Data

1. For each of the following parts, use a complete sentence to comment on how the value obtained from the data compares to the theoretical value you expected from the distribution in the section titled [Theoretical Distribution](#).
 - a. minimum value: _____
 - b. first quartile: _____
 - c. median: _____
 - d. third quartile: _____
 - e. maximum value: _____
 - f. width of IQR: _____
 - g. overall shape: _____
2. Based on your comments in the section titled [Collect the Data](#), how does the box plot fit or not fit what you would expect of the distribution in the section titled [Theoretical Distribution](#)?

Discussion Question

1. Suppose that the number of values generated was 500, not 50. How would that affect what you would expect the empirical data to be and the shape of its graph to look like?

Data Collection Experiment

Data Collection Experiment

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

Movie SurveyAsk five classmates from a different class how many movies they saw at the theater last month. Do not include rented movies.

1. Record the data.
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four names on the class list. Mark that person's name. Continue doing this until you have marked 12 names. You may need to go back to the start of the list. For each marked name record the five data values. You now have a total of 60 data values.
3. For each name marked, record the data.

Order the DataComplete the two relative frequency tables below using your class data.

Frequency of Number of Movies Viewed			
Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1			
2-3			
4-5			
6-7+			

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

Discussion Questions

1. Is one of the tables “more correct” than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

Sampling Experiment

Sampling Experiment

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

Note

The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

Restaurants Stratified by City and Entree Cost

10 and 15, *the fourth column between* 15 and 20, *and the fifth column restaurants with entrees over* 20."> Restaurants Used in Sample

Entree Cost	Under \$10	\$10 to under \$15	\$15 to under \$20	Over \$20
San Jose	El Abuelo	Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's
Blake's	Eulipia, Hayes Mansion, Germania	Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's
Scott's	Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's	Los Gatos	Mary's
Patio	Mount Everest, Sweet Pea's, Andele	Taqueria Lindsey's, Willow Street	Toll House	Charter House, La Maison Du
Cafe	Mountain View	Maharaja, New Ma's, Thai-Rific,	Garden Fresh	Amber Indian, La Fiesta, Fiesta del
Mar	Dawit Austin's, Shiva's, Mazeh	Le Petit Bistro	Cupertino	Hobees, Hung Fu, Samrat, Panda
Express	Santa Barb. Grill, Mand. Gourmet, Bombay	Oven, Kathmandu	West Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana,	Lemon Grass	Pacific Fresh, Charley Brown's, Cafe	Cameroon, Faz, Aruba's
Lion & Compass, The Palace, Beau	Sejour	Santa Clara	Rangoli, Armadillo	Willy's, Thai Pepper, Pasand
Arthur's, Katie's	Cafe, Pedro's, La Galleria	Birk's, Truya	Sushi, Valley Plaza	Lakeside, Mariani's

A Simple Random Sample Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Systematic SamplePick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Stratified SamplePick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

A Stratified SamplePick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

A Cluster Sample Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4. _____	9. _____	14. _____	19. _____	24. _____
5. _____	10. _____	15. _____	20. _____	25. _____

Hypothesis Testing of a Single Mean and Single Proportion

Hypothesis Testing of a Single Mean and Single Proportion

Class Time:

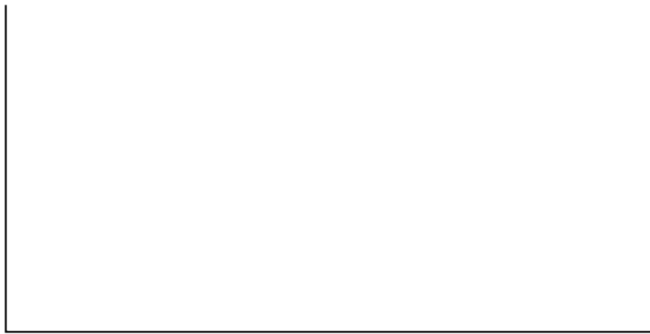
Names:

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0 : _____
2. H_a : _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
 - a. Graph:



- b. Determine the p -value.
7. Do you or do you not reject the null hypothesis? Why?
 8. Write a clear conclusion using a complete sentence.

Language Survey About 42.3% of Californians and 19.6% of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school who speak a language other than English at home is different from 42.3%.

1. H_0 : _____
2. H_a : _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:



b. Determine the p -value.

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey Suppose that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0 : _____
2. H_a : _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:



- b. Determine the p -value.
- 7. Do you or do you not reject the null hypothesis? Why?
- 8. Write a clear conclusion using a complete sentence.

Probability Topics

Probability Topics

Class time:

Names:

Student Learning Outcomes

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment Count out 40 mixed-color M&Ms® which is approximately one small bag's worth. Record the number of each color in (Figure). Use the information from this table to complete (Figure). Next, put the M&Ms in a cup. The experiment is to pick two M&Ms, one at a time. Do **not** look at them as you pick them. The first time through, replace the first M&M before picking the second one. Record the results in the "With Replacement" column of (Figure). Do this 24 times. The second time through, after picking the first M&M, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the "Without Replacement" column section of (Figure). After you record the pick, put **both** M&Ms back. Do this a total of 24 times, also. Use the data from (Figure) to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Population	
Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Theoretical Probabilities

With Replacement	Without Replacement
$P(2 \text{ reds})$	
$P(R_1B_2 \text{ OR } B_1R_2)$	
$P(R_1 \text{ AND } G_2)$	
$P(G_2 R_1)$	
$P(\text{no yellows})$	
$P(\text{doubles})$	
$P(\text{no doubles})$	

Note

G_2 = green on second pick; R_1 = red on first pick; B_1 = brown on first pick; B_2 = brown on second pick; doubles = both picks are the same colour.

Empirical Results

With Replacement	Without Replacement
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)

Empirical Probabilities

With Replacement	Without Replacement
$P(2 \text{ reds})$	
$P(R_1B_2 \text{ OR } B_1R_2)$	
$P(R_1 \text{ AND } G_2)$	
$P(G_2 R_1)$	
$P(\text{no yellows})$	
$P(\text{doubles})$	
$P(\text{no doubles})$	

Discussion Questions

1. Why are the “With Replacement” and “Without Replacement” probabilities different?
2. Convert $P(\text{no yellows})$ to decimal format for both Theoretical “With Replacement” and for Empirical “With Replacement”. Round to four decimal places.
 - a. Theoretical “With Replacement”: $P(\text{no yellows}) = \underline{\hspace{2cm}}$
 - b. Empirical “With Replacement”: $P(\text{no yellows}) = \underline{\hspace{2cm}}$
 - c. Are the decimal values “close”? Did you expect them to be closer together or farther apart? Why?
3. If you increased the number of times you picked two M&Ms to 240 times, why would empirical probability values change?
4. Would this change (see part 3) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
5. Explain the differences in what $P(G_1 \text{ AND } R_2)$ and $P(R_1|G_2)$ represent. Hint: Think about the sample space for each probability.

Discrete Distribution (Playing Card Experiment)

Discrete Distribution (Playing Card Experiment)

Class Time:

Names:

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if an everyday experiment fits a discrete distribution.
- The student will compare technology-generated simulation and a theoretical distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- One full deck of playing cards
- One programming calculator

ProcedureThe experimental procedure for empirical data is to pick one card from a deck of shuffled cards.

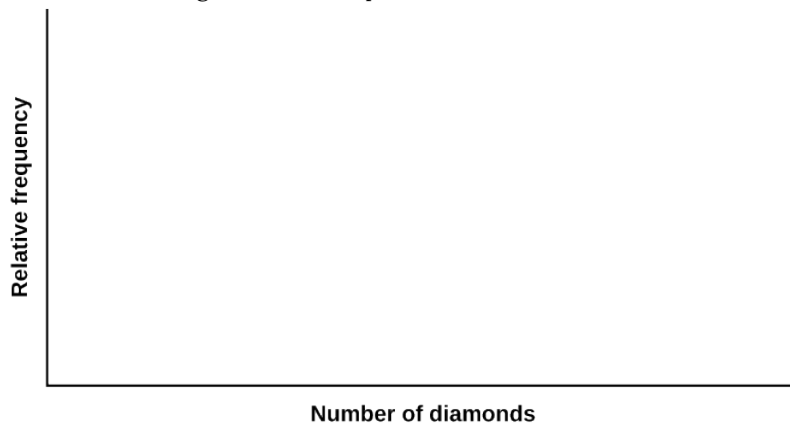
1. The theoretical probability of picking a diamond from a deck is _____.
2. Shuffle a deck of cards.
3. Pick one card from it.
4. Record whether it was a diamond or not a diamond.
5. Put the card back and reshuffle.
6. Do this a total of ten times.
7. Record the number of diamonds picked.
8. Let X = number of diamonds. Theoretically, $X \sim B(\text{_____, _____})$

Organize the Data

1. Record the number of diamonds picked for your class with playing cards in [\(Figure\)](#). Then calculate the relative frequency.

x	Frequency	Relative Frequency
0	-----	-----
1	-----	-----
2	-----	-----
3	-----	-----
4	-----	-----
5	-----	-----
6	-----	-----
7	-----	-----
8	-----	-----
9	-----	-----
10	-----	-----

2. Calculate the following:
 - a. \overline{x} = -----
 - b. s = -----
3. Construct a histogram of the empirical data.

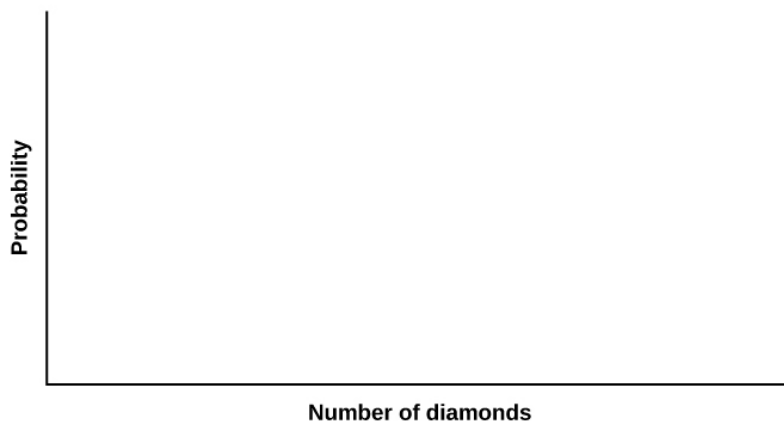


Theoretical Distribution

- a. Build the theoretical PDF chart based on the distribution in the [Procedure](#) section.

x	$P(x)$
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

- b. Calculate the following:
- $\mu =$ _____
 - $\sigma =$ _____
- c. Construct a histogram of the theoretical distribution.



Using the Data

NOTE

RF = relative frequency

Use the table from the [Theoretical Distribution](#) section to calculate the following answers. Round your answers to four decimal places.

- $P(x = 3) =$ _____
- $P(1 < x < 4) =$ _____
- $P(x \geq 8) =$ _____

Use the data from the [Organize the Data](#) section to calculate the following answers. Round your answers to four decimal places.

- $RF(x = 3) =$ _____
- $RF(1 < x < 4) =$ _____
- $RF(x \geq 8) =$ _____

Discussion Questions For questions 1 and 2, think about the shapes of the two graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical, empirical, and simulation distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical, empirical, and simulation distributions.
3. Using your answers from questions 1 and 2, does it appear that the two sets of data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect [\(Figure\)](#) or [\(Figure\)](#) to change, and how would it change? Why? Why wouldn't the other table(s) change?

Discrete Distribution (Lucky Dice Experiment)

Discrete Distribution (Lucky Dice Experiment)

Class Time:

Names:

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if a Tet gambling game fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- one “Lucky Dice” game or three regular dice

Procedure

Round answers to relative frequency and probability problems to four decimal places.

1. The experimental procedure is to bet on one object. Then, roll three Lucky Dice and count the number of matches. The number of matches will decide your profit.
2. What is the theoretical probability of one die matching the object?
3. Choose one object to place a bet on. Roll the three Lucky Dice. Count the number of matches.
4. Let X = number of matches. Theoretically, $X \sim B(\text{-----}, \text{-----})$
5. Let Y = profit per game.

Organize the DataIn (Figure), fill in the y value that corresponds to each x value. Next, record the number of matches picked for your class. Then, calculate the relative frequency.

1. Complete the table.

x	y	Frequency	Relative Frequency
0			
1			
2			
3			

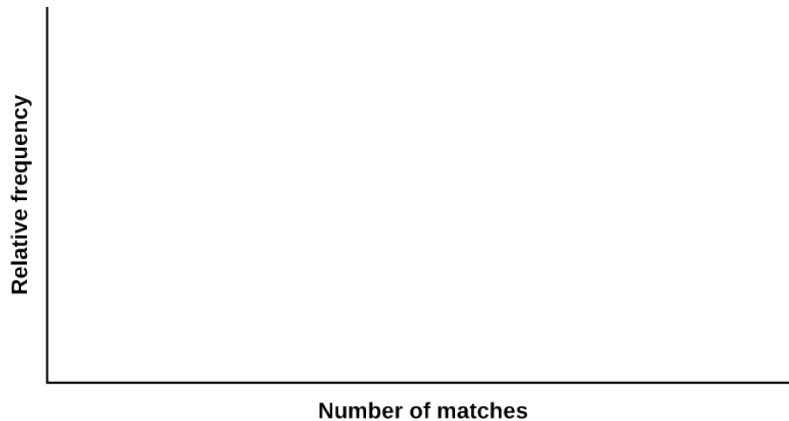
2. Calculate the following:

a. $\bar{x} = \text{-----}$

b. $s_x = \text{-----}$

c. $\bar{y} = \text{-----}$

- d. $s_y = \text{-----}$
3. Explain what \bar{x} represents.
 4. Explain what \bar{y} represents.
 5. Based upon the experiment:
 - a. What was the average profit per game?
 - b. Did this represent an average win or loss per game?
 - c. How do you know? Answer in complete sentences.
 6. Construct a histogram of the empirical data.

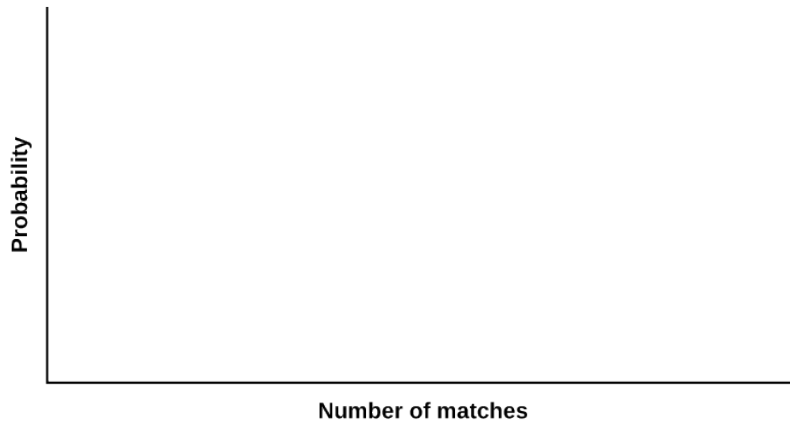


Theoretical Distribution Build the theoretical PDF chart for x and y based on the distribution from the [Procedure](#) section.

1.

x	y	$P(x) = P(y)$
0		
1		
2		
3		

2. Calculate the following:
 - a. $\mu_x = \text{-----}$
 - b. $\sigma_x = \text{-----}$
 - c. $\mu_y = \text{-----}$
3. Explain what μ_x represents.
4. Explain what μ_y represents.
5. Based upon theory:
 - a. What was the expected profit per game?
 - b. Did the expected profit represent an average win or loss per game?
 - c. How do you know? Answer in complete sentences.
6. Construct a histogram of the theoretical distribution.



Use the Data

Note

RF = relative frequency

Use the data from the [Theoretical Distribution](#) section to calculate the following answers. Round your answers to four decimal places.

1. $P(x = 3) =$ _____
2. $P(0 < x < 3) =$ _____
3. $P(x \geq 2) =$ _____

Use the data from the [Organize the Data](#) section to calculate the following answers. Round your answers to four decimal places.

1. $RF(x = 3) =$ _____
2. $RF(0 < x < 3) =$ _____
3. $RF(x \geq 2) =$ _____

Discussion Question For questions 1 and 2, consider the graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions.
3. Thinking about your answers to questions 1 and 2, does it appear that the data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect [\(Figure\)](#) or [\(Figure\)](#) to change, and how would it change? Why? Why wouldn't the other table change?

Regression (Distance from School)

Regression (Distance from School)

Class Time:

Names:

Student Learning Outcomes

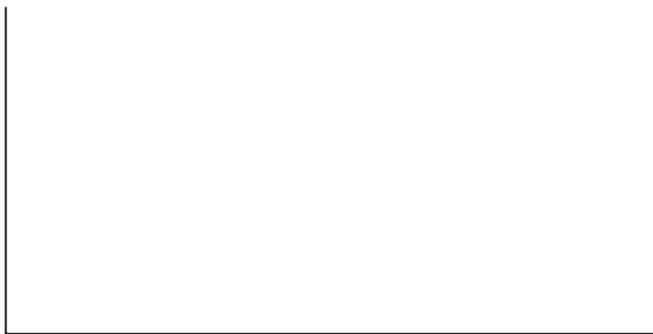
- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the DataUse eight members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from school	Cost of supplies this term
----------------------	----------------------------

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “distance” vs. “cost.” Plot the points on the graph. Label both axes with words. Scale both axes.



Analyze the DataEnter your data into your calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:

a. $a =$ _____

- b. $b =$ _____
 - c. correlation = _____
 - d. $n =$ _____
 - e. equation: $\hat{y} =$ _____
 - f. Is the correlation significant? Why or why not? (Answer in one to three complete sentences.)
2. Supply an answer for the following scenarios:
 - a. For a person who lives eight miles from campus, predict the total cost of supplies this term:
 - b. For a person who lives eighty miles from campus, predict the total cost of supplies this term:
 3. Obtain the graph on your calculator or computer. Sketch the regression line.



Discussion Questions

1. Answer each question in complete sentences.
 - a. Does the line seem to fit the data? Why?
 - b. What does the correlation imply about the relationship between the distance and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

Regression (Textbook Cost)

Regression (Textbook Cost)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data Survey ten textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Number of pages	Cost of textbook
-----------------	------------------

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “pages” vs. “cost.” Plot the points on the graph in [Analyze the Data](#). Label both axes with words. Scale both axes.

Analyze the Data Enter your data into your calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
 - a. $a =$ _____
 - b. $b =$ _____
 - c. correlation = _____
 - d. $n =$ _____
 - e. equation: $y =$ _____
 - f. Is the correlation significant? Why or why not? (Answer in complete sentences.)
2. Supply an answer for the following scenarios:
 - a. For a textbook with 400 pages, predict the cost.
 - b. For a textbook with 600 pages, predict the cost.
3. Obtain the graph on your calculator or computer. Sketch the regression line.



Discussion Questions

1. Answer each question in complete sentences.
 - a. Does the line seem to fit the data? Why?
 - b. What does the correlation imply about the relationship between the number of pages and the cost?
2. Are there any outliers? If so, which point(s) is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

Regression (Fuel Efficiency)

Regression (Fuel Efficiency)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the DataFind a reputable source that provides information on total fuel efficiency (in miles per gallon) and weight (in pounds) of new model cars with automatic transmissions. We will use this data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Using your random number generator, randomly select 20 cars from the list and record their weights and fuel efficiency into [\(Figure\)](#).

Weight Fuel Efficiency

-
2. Which variable should be the dependent variable and which should be the independent variable? Why?
 3. By hand, do a scatterplot of “weight” vs. “fuel efficiency”. Plot the points on graph paper. Label both axes with words. Scale both axes accurately.



Analyze the Data Enter your data into your calculator or computer. Write the linear equation, rounding to 4 decimal places.

1. Calculate the following:
 - a. $a =$ _____
 - b. $b =$ _____
 - c. correlation = _____
 - d. $n =$ _____
 - e. equation: $\hat{y} =$ _____
2. Obtain the graph of the regression line on your calculator. Sketch the regression line on the same axes as your scatter plot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4,000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10,000 pounds using the least squares line? Explain why or why not.
6. Answer each question in complete sentences.
 - a. Does the line seem to fit the data? Why or why not?
 - b. What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?
7. Are there any outliers? If so, which point is an outlier?

Descriptive Statistics

Descriptive Statistics

Class Time:

Names:

Student Learning Outcomes

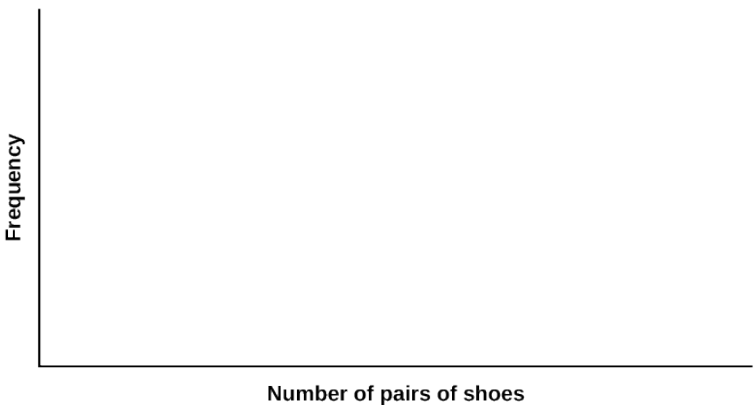
- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data Record the number of pairs of shoes you own.

1. Randomly survey 30 classmates about the number of pairs of shoes they own. Record their values.

Survey Results				
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----
-----	-----	-----	-----	-----

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil and scale the axes.



3. Calculate the following values.
 - a. \bar{x} = -----
 - b. s = -----
4. Are the data discrete or continuous? How do you know?

5. In complete sentences, describe the shape of the histogram.
6. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following values.
 - a. Min = _____
 - b. M = _____
 - c. Max = _____
 - d. Q_1 = _____
 - e. Q_3 = _____
 - f. IQR = _____
2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:
 - a. above the mean.
 - b. below the mean.

Review Exercises (Ch 1-13)

Practice Tests (1-4) and Final Exams

Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

Race Lap Times (in seconds)							
	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Practice Lap Times (in seconds)							
	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

IPO Offer Prices

\$17.00	\$23.00	\$14.00	\$16.00	\$12.00	\$26.00
\$20.00	\$22.00	\$14.00	\$15.00	\$22.00	\$18.00
\$18.00	\$21.00	\$21.00	\$19.00	\$15.00	\$21.00
\$18.00	\$17.00	\$15.00	\$25.00	\$14.00	\$30.00
\$16.00	\$10.00	\$20.00	\$12.00	\$16.00	\$17.44
\$16.00	\$14.00	\$15.00	\$20.00	\$20.00	\$16.00
\$17.00	\$16.00	\$15.00	\$15.00	\$19.00	\$48.00
\$16.00	\$18.00	\$9.00	\$18.00	\$18.00	\$20.00
\$8.00	\$20.00	\$17.00	\$14.00	\$11.00	\$16.00
\$19.00	\$15.00	\$21.00	\$12.00	\$8.00	\$16.00
\$13.00	\$14.00	\$15.00	\$14.00	\$13.41	\$28.00
\$21.00	\$17.00	\$28.00	\$17.00	\$19.00	\$16.00
\$17.00	\$19.00	\$18.00	\$17.00	\$15.00	
\$14.00	\$21.00	\$12.00	\$18.00	\$24.00	
\$15.00	\$23.00	\$14.00	\$16.00	\$12.00	
\$24.00	\$20.00	\$14.00	\$14.00	\$15.00	
\$14.00	\$19.00	\$16.00	\$38.00	\$20.00	
\$24.00	\$16.00	\$8.00	\$18.00	\$17.00	
\$16.00	\$15.00	\$7.00	\$19.00	\$12.00	
\$8.00	\$23.00	\$12.00	\$18.00	\$20.00	
\$21.00	\$34.00	\$16.00	\$26.00	\$14.00	

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

Group and Partner Projects

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

_____ Decide what data you are going to study.

Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

_____ Are your data discrete or continuous? How do you know?

_____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).

_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?

_____ Conduct your survey. **Your data size must be at least 30.**

_____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**

Answer the following (rounded to two decimal places):

a. \bar{x} = _____

b. s = _____

c. First quartile = _____

d. Median = _____

e. 70th percentile = _____

_____ What value is two standard deviations above the mean?

_____ What value is 1.5 standard deviations below the mean?

_____ Construct a histogram displaying your data.

- _____ In complete sentences, describe the shape of your graph.
- _____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
- _____ Construct a box plot displaying your data.
- _____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
- _____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- **Cover sheet:** name, class time, and name of your study
- **Summary page:** This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- **URL** for data, if your data are from the World Wide Web
- **Chart of data, frequency, relative frequency, and cumulative relative frequency**
- **Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

_____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance

telephone call lasts.)

_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?

_____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.

_____ Define (in words) the random variable for your data. $X =$ _____

_____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.

_____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\bar{x} =$ _____

b. $s =$ _____

_____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

_____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.

_____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b .

_____ Normal: $X \sim N$ _____ Use \bar{x} to estimate for μ and s to estimate for σ .

_____ **Must** your data fit one of the above distributions? Explain why or why not.

_____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.

_____ Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. $k =$ _____ (rounded to two decimal places) Note: $k = \bar{x} + (1.75)s$

_____ Determine the relative frequencies (RF) rounded to four decimal places.

Note

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k) =$ _____

b. $RF(X > k) =$ _____

c. $RF(X = k) =$ _____

Note

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

_____ State the distribution: $X \sim$ _____

_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.

_____ Find the following theoretical probabilities (rounded to four decimal places).

- a. $P(X < k) =$ _____
- b. $P(X > k) =$ _____
- c. $P(X = k) =$ _____

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?

_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.

_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.

_____ List the 40 averages in order from smallest to largest.

_____ Define the random variable, \bar{X} , in words. $\bar{X} =$ _____

_____ State the approximate theoretical distribution of \bar{X} . $\bar{X} \sim$ _____

_____ Base this on the mean and standard deviation from your original data.

_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it. Calculate the value \bar{k} (an \bar{X} value) that is 1.75 standard deviations above the sample mean. $\bar{k} =$ _____ (rounded to two decimal places)

Determine the relative frequencies (RF) rounded to four decimal places.

- a. $RF(\bar{X} < \bar{k}) =$ _____
- b. $RF(\bar{X} > \bar{k}) =$ _____
- c. $RF(\bar{X} = \bar{k}) =$ _____

Find the following theoretical probabilities (rounded to four decimal places).

- a. $P(\bar{X} < \bar{k}) =$ _____
- b. $P(\bar{X} > \bar{k}) =$ _____
- c. $P(\bar{X} = \bar{k}) =$ _____

_____ Draw the graph of the theoretical distribution of \bar{X} .

_____ Compare the relative frequencies to the probabilities. Are the values close?

_____ Does it appear that the data of averages fit the distribution of \bar{X} well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.

In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.

_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.

_____ What happened to the shape and distribution when you averaged your data? **In theory**, what should have happened? In theory, would “it” always happen? Why or why not?

_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or \bar{X} distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

_____ **Cover sheet:** name, class time, and name of your study

_____ **Summary pages:** These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions

_____ **URL** for data, if your data are from the World Wide Web

_____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered

_____ **Pages of the data requested**

_____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.

_____ **Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.

_____ **Copy or print out the article** and include a copy in your project, along with the source.

_____ **State how you will collect your data.** (Convenience sampling is not acceptable.)

_____ **Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your

final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.

_____ **State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

_____ **Make two copies of the appropriate solution sheet.**

_____ **Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

_____ **Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.

_____ **Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

- a. Brief discussion of the article, including the source
- b. Statement of the claim made in the article (one of the hypotheses).
- c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
- d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
- e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

_____ **Cover sheet** containing your name(s), class time, and the name of your study

_____ **Summary**, which includes all items listed on summary checklist

_____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.

_____ **Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

_____ **Raw data collected AND a table summarizing the sample data** (n , \bar{x} and s ; or x , n , and p), as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction_____State the bivariate data your group is going to study.

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

_____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.

_____Conduct your survey. Your number of pairs must be at least 30.

_____Print out a copy of your data.

Analysis _____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.

_____State the least squares line and the correlation coefficient.

_____On your scatter plot, in a different color, construct the least squares line.

_____Is the correlation coefficient significant? Explain and show how you determined this.

_____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.

_____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.

_____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula

in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your x -values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____ Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.

_____ Answer the following question, rounded to two decimal places:

- a. Sample mean = _____
- b. Sample standard deviation = _____
- c. First quartile = _____
- d. Third quartile = _____
- e. Median = _____
- f. 70th percentile = _____
- g. Value that is 2 standard deviations above the mean = _____
- h. Value that is 1.5 standard deviations below the mean = _____

_____ Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26, 27-33, 34-40, 41-47, 48-54, 55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .

_____ In complete sentences, describe the shape of your histogram.

_____ Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in [Descriptive Statistics](#) (since you are now using univariate data) to determine which values might be outliers.

_____ Construct a box plot of your data.

_____ Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_____ Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____
 - _____ Cover sheet: names, class time, and name of your study
 - _____ Part I: label the sections “Intro” and “Analysis.”
 - _____ Part II:
 - _____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
 - _____ All graphs requested in the project
 - _____ All calculations requested to support questions in data
 - _____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

Note

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

Solution Sheets

Hypothesis Testing with One Sample

Class Time: _____

Name: _____

a. H_0 : _____

b. H_a : _____

c. In words, **CLEARLY** state what your random variable \bar{X} or P' represents.

d. State the distribution to use for the test.

e. What is the test statistic?

f. What is the p -value? In one or two complete sentences, explain what the p -value means for this problem.

g. Use the previous information to sketch a picture of this situation. **CLEARLY**, label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

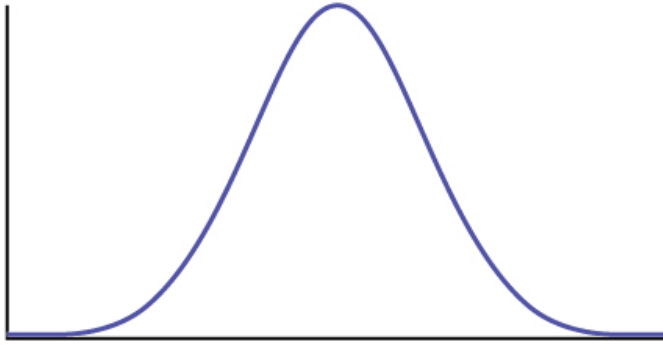
i. Alpha: _____

ii. Decision: _____

iii. Reason for decision: _____

iv. Conclusion: _____

i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.



Hypothesis Testing with Two Samples

Class Time: _____

Name: _____

- H_0 : _____
- H_a : _____
- In words, **clearly** state what your random variable $\bar{X}_1 - \bar{X}_2$, $P'_1 - P'_2$ or \bar{X}_d represents.
- State the distribution to use for the test.
- What is the test statistic?
- What is the p -value? In one to two complete sentences, explain what the p -value means for this problem.
- Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
 - Alpha: _____
 - Decision: _____
 - Reason for decision: _____

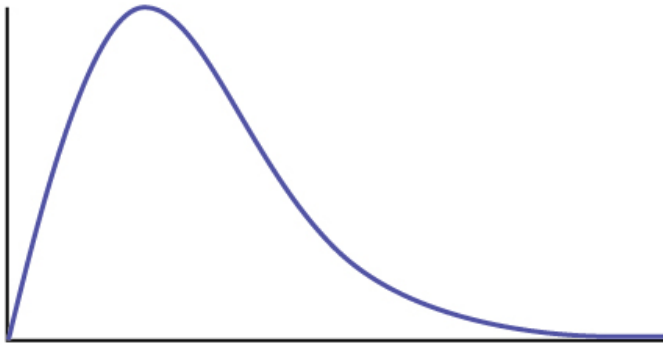
- d. Conclusion: _____
- i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: _____

Name: _____

- H_0 : _____
- H_a : _____
- What are the degrees of freedom?
- State the distribution to use for the test.
- What is the test statistic?
- What is the p -value? In one to two complete sentences, explain what the p -value means for this problem.
- Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences**.
 - Alpha: _____
 - Decision: _____
 - Reason for decision: _____
 - Conclusion: _____

F Distribution and One-Way ANOVA

Class Time: _____

Name: _____

- a. H_0 : _____
- b. H_a : _____
- c. $df(n) = \underline{\hspace{2cm}}$ $df(d) = \underline{\hspace{2cm}}$
- d. State the distribution to use for the test.
- e. What is the test statistic?
- f. What is the p -value?
- g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences**.
 - a. Alpha: _____
 - b. Decision: _____
 - c. Reason for decision: _____
 - d. Conclusion: _____

Mathematical Phrases, Symbols, and Formulas

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer X than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Formulas

Formula 1: Factorial

$$n! = n(n-1)(n-2) \dots (1)$$
$$0! = 1$$

Formula 2: Combinations

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Formula 3: Binomial Distribution

$$X \sim B(n, p) \\ P(X = x) = \binom{n}{x} p^x q^{n-x}, \text{ for } x = 0, 1, 2, \dots, n$$

Formula 4: Geometric Distribution

$$X \sim G(p) \\ P(X = x) = q^{x-1}p, \text{ for } x = 1, 2, 3, \dots$$

Formula 5: Hypergeometric Distribution

$$X \sim H(r, b, n) \\ P(X = x) = \frac{\binom{r}{x} \binom{b}{n-x}}{\binom{r+b}{n}}$$

Formula 6: Poisson Distribution

$$X \sim P(\mu) \\ P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula 7: Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}, a < x < b$$

Formula 8: Exponential Distribution

$$X \sim \text{Exp}(m)$$

$$f(x) = me^{-mx} \quad m > 0, x \geq 0$$

Formula 9: Normal Distribution $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

Formula 10: Gamma Function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad z > 0$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(m+1) = m! \text{ for } m, \text{ a nonnegative integer}$$

$$\text{otherwise: } \Gamma(a+1) = a\Gamma(a)$$

Formula 11: Student's *t*-distribution

$$X \sim t_{df}$$

$$f(x) = \frac{\left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$$Z \sim N(0, 1), Y \sim \chi^2_{df}, n = \text{degrees of freedom}$$

Formula 12: Chi-Square Distribution

$$X \stackrel{2}{\sim} \chi^2_{df}$$

$$f(x) = \frac{x^{\frac{n-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}, x > 0, n = \text{positive integer and degrees of freedom}$$

Formula 13: F Distribution

$$X \sim F_{df(n), df(d)}$$

$df(n)$ =degrees of freedom for the numerator
 $df(d)$ =degrees of freedom for the denominator

$$f(x) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})} \left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\frac{u}{2}-1} \left[1 + \left(\frac{u}{v}\right) x\right]^{-0.5(u+v)}$$

$$X = \frac{Y_u}{W_v}, Y, W \text{ are chi-square}$$

Symbols and Their Meanings

Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\quad}$	The square root of	same
Sampling and Data	π	Pi	3.14159... (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	\bar{x}	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	s_{sx}	s	sample standard deviation
Descriptive Statistics	$s^2 s_x^2$	s squared	sample variance
Descriptive Statistics	$\sigma \sigma_{sx}$	sigma	population standard deviation
Descriptive Statistics	$\sigma^2 \sigma_x^2$	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{ \}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A \text{—} B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same

Chapter (1st used)	Symbol	Spoken	Meaning
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	\geq	greater than or equal to	same
Discrete Random Variables	\leq	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	\neq	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	pdf	prob. density function	same
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	Exp	exponential distribution	same
Continuous Random Variables	k	k	critical value
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z -score	same
The Normal Distribution	Z	standard normal dist.	same


Chapter (1st used)	Symbol	Spoken	Meaning
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	\overline{X}	X-bar	the random variable X-bar
The Central Limit Theorem	μ_x	mean of X	the average of X
The Central Limit Theorem	$\mu_{\overline{x}}$	mean of X-bar	the average of X-bar
The Central Limit Theorem	σ_x	standard deviation of X	same
The Central Limit Theorem	$\sigma_{\overline{x}}$	standard deviation of X-bar	same
The Central Limit Theorem	$\sum X$	sum of X	same
The Central Limit Theorem	$\sum x$	sum of x	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same
Confidence Intervals	t	Student's t-distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with $\alpha/2$ area in right tail	same
Confidence Intervals	$p'; p$	p-prime; p-hat	sample proportion of success
Confidence Intervals	$q'; q$	q-prime; q-hat	sample proportion of failure
Hypothesis Testing	H_0	H-naught, H-sub 0	null hypothesis
Hypothesis Testing	H_a	H-a, H-sub a	alternate hypothesis
Hypothesis Testing	H_1	H-1, H-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error

Chapter (1st used)	Symbol	Spoken	Meaning
Hypothesis Testing	$\overline{X_1} - \overline{X_2}$	X1-bar minus X2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	mu-1 minus mu-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	P1-prime minus P2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	p1 minus p2	difference in population proportions
Chi-Square Distribution	χ^2	Ky-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	y equals a plus b-x	equation of a line
Linear Regression and Correlation	\hat{y}	y-hat	estimated value of y
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ϵ	error	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
Linear Regression and Correlation	1.9s	1.9 times s	cut-off value for outliers
F-Distribution and ANOVA	F	F-ratio	F-ratio





Notes for the TI-83, 83+, 84, 84+ Calculators


Quick Tips

Legend

-  represents a button press
- [] represents yellow command or green letter behind a key
- < > represents items on the screen


To adjust the contrast Press , then hold  to increase the contrast or  to decrease the contrast.

To capitalize letters and words Press  to get one capital letter, or press , then  to set all button presses to capital letters. You can return to the top-level button values by pressing  again.


To correct a mistake If you hit a wrong button, just hit  and start again.

To write in scientific notation Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...



- $4.321 \text{ E } 4 = 4.32110^4$
- $4.321 \text{ E } -4 = 4.32110^{-4}$

To transfer programs or equations from one calculator to another: **Both calculators:** Insert your respective end of the link cable and press , then [LINK].

Calculator receiving information:


1. Use the arrows to navigate to and select <RECEIVE>
2. Press .

Calculator sending information:

1. Press appropriate number or letter.
2. Use up and down arrows to access the appropriate item.
3. Press  to select item to transfer.
4. Press right arrow to navigate to and select <TRANSMIT>.
5. Press .

Note

ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable. Both calculators: press , then [QUIT] to exit when done.

Manipulating One-Variable Statistics

Note

These directions are for entering data with the built-in statistical program.

**Sample Data We
are manipulating
one-variable
statistics.**

Data	Frequency
-2	10
-1	3
0	4
1	5
3	8

To begin:

1. Turn on the calculator.





2. Access statistics mode.



3. Select <4:ClrList> to clear data from lists, if desired.

, 

4. Enter list [L1] to be cleared.

, [L1], 

5. Display last instruction.

, [ENTRY]

6. Continue clearing remaining lists in the same fashion, if desired.



7. Access statistics mode.

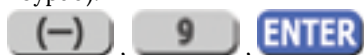


8. Select <1:Edit . . .>



9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).

- Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).



- Continue in the same manner until all data values are entered.

10. In [L2], enter the frequencies for each data value in [L1].

- Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").



- Continue in the same manner until all data values are entered.

11. Access statistics mode.

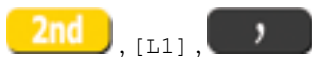


12. Navigate to <CALC>.

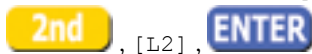
13. Access <1:1-var Stats>.



14. Indicate that the data is in [L1]...



15. ...and indicate that the frequencies are in [L2].



16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.


Drawing Histograms

Note

We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

 , [STAT PLOT]

2. Select <1:plot 1> to access plotting – first graph.



3. Use the arrows navigate go to <ON> to turn on Plot 1.



<ON> , 

4. Use the arrows to go to the histogram picture and select the histogram.





5. Use the arrows to navigate to <Xlist>.

6. If “L1” is not selected, select it.


 , [L1] , 

7. Use the arrows to navigate to <Freq>.

8. Assign the frequencies to [L2].

 , [L2] , 

9. Go back to access other graphs.

 , [STAT PLOT]

10. Use the arrows to turn off the remaining plots.

11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.



2. Select each equal sign (=).

3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.



2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.



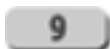
3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.



2. Select <9:ZoomStat>.



3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.



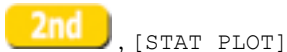
2.
 - $X_{\min} = -2.5$
 - $X_{\max} = 3.5$
 - $X_{scl} = 1$ (width of bars)
 - $Y_{\min} = 0$
 - $Y_{\max} = 10$
 - $Y_{scl} = 1$ (spacing of tick marks on y-axis)
 - $X_{res} = 1$

3. Access graphing mode to see the histogram.



To draw box plots:

1. Access graphing mode.



, [STAT PLOT]

2. Select <1:Plot 1> to access the first graph.



3. Use the arrows to select <ON> and turn on Plot 1.




4. Use the arrows to select the box plot picture and enable it.



5. Use the arrows to navigate to <Xlist>.

6. If “L1” is not selected, select it.


 , [L1] , 

7. Use the arrows to navigate to <Freq>.

8. Indicate that the frequencies are in [L2].


 , [L2] , 

9. Go back to access other graphs.

 , [STAT PLOT]

10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.

11. View the box plot.

 , [STAT PLOT]

Linear Regression

Sample Data

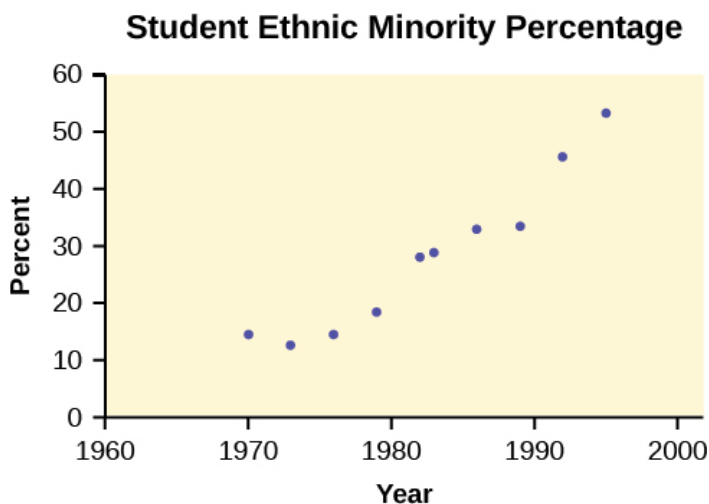
The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

The independent variable is “Year,” while
the independent variable is “Student Ethnic
Minority Percent.”

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

Student Ethnic Minority Percentage

By hand, verify the scatterplot above.



Note

The TI-83 has a built-in linear regression feature, which allows the data to be edited. The x-values will be in [L1]; the y-values in [L2].


To enter data and do linear regression:

1. ON Turns calculator on.



2. Before accessing this program, be sure to turn off all plots.

- Access graphing mode.


 , [STAT PLOT]

- Turn off all plots.

 , 

3. Round to three decimal places. To do so:

- Access the mode menu.

 , [STAT PLOT]

- Navigate to <Float> and then to the right to <3>.

- All numbers will be rounded to three decimal places until changed.




4. Enter statistics mode and clear lists [L1] and [L2], as describe previously.

 , 

5. Enter editing mode to insert values for x and y.

 , 

6. Enter each value. Press  to continue.

To display the correlation coefficient:

1. Access the catalog.

 , [CATALOG]

2. Arrow down and select <DiagnosticOn>

 ... ,  , 

3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

5. Select the form of $y = a + bx$.

 , 

The display will show:

LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r = 2\ 0.924$
- $r = 0.961$


This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- Percent = $-3176.909 + 1.617$ (year #)

The correlation coefficient $r = 0.961$

To see the scatter plot:

1. Access graphing mode.

 , [STAT PLOT]

2. Select <1:plot 1> To access plotting – first graph.



3. Navigate and select <ON> to turn on Plot 1.

<ON> 

4. Navigate to the first picture.

5. Select the scatter plot.



6. Navigate to <Xlist>.



7. If [L1] is not selected, press  , [L1] to select it.

8. Confirm that the data values are in [L1].


<ON> 

9. Navigate to <Ylist>.

10. Select that the frequencies are in [L2].

 , [L2] , 


11. Go back to access other graphs.

 , [STAT PLOT]

12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.



- $X_{\min} = 1970$
- $X_{\max} = 2000$
- $X_{scl} = 10$ (spacing of tick marks on x-axis)
- $Y_{\min} = -0.05$
- $Y_{\max} = 60$
- $Y_{scl} = 10$ (spacing of tick marks on y-axis)
- $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press the graph button to see the scatter plot. 

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.



2. Access the vars menu and navigate to <5: Statistics>.

 , 

3. Navigate to <EQ>.

4. <1: RegEQ> contains the regression equation which will be entered in Y1.



5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.



To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

 , [LIST], <RESID>

2. Confirm twice to view the list of residuals. Use the arrows to select them.

 , 

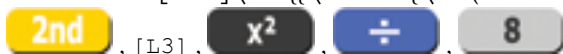
3. The critical point for an outlier is: $1.9V \frac{SSE}{n-2}$ where:

- n = number of pairs of data
- SSE = sum of the squared errors
- $\sum \text{residual}^2$

4. Store the residuals in [L3].



5. Calculate the $\frac{\sum (\text{residual})^2}{n-2}$. Note that $n - 2 = 8$



6. Store this value in [L4].



7. Calculate the critical value using the equation above.



8. Verify that the calculator displays: 7.642669563. This is the critical value.

9. Compare the absolute value of each residual value in [L3] to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of y for various x -values: There are various ways to determine estimates for “ y .” One way is to substitute values for “ x ” in the equation. Another way is to use the **TRACE** on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access DISTR (for “Distributions”).

For technical assistance, visit the Texas Instruments website at <http://www.ti.com> and enter your calculator model into the “search” box.

Binomial Distribution

- `binompdf(n, p, x)` corresponds to $P(X = x)$
- `binomcdf(n, p, x)` corresponds to $P(X \leq x)$
- To see a list of all probabilities for x : 0, 1, \dots , n , leave off the “ x ” parameter.

Poisson Distribution

- `poissonpdf(λ , x)` corresponds to $P(X = x)$
- `poissoncdf(λ , x)` corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value `-1EE99` for left bound
- ∞ uses the value `1EE99` for right bound

Normal Distribution

- `normalpdf(x , μ , σ)` yields a probability density function value (only useful to plot the normal curve, in which case “ x ” is the variable)
- `normalcdf(left bound, right bound, μ , σ)` corresponds to $P(\text{left bound} < X < \text{right bound})$
- `normalcdf(left bound, right bound)` corresponds to $P(\text{left bound} < Z < \text{right bound})$ – standard normal
- `invNorm(p , μ , σ)` yields the critical value, k : $P(X < k) = p$
- `invNorm(p)` yields the critical value, k : $P(Z < k) = p$ for the standard normal

Student's t-Distribution

- `tpdf(x , df)` yields the probability density function value (only useful to plot the student-t curve, in which case “ x ” is the variable)
- `tcdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < t < \text{right bound})$

Chi-square Distribution

- `χ^2 pdf(x , df)` yields the probability density function value (only useful to plot the χ^2 curve, in which case “ x ” is the variable)
- `χ^2 cdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < \chi^2 < \text{right bound})$

F Distribution

- `Fpdf(x , $dfnum$, $dfdenom$)` yields the probability density function value (only useful to plot the F curve, in which case “ x ” is the variable)
- `Fcdf(left bound, right bound, $dfnum$, $dfdenom$)` corresponds to $P(\text{left bound} < F < \text{right bound})$

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press **DATA** to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing **STAT** once in the appropriate tests.

Confidence Intervals

- **ZInterval** is the confidence interval for mean when σ is known.
- **TInterval** is the confidence interval for mean when σ is unknown; s estimates σ .
- **1-PropZInt** is the confidence interval for proportion.

Note

The confidence levels should be given as percents (ex. enter “95” or “. 95” for a 95% confidence level).

Hypothesis Tests

- **Z-Test** is the hypothesis test for single mean when σ is known.
- **T-Test** is the hypothesis test for single mean when σ is unknown; s estimates σ .
- **2-SampZTest** is the hypothesis test for two independent means when both σ 's are known.
- **2-SampTTest** is the hypothesis test for two independent means when both σ 's are unknown.
- **1-PropZTest** is the hypothesis test for single proportion.
- **2-PropZTest** is the hypothesis test for two proportions.
- **χ^2 -Test** is the hypothesis test for independence.
- **χ^2 GOF-Test** is the hypothesis test for goodness-of-fit (TI-84+ only).
- **LinRegTTEST** is the hypothesis test for Linear Regression (TI-84+ only).

Note

Input the null hypothesis value in the row below “Inpt.” For a test of a single mean, “ $\mu\emptyset$ ” represents the null hypothesis. For a test of a single proportion, “ $p\emptyset$ ” represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

From 2.3: Displaying Quantitative Data

Here are calculator instructions for entering data and for creating a customized histogram. Create a histogram.

Calculator Instructions

- Press **Y=**. Press **CLEAR** to delete any equations.
- Press **STAT 1:EDIT**. If L1 has data in it, arrow up into the name L1, press **CLEAR** and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press **WINDOW**. Set $X_{\min} = .5$, $X_{\max} = 6.5$, $X_{\text{scl}} = (6.5 - .5)/6$, $Y_{\min} = -1$, $Y_{\max} = 20$, $Y_{\text{scl}} = 1$, $X_{\text{res}} = 1$.

- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram. Frequency Polygons Frequency polygons are analogous to line graphs, but instead utilize binning techniques to make continuous data visually easy to interpret. It is essentially a combination of a histogram and line graph.

Tables

This module contains links to government site tables used in statistics.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, January 3, 2009)

- [Student t table](#)
- [Normal table](#)
- [Chi-Square table](#)
- [F-table](#)
- All [four tables](#) can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- [95% Critical Values of the Sample Correlation Coefficient](#)

Glossary

Alternative hypothesis

A working hypothesis that is contradictory to the null hypothesis

Anecdotal evidence

Evidence that is based on personal testimony and collected informally

Association

A relationship between variables

Bernoulli trial

An experiment with the following characteristics:

- There are only two possible outcomes called “success” and “failure” for each trial
- The probability (p) of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial)

Bimodal distribution

A distribution that has 2 modes

Binomial distribution

A random variable that counts the number of successes in a fixed number (n) of independent Bernoulli trials each with probability of a success (p)

Bivariate data

Data consisting of two variables, often in search of an association

Blinding

Not telling participants which treatment they are receiving

Block design study

Grouping individuals based on a variable into "blocks" and then randomizing cases within each block to the treatment groups

Case-control study

A study that compares a group that has a certain characteristic to a group that does not, often a retrospective study for rare conditions

Center

The central tendency or most typical value of a dataset

Central limit theorem (CLT)

States that if there is a population with mean μ and standard deviation σ and you take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed

Class midpoint

Found by adding the lower limit and upper limit, then dividing by 2

Class width

The difference in consecutive lower class limits

Cluster sampling

A method of sampling where the population has already sorted itself into groups (clusters), randomly selecting a cluster, and using every individual in the chosen cluster as the sample

Coefficient of determination

A numerical measure of the percentage or proportion of variation in the dependent variable (y) that can be explained by the independent variable (x)

Cohort study

Longitudinal study where a group of people (typically having a common factor) are studied and data is collected for a purpose

Complement

The complement of an event consists of all outcomes in a sample space that are NOT in the event

Completely randomized study

Dividing participants into treatment groups randomly

Conditional probability

The likelihood that an event will occur given knowledge of another event

Confidence interval

An interval built around a point estimate for an unknown population parameter

Confounding (lurking, conditional) variable

A variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Contingency (two-way) table

A table in a matrix format that displays the frequency distribution of different variables

Continuity correction

When statisticians add or subtract .5 to values to improve approximation

Continuous random variable

A random variable (RV) whose outcomes are measured as an uncountable, infinite, number of values

Control group

A group in a randomized experiment that receives no (or an inactive) treatment but is otherwise managed exactly as the other groups

Controlled (designed) experiment

Type of experiment where variables are manipulated; data is collected in a controlled setting

Convenience sampling

Selecting individuals that are easily accessible and may result in biased data

Correlation coefficient

A numerical measure that provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y

Critical value

Point that lies on a distribution that acts as a cut-off value for accepting or rejecting the null hypothesis

Cross-sectional study

Data collection on a population at one point in time (often prospective)

Cumulative distribution function (CDF)

A function that gives the probability that a random variable takes a value less than or equal to x

Cumulative relative frequency

The sum of the relative frequencies for all values that are less than or equal to the given value

Data

Actual values (numbers or words) that are collected from the variables of interest

Data analysis process

Process of collecting, organizing, and analyzing data

Degrees of freedom

The number of objects in a sample that are free to vary

Descriptive statistics

Methods of organizing, summarizing, and presenting data

Designed (controlled) experiment

Data collection where variables are manipulated in a controlled setting

Difference in means

The difference in the means of two independent populations

Discrete random variable

A random variable that produces discrete data

Distribution

The possible values a variable can take on, and how often it does so

Double-blind study

The act of blinding both the subjects of an experiment and the researchers who work with the subjects

Empirical rule

Roughly 68% of values are within 1 standard deviation of the mean, roughly 95% of values are within 2 standard deviations of the mean, and 99.7% of values are within 3 standard deviations of the mean

Event

A single outcome, or subset of outcomes, of an experiment that you are interested in

Expected value

Mean of a random variable

Experimental unit

Any individual or object to be measured

Explanatory variable

The independent variable in an experiment; the value controlled by researchers

Extrapolation

The process of predicting outside of the observed x values

Factors

Variables in an experiment

Frequency

The number of times a value of the data occurs

Graphical descriptive methods

Organizing, summarizing, or presenting data visually in graphs, figures, or charts

Hypothesis testing

A decision making procedure for determining whether sample evidence supports a hypothesis

Independent

The occurrence of one event has no effect on the probability of the occurrence of another event

Individuals

The person, animal, item, thing, place, etc. that we collect information about

Inferential statistics

The facet of statistics dealing with using a sample to generalize (or infer) about the population

Influential points

Observed data points that do not follow the trend of the rest of the data and have a large influence on the calculation of the regression line

Intersection (AND)

The shared or common outcomes of two events

Interval scale level

Quantitative data where the difference or gap between values is meaningful

Law of large numbers

As the number of trials in a probability experiment increases, the relative frequency of an event approaches the theoretical probability

Levels

Certain values of variables in an experiment

Linear regression

A mathematical model of a linear association

Longitudinal study

Collecting data multiple times on the same individuals, usually at fixed increments, over a period of time

Lower class limit

The lower end of a bin or class in a frequency table or histogram

Margin of error (MoE)

How much a point estimate can be expected to differ from the true population value; made up of the standard error multiplied by the critical value

Matched pairs design

Very similar individuals (or even the same individual) receive two different two treatments (or treatment vs. control) then the difference in results are compared

Mean (average)

A number that measures the central tendency of the data

Measures of location

A measure of an observation's standing relative to the rest of the dataset

Median

The middle number in a sorted list

Modality

How many peaks or clusters there appear to be in a quantitative distribution

Mode

The most frequently occurring value

Mutually exclusive (disjoint)

Two events that cannot happen at the same; they share no common outcomes

Nominal scale level

Categorical data where the the categories have no natural, intuitive, or obvious order

Normal (Gaussian) distribution

A commonly used symmetric, unimodal, bell-shaped, continuous probability distribution

Null hypothesis

The claim that is assumed to be true and is tested in a hypothesis test

Numerical descriptive methods

Numbers that summarize some aspect of a dataset, often calculated

Observational study

Data collection where no variables are manipulated

Ordinal scale level

Categorical data where the the categories have a natural or intuitive order

Outcome

A particular result of an experiment

Outlier

An observation that stands out from the rest of the data significantly

P-value

The probability that an event will occur, assuming the null hypothesis is true

Parameter

A number that is used to represent a population characteristic and can only be calculated as the result of a census

Placebo

An inactive treatment that has no real effect on the explanatory variable

Point estimate

The value that is calculated from a sample used to estimate an unknown population parameter

Point estimation

Using sample data to calculate a single statistic as an estimate of an unknown population parameter

Pooled proportion

Estimate of the common value of p_1 and p_2

Population

The whole group of individuals who can be studied to answer a research question

Population mean

The arithmetic mean, or average of a population

Population mean difference

The mean of the differences in a matched pairs design

Population proportion

The number of individuals that have a characteristic we are interested in divided by the total number in the population

Power

The probability of failing to reject a true hypothesis

Probability

The study of randomness; a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Probability density function (PDF)

A function that defines a continuous random variable, and the likelihood of an outcome

Probability experiment

A random experiment where the result is not predetermined

Probability mass function (PMF)

A function that gives the probability that a discrete random variable is exactly equal to some value (x)

Probability model

A mathematical representation of a random process that lists all possible outcomes and assigns probabilities to each of them

Prospective study

Collecting information as events unfold

Qualitative (categorical) data

Data that describes qualities, or puts individuals into categories

Quantile

Points in a distribution that relate to the rank order of values in that distribution

Quantitative (numerical) data

Numerical data with a mathematical context

Quantitative continuous data

Data produced by a variable that takes on an uncountable, infinite, number of values

Quantitative discrete data

Data produced by a variable that takes on a countable number of values

Random variable

A representation of a probability model

Ratio scale level

Quantitative data where the difference or gap between values is meaningful AND has a true 0 value

Relative frequency

The percentage, proportion, or ratio of the frequency of a value of the data to the total number of outcomes

Repeated measures

When an individual goes through a single treatment more than once

Residual (error)

A residual measures the vertical distance between an observation and the predicted point on a regression line

Response variable

The dependent variable in an experiment; the value that is measured for change at the end of an experiment

Retrospective study

Collecting or using data after events have taken place

Robust

Not affected by violations of assumptions such as outliers

Sample

A subset of the population studied

Sample mean

The arithmetic mean, or average of a dataset

Sample proportion

The number of individuals that have a characteristic we are interested in divided by the total number in the sample, often found from categorical data

Sample space

The set of all possible outcomes of an experiment

Sampling bias

Bias resulting from all members of the population not being equally likely to be selected

Sampling distribution

The probability distribution of a statistic at a given sample size

Sampling variability

The idea that samples from the same population can yield different results

Shape

What a dataset looks like visually

Significance level

Probability that a true null hypothesis will be rejected, also known as Type I error and denoted by α

Simple random sample (SRS)

Each member of the population is equally likely to be chosen for a sample of a given sample size *and* each sample is equally likely to be chosen

Slope

Tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average

Spread (variation, variability)

The level of variability or dispersion of a dataset; also commonly known as variation/variability

Standard deviation

The average distance (deviation) of each observation from the mean

Standard error

The standard deviation of a sampling distribution

Standard normal distribution (SND)

A normal random variable with a mean of 0 and standard deviation of 1 which z-scores follow; denoted $N(0, 1)$

Statistic

A number calculated from a sample

Statistical inference

Using information from a sample to answer a question, or generalize, about a population

Statistically significant

Finding sufficient evidence that the effect we see is not just due to variability, often from rejecting the null hypothesis

Stratified sampling

Dividing a population into groups (strata), and then using simple random sampling to identify a proportionate number of individuals from each

Systematic (probability) sampling

Using some sort of pattern or probability based method for choosing your sample

T-distribution

A family of t-distributions, dependent on degrees of freedom, similar to the normal distribution but with more variability built in

Test statistic

A measure of how far what you observed is from the hypothesized (or claimed) value

Treatment combinations (interactions)

Combinations of levels of variables in an experiment

Treatments

Different values or components of the explanatory variable applied in an experiment

Tree diagram

Diagram that helps calculate and organize the number of possible outcomes of an event or problem

Type I error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true

Type II error

Erroneously rejecting a true null hypothesis, or erroneously failing to reject a false null hypothesis

Uniform distribution

A probability distribution in which all outcomes are equally likely

Union (OR)

The set of all outcomes in two (or more) events

Upper class limit

The upper end of a bin or class in a frequency table or histogram

Values

Possible observations of the variable

Variable

A characteristic of interest for each person or object in a population

Variance

The square of the standard deviation; a computational step along the way to calculating the standard deviation

Variation (variability, spread)

The level of variability or dispersion of a dataset; also commonly known as 'spread'

Venn diagram

A diagram that shows all possible relations between a collection of different sets

y-intercept

The value of y when x is 0 in your regression equation

z-score

A measure of location that tells us how many standard deviations a value is above or below the mean