# Statistical Thinking

2.1

## A Simulation Approach to Modeling Uncertainty

Catalysts for Change

CATALYSTS FOR CHANGE

# Statistical Thinking

## A Simulation Approach to Modeling Uncertainty

CATALYST PRESS

PUBLISHED BY CATALYST PRESS



Zieffler, A., & Catalysts for Change. (2013). *Statistical Thinking: A simulation approach to uncertainty* (second edition). Minneapolis, MN: Catalyst Press.

The work to create the material appearing in the book was made possible by the National Science Foundation (DUE-0814433). All graphics in the book were in the public domain and were obtained from Wikimedia Commons, Public Domain Images, Easy Vectors, Clker.com, and Open Clip Art Library. The TinkerPlots™ icons and screenshots were used by permission from Key Curriculum Press.

Printed in the United States of America

ISBN 978-0615691305

Catalyst Press
Minneapolis, MN 55455

http://catalystsumn.blogspot.com

*Third printing, January 2015*

# *Contents*

 **Introduction** **17**

Graphical Icons . . . . . 18

TinkerPlots™ Software . . . . . 18

Data, Errata and Other Book Resources . . . . . 19

Mac Users . . . . . 19

Participation in the Learning Process . . . . . 19

 **Unit I: Modeling & Simulation** **21**

Simulation . . . . . 22

Policy and Population . . . . . 23

Outline of the Unit . . . . . 25

 **How Random is the iPod's Shuffle?** **27**

 **iPod Shuffle** **29**

Group Task . . . . . 30

Explore and Describe . . . . . 31
Develop Rules . . . . . 32
Test Rules . . . . . 32
Evaluate . . . . . 33
Summarize . . . . . 33
Discussion . . . . . 34

 **Randomness** **35**

 **Can You "Beat" Randomness—Part I** **36**

 **Modeling Random Behavior—Part I** **39**

Intuitions about Coin Flips . . . . . 40
Modeling Coin Flips . . . . . 42
Setting Up the Model . . . . . 42
Collecting the Results from Many Trials . . . . . 44
Plotting the Results from Many Trials . . . . . 45
Intuitions about Dice . . . . . 48
Modeling Dice Rolls . . . . . 49
Setting Up the Model . . . . . 49
Extensions . . . . . 51

 **Probability Simulation** **53**

 **Modeling Random Behavior—Part II** **54**

Modeling Coin Flips . . . . . 54

Automating the Collection of Trial Results . . . . 55

Plotting the Outcomes from a Trial . . . . . 56

Computing the Trial Result in the Plot . . . . . 56

Collecting the Results from Many Trials . . . . . 57

Modeling Dice Rolls . . . . . 59

Extensions . . . . . 61

 **Pregnancy Tests Readiness** **62**

 **Pregnancy Tests** **63**

Modeling the Accuracy of the First Response Gold® Digital Pregnancy Test . . . . . 65

Modeling the Accuracy for the "Low Pregnancy" Group 66

Simulate Selecting 100 Women at Random . . . . 67

Modeling the Accuracy for the "High Pregnancy" Group . . . . . 68

 **Introduction to Statistical Hypothesis Testing** **69**

Models and Hypotheses . . . . . 70

Connections . . . . . 72

 **Matching Dogs to Owners** **73**

The Random Chance Model . . . . . 76

Matching Dogs and Owners Under the Random Chance Model . . . . . 77

Using TinkerPlots™ to Match Dogs and Owners Under the Random Chance Model . . . . . 78

Random Assignment: Sampling without Replacement 78

Creating Attributes Using Formulas . . . . . 80

Plotting and Collecting Results . . . . . 82

Evaluating the Observed Result . . . . . 83

 **Helper or Hinderer Readiness** **88**

 **Helper or Hinderer** **89**

Selecting the Helper or Hinderer Under the Random Chance Model . . . . . 91

Carrying Out the Simulation . . . . . 92

Evaluating the Observed Result . . . . . 93

Evidence Against the Model . . . . . 93

 **Learning Goals: Unit 1** **97**

Literacy/Understanding (Terms and Concepts) . . . . . 97

Selecting/Using Models . . . . . 98

TinkerPlots™ Skills . . . . . 98

 **Unit 1 Wrap-up & Review** **100**

Terminology for Unit 1 . . . . . 100

Modeling Random Behavior . . . . . 101

Matching Dogs to Owners . . . . . 101

Helper or Hinderer . . . . . 102

Further Practice . . . . . 103

 **Unit II: Comparing Groups** **105**

Comparisons in the Media . . . . . 106

Statistical Comparisons . . . . . 107

Operationalization . . . . . 108

Summarization . . . . . 110

To Infer or Not to Infer . . . . . 113

Outline of the Unit . . . . . 115

 **America's Most Reliable Airlines** **116**

 **Comparing Airlines** **118**

Group Task . . . . . 119

Explore and Describe . . . . . 120

Develop Rules . . . . . 121

Test Rules . . . . . 121

Summarize . . . . . 122

Discussion . . . . . 123

 **Characteristics of Distributions** **124**

Cell Phone Bills . . . . . 127

Number of Hours Studied . . . . . 128

 **Describing Distributions** **130**

Shape . . . . . 131

Location . . . . . 131

Variation . . . . . 132

Putting It All Together . . . . . 133

 **Memorization** **135**

Examining the Observed Data . . . . . 136

Summarizing the Difference Between the Two Conditions 137

Considering Chance Variation as an Explanation for the Difference in Means . . . . . 138

The Null Model . . . . . 138

Re-randomization: Inspecting Other Possible Random Assignments of the Subjects . . . . . 139

Physical Simulation of the Re-Randomization . . . . . 140

Examining the Distribution of the Difference in Means . 141

 **What the p-value?** **144**

**Sleep Deprivation** 145

Modeling the Differences in Improvent Under the Null Model . . . . . 148

Randomization Tests in TinkerPlots™ . . . . . 149

Modeling a Set of Fixed Responses Under the Null Model 149

Modeling the Random Assignment of the Treatment Condition Labels by Linking Multiple Devices . . . . 150

Computing the Difference in Means . . . . . 152

Collecting the Difference in Means . . . . . 152

Evaluating the Observed Result . . . . . 152

Example Write-Up for Sleep Deprivation Study . . . . 155

**Strength Shoe®** 158

Random Assignment . . . . . 160

Randomization Applet . . . . . 165

**Random Assignment** 169

**Dolphin Therapy** 170

Modeling the Improvement Under the Null Model . . 174

Plotting and Collecting the Results . . . . . 175

Evaluating the Observed Result . . . . . 176

**Latino Achievement** 179

Modeling Achievement Under the Null Model . . . . . 181

Plotting and Collecting the Results . . . . . 182

Evaluating the Observed Result . . . . . 182

 **Random Selection** **185**

Studies that Use Random Sampling . . . . . 185

Random Sampling . . . . . 187

 **Murderous Nurse** **189**

Discuss the Following Questions . . . . . 190

Modeling the Chance Variation Under the Assumption of No Difference . . . . . 192

Plotting and Collecting the Results . . . . . 193

Evaluating the Observed Result . . . . . 193

 **Observational Studies** **196**

 **Pregnancy Tests** **198**

Sensitivity and Specificity . . . . . 200

False Positives and False Negatives . . . . . 201

 **Learning Goals: Unit 2** **204**

Literacy/Understanding (Terms and Concepts) . . . . . 204

Selecting/Using Models . . . . . 205

Evaluation . . . . . 205

TinkerPlots™ Skills . . . . . 206

 **Unit 2 Wrap-up & Review** **207**

Terminology for Unit 2 . . . . . 207

Sleep Deprivation . . . . . 210

Dolphin Therapy . . . . . 212

Teen Hearing . . . . . 213

Mammography . . . . . 215

Native Californians . . . . . 217

Blood Pressure . . . . . 218

Social Fibbing . . . . . 220

 **Unit III: Sampling & Estimation** **222**

Outline of the Unit . . . . . 223

 **Sampling** **225**

Unbiasedness . . . . . 227

Simple Random Sampling . . . . . 232

Use TinkerPlots™ to Draw a SRS . . . . . 233

Examining the Sampling Variation . . . . . 235

Sample Size . . . . . 236

Population Size . . . . . 238

 **Comparing Hand Spans** 241

The Standard Deviation . . . . . 243

Using Both the Mean and Standard Deviation: A Complete Summary . . . . . 246

 **Understanding the Standard Deviation** 247

 **Kissing the 'Right' Way** 248

Precision of the Estimate . . . . . 250

Modeling the Variation Due to Random Sampling . . . . 251

Nonparametric Bootstrapping Using TinkerPlots™ . . . 252

Evaluating the Bootstrap Distribution . . . . . 252

Margin of Error . . . . . 253

Interval Estimates . . . . . 254

 **Margin of Error** 256

 **Memorization—Part II** 257

Size of the Effect . . . . . 258

Precision of the Size of the Effect . . . . . 259

Using TinkerPlots™ to Measure the Precision . . . . 261

Interval Estimate for the Size of the Effect . . . . . 262

 **Why Two Standard Errors?** 264

Exploring Random Samples from an Unknown Population . . . . . 264

Exploring a Single Random Sample . . . . . 267

Exploring the Bootstrap Distribution from Different Random Samples . . . . . 269

Exploring the Bootstrap Intervals Computed from Different Random Samples . . . . . 269

What does it Mean to be 95% Confident? . . . . . 271

 **Learning Goals: Unit 3** **274**

Literacy/Understanding (Terms and Concepts) . . . . . 274

Selecting/Using Models . . . . . 275

Evaluation . . . . . 276

TinkerPlots™ Skills . . . . . 276

 **Unit 3 Wrap-up & Review** **277**

Terminology for Unit 3 . . . . . 277

Sampling . . . . . 279

Comparing Airlines—Revisited . . . . . 279

Rating Chain Restaurants . . . . . 281

Emotional Support . . . . . 282

Balsa Wood . . . . . 283

Microsort® . . . . . 284

Marijuana . . . . . 286

Bibliography 287

*The material in this book is a direct reflection of the ideas, work, and effort of Robert delMas, Joan Garfield, Robert Gould, Rebekah R. Isaak, Laura Le, Andrew Zieffler, and Laura Ziegler as a part of the NSF-funded Change Agents for Teaching and Learning Statistics (CATALST; NSF DUE-0814433) project. Some of the activities presented here were originally developed by Beth Chance, George Cobb, John Holcomb, and Allan Rossman as part of their NSF-funded project Concepts of Statistical Inference: A Randomization-Based Curriculum (NSF CCLI-DUE-0633349). The second and third editions of the book also includes contributions from Ethan Brown, Katherine Edwards, Michelle Everson, Elizabeth Fry, Nicola Justice, and Anelise Sabbag.*

# *Introduction*

Learning statistics is sexy. Hal Varian, Google's chief economist, believes this. During an interview in *McKinsey Quarterly*, Varian stated, "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?" Varian is not the only person to express this sentiment either. Hans Rosling in *Joy of Stats* referred to statistics as the "sexiest subject around".

http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286

Watch *Joy of Stats* online at http://www.gapminder.org/videos/the-joy-of-stats/

Whether you believe it is the sexiest subject or not, it is incontrovertible that the use of statistics and data are prevalent in today's information age. Almost every person on earth will benefit from learning some foundational ideas of statistics. Concepts and ideas related to statistical modeling, inference, comparison, sampling, and estimation are as germaine to our daily lives as they are to scientists and researchers. Google, Netflix, Twitter, Facebook, OKCupid, Match.com, Amazon, iTunes, and the Federal Government are just a handful of the companies and organizations that use statistics on a daily basis. Journalism, political science, biology, sociology, psychology, graphic design, economics, sports science, and dance are all disciplines that have made use of statistical methodology.

The materials in this book will introduce you to the seminal ideas underlying the discipline of statistics. In addition, they have been designed with your learning in mind. For example, many of the in-class activities were developed using pedagogical principles, such as small group activities and discussion, that have been shown in research to improve student learning.

## *Graphical Icons*

The use of graphical icons throughout the text are intended to help you make your way through the material. In the table of contents we denote each chapter as an in-class activity, course reading, or an assignment. The graphical icon to denote these are

 In-class Activity

 Course Reading

 Assignment

The course readings and assignments should be completed outside of class and are intended to help you both learn and extend the ideas, skills, and concepts you learn in the classroom. The readings themselves are not all "traditional" readings in the sense of words written on paper, but instead often link to video clips, blogs and other multimedia material.

## *TinkerPlots™ Software*

Much of the material presented in the book requires the use of TinkerPlots™. This software can be downloaded (for Mac or PC) from the McGraw Hill website. As of the time of this writing, a non-expiring, home-use license can be purchased and downloaded for $49.95 from the publisher's website. (Note also that a student one-year license can be purchased for $7.95.)

### *Data, Errata and Other Book Resources*

The data sets used in the materials, as well as the book errata, are available at `https://github.com/zief0002/Statistical-Thinking`. To download these resources, click the button labeled `Download ZIP` on the right-hand side of the webpage.

### *Mac Users*


If you are using a Mac and seem to have problems downloading the TinkerPlots™ data files, hold the option-key while clicking on the link. This should download the file to your desktop (or your *Downloads* folder). You then need to erase the `.txt` suffix that is appended to the end of the downloaded file in order to use the file with TinkerPlots™. The file suffix should be `.tp`, and not `.tp.txt`.

### *Participation in the Learning Process*

The textbook, instructors, and TAs are all in place to help you learn the material. In the end, however, you will have to do all of the hard work associated with actually learning that material. To most succesfully navigate this, it is vital that you be an active participant in the learning process. Coming to class, participating in the activities and discussions, reading, completing the assignments, and asking questions are essential to successful learning.

Learning anything new takes time and effort and this is especially true of learning statistics, as you are not just learning a set of methods, but rather a disciplined way of thinking about the world. Changing your habits of mind will take continual practice. It will also take a great deal of patience and persistence.

As you engage in and use the skills, concepts and ideas introduced in the material, you will find yourself thinking about

data and evidence in a different way. This may lead you to make different decisions or choices. But, even if this course doesn't change your world overnight, you will at the very least be able to critically think about inferences and conclusions drawn from data and the underlying assumptions that are inherent therein.

# *Unit I: Modeling & Simulation*

*There is mounting evidence that the "model-building era" that dominated the theoretical activities of the sciences for a long time is about to be succeeded or at least lastingly supplemented by the "simulation era".*

—S. Hartmann (1996)

Modeling is one of the most important subjects you may ever learn. It is used in microbiology, macroeconomics, urban studies, sociology, psychology, public health, computer science, and of course, statistics. In fact, modeling is a method that is used in almost every discipline. Many think that it is an important skill to learn because it is so pervasive. While this is true, even more importantly is how closely the skills of modeling tie to the more general skills of problem solving. Starfield, Smith, and Bleloch (1994) summed this sentiment up nicely when they wrote, "learning to model is bound up with learning to solve problems and to think imaginatively and purposefully" (p. x).

Starfield, A. M., Smith, K. A., & Bleloch, A. L. (1994). *How to model it: Problem solving for the computer age.* Edina, MN: Burgess International Group, Inc.

A model is a simplified representation of a system that can be used to promote an understanding of a more complex system. For example, meteorologists use computers to build models of the climate to understand and predict the weather. The computer model includes behaviors or properties which correspond, in some way, to the particular real-world system of climate. The computer models, however, do not include every possible detail about climate. All models leave things out and get some things—many things—wrong. This is because

all models are simplifications of reality. Since all models are simplifications of reality there is always a trade-off as to what level of detail is included in the model. If too little detail is included in the model one runs the risk of missing relevant interactions and the resultant model does not promote understanding. If too much detail is included in the model, the model may become overly complicated and actually preclude the development of understanding.

Models. have many purposes, but are primarily used to better understand phenomena in the real-world. Common uses of models are for description, exploration, prediction, and classification. For example, Google builds models to understand and predict peoples' internet searching tendencies. These models are then used to help Google carry out more efficient and better searches of information. As another example, Netflix builds models to understand the characteristics of movies that their customers have rated highly so that they can then recommend other movies that the person may enjoy. Amazon and Apple iTunes both use models in similar manners.

Joshua Epstein in his keynote address, Why Model?, offers several purposes for building models.

## *Simulation*

One method that statisticians use to understand real-world phenomena is to conduct a simulation. A simulation is the manipulation of a model to enable a person to understand the potential outcomes and interactions of the system being modeled. In a simulation, a model is used to generate data under a particular set of conditions or assumptions. These conditions and assumptions allow the model to mimic processes and events in the real-world. By examining the data produced from the simulation, researchers can draw insight about and predict what might happen in the real-world under a given set of circumstances. Consider the following example.

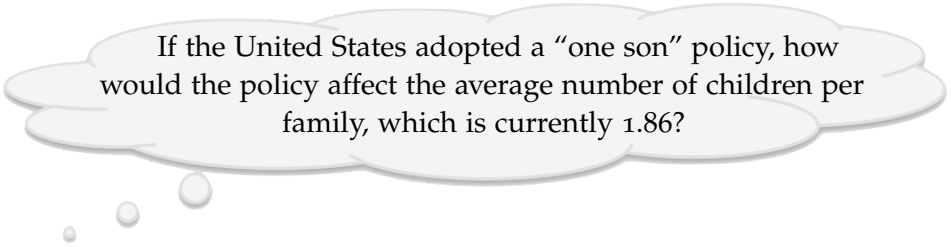
*Policy and Population*

In 1978, China introduced the "one-child" policy in order to alleviate social, economic, and environmental problems in China. According to Wikipedia,

> The policy officially restricts the number of children married urban couples can have to one, although it allows exemptions for several cases, including rural couples, ethnic minorities, and parents without any siblings themselves. A spokesperson of the Committee on the One-Child Policy has said that approximately 35.9% of China's population is currently subject to the one-child restriction.

Wikipedia. One-child policy—Wikipedia, the free encyclopedia, 2004.

Although the Chinese government has suggested that the policy has prevented more than 250 million births from its implementation to 2000, the policy is controversial both within and outside of China because of the manner in which the policy has been implemented. There have also been concerns raised about potential negative economic and social consequences, in part because many families were determined to have a son. Scholars have wondered how things would change if instead of a one-child policy, a country adopted a "one son" policy. A "one son" policy would allow families to keep having children until they had a son. If a family's first child is a boy, they would be restricted from having more children. If, however, the first child was a daughter, the family could continue having children until a son was born.




One way in which this question could be studied (without actually implementing the policy) would be to conduct a simulation study by modeling this situation. Consider for a minute

how you might model the number of children a particular family would have. One way to model this is to write the word `boy` on one index card and the word `girl` on another index card and to place those two index cards in a hat. After mixing up the index cards, you could draw a single card from the hat. If the card has the word `boy` written on it, the simulated "family" would be reported to have one child. If the card has the word `girl` written on it, a tally mark could be recorded and the index card would be replaced in the hat. The cards could then be remixed and another card would be drawn. If the card drawn has the word `boy` written on it, the simulated "family" would be reported to have two children. If the card has the word `girl` written on it, another tally mark could be recorded and the index card would again be replaced in the hat. This process would continue until the `boy` card was drawn. The table below shows the results after carrying out this process for three simulated families.

| Family | Girl | Boy |
|---|---|---|
| Family 1 | ✓ | ✓ |
| Family 2 | | ✓ |
| Family 3 | ✓✓ | ✓ |

The recorded number of girls and boys for three simulated families.

We could carry out this simulation for many families, say 500 families, and use the results to provide an answer to the research question. You can imagine that carrying out even this simple simulation would quickly become quite tedious. Simulation studies, such as this, are typically carried out using computer programs. In this unit, you will learn to use a computer program called TinkerPlots™ to model processes in the real-world and carry out simulation studies.

"Wait," you say. "Even if I carried out this simulation, I still would not be able to provide an answer to the research question! It doesn't reflect reality! Some families may not want to have any children, while others might be happy to stop after a girl was born. What about multiple births?" Maybe you are even questioning whether the probability of having a boy or having a girl is really 50:50. These are all valid points, and all

would likely affect the results of the simulation, which in turn affects the inferences and conclusions that are drawn.

Often incredibly complex models are used in carrying out research. As an example, Electronic Arts, the video game company behind titles such as `Madden`, *NHL* and *FIFA*, uses game telemetry to model the gameplay patterns of players and identify the elements of their games that are highly correlated with player retention. By understanding the behavior of players and the common patterns that are used, Electronic Arts game developers can focus their attention on more relevant features in future iterations of the game and ultimately reduce production costs. In their examination of `Madden NFL 11`, Electronic Arts used 46 features to model a player's mode preferences, control usage, performance, and playcalling style.

B. G. Weber, M. John, M. Mateas, & A. Jhala (2011). Modeling player retention in Madden NFL 11. Presented at *Innovative Applications of Artificial Intelligence*.

Game telemetry is the transmission of data from a game executable for recording and analysis.

To see other applications of how data are being used in video game design, watch the webinar, How Big Data and Statistical Modeling are Changing Video Games.

While the model we used in the previous example is overly simplistic for drawing any sorts of conclusions about implementing a "one-son" policy in the United States, it could however, provide a useful starting point for introducing additional complexity. Even in the most enormously complicated modeling problem, researchers often make many simplifying assumptions. With enough simplification, a model can be constructed and studied. The model is evaluated and often revised or updated as certain assumptions are deemed tenable and others are not. Because of this process, simulation studies are generally iterative in their development. This iteration process continues until an adequate level of understanding is developed and the research question can be answered.

Remember that *all* models—even those that seem quite complex—are simplifications of reality and get many things wrong.

## *Outline of the Unit*

In this unit, you will begin by exploring ideas related to randomness. Randomness permeates, and is in fact fundamental to understanding the models that are used in statistics. In learning about these ideas, you will also be confronted with

common human intuitions about randomness that are incorrect and can be misleading.

After examining the 'behavior' of randomness, you will be introduced to the TinkerPlots™ software, and learn how to model simple random processes such as coin flips and dice rolls. You will also learn how to carry out a simulation using TinkerPlots™. Later in the unit, you will have the opportunity to model more complex phenomena such as the effect of implementing a "one-son" policy.

At the end of this unit, you will learn how the models introduced in this unit are used to examine research hypotheses about the world. In particular, you will use a chance model to evaluate results that have been observed in research studies to judge the strength of evidence for particular tested hypotheses.

As you progress through the unit, remember that the modeling process is a creative process that can often be very challenging. At times, this might lead to frustration as you are learning and practicing some of the material. But, as Mosteller et al. remind us, it is also a profitable experience since "modeling is not only a technique of statistics . . . it is a method of study which can be applied in many other fields as well" (p. xii).

Mosteller, F., Kruskal, W. H., Link, R. F., Pieters, R. S., & Rising, G. R. (1973). *Statistics by example: Finding models*. Reading, MA: Addison-Wesley.

# *How Random is the iPod's Shuffle?*

Based on excerpts from an article by Carl Bialik in the September 21, 2006 *Wall Street Journal*, the *iPod User's Manual* and a Keynote Address by Steve Jobs

At the original iPod product launch, Steve Jobs, CEO of Apple, Inc. stated "...iPod, a thousand songs in your pocket. This is a major, major breakthrough." One of the amazing aspects about storing 1000 songs in your pocket is the ability to become your own disc jockey.

A feature built into the iPod software called "shuffle" further enhances this aspect. The shuffle feature takes a list of songs, called a playlist, and rearranges them in a random order. Each song will appear in the shuffled playlist only once.

In 2006, Mads Haahr ordered a customized iPod with "God Plays Dice" engraved on its back. Mr. Haahr—a random-number enthusiast, lecturer in computer science at Trinity College in Dublin and keeper of the Web site Random.org, a popular source of random numbers—intends to answer a question that has long bedeviled users of Apple's popular music player: Does the shuffle function really play users' songs in random order?

Since Apple Computer Inc. added the shuffle function to the main menu of iPods two years ago, the question has been raised by the *New York Times* and *Newsweek*; debated on Slashdot and other Web sites; and inspired a regular feature in *The Onion*.

The iPod's shuffle feature also has sparked interest from a cadre of random-number experts and enthusiasts such as Mr. Haahr. Just what makes a string of numbers random? Say you have ten songs in your iPod, numbered one to ten. A random

sequence must contain each number in equal frequency, so that, in the iPod example, none of your songs plays much more than any other in the long run. Also, it must be impossible to predict which number comes next, so song No. 5 can't always follow song No. 3.

How would you know if your sequence of numbers is random? Just looking at them wouldn't help. "People are notoriously bad at being a random number generator or recognizing something as random," said Landon Curt Noll, one of the creators of LavaRnd. People tend to seek patterns and order where none exist—perhaps even in a shuffled iPod playlist, where they might pay more attention when their favorite songs are playing, and thus assume that those songs are in heavier rotation.

***Be ready to share and discuss your responses to each of the following questions with your group.***

1. Do you have an iPod or some other digital music player? Have you used the shuffle feature? If you have used the shuffle feature, have you ever wondered how truly random it is?
2. What comes to mind when you hear the word, 'random'?
3. If the iPod shuffle feature is not producing a random sequence of songs, then what might the sequence of songs look like? What would you expect to see?
4. Do you think you can be 100% certain that a sequence of songs was not randomly generated? Explain your answer.

# *Course Activity: iPod Shuffle*

***Share and discuss your responses to each of the following questions with your group.***


1. Do you have an iPod or some other digital music player? Have you used the shuffle feature? If you have used the shuffle feature, have you ever wondered how truly random it is?

2. What comes to mind when you hear the word, 'random'?

3. If the iPod shuffle feature is not producing a random sequence of songs, then what might the sequence of songs look like? What would you expect to see?

4. Do you think you can be 100% certain that a sequence of songs was not randomly generated? Explain your answer.

## *Group Task*

Albert Hoffman, an iPod owner, has written a letter to Apple to complain about the iPod shuffle feature. He writes that every day he takes an hour-long walk and listens to his iPod using the shuffle feature. **He believes that the shuffle feature is producing playlists in which some artists are played too often and others are not played enough.**

He has claimed that the iPod Shuffle feature is not generating random playlists. As evidence, Mr. Hoffman has provided both his music library (8 artists with 10 songs each) and three playlists (20 songs each) that his iPod generated using the shuffle feature.

Tim Cook, the CEO of Apple, Inc., has contacted your group to respond to Mr. Hoffman's complaint. He has provided your group with several playlists of 20 songs each using the same

songs as Mr. Hoffman's library but generating them using a genuine random number generation method.

To help your group respond to Mr. Hoffman, the next four sections of the problem are designed to help your group explore properties of the randomly generated lists to develop rules that could help determine whether a set of playlists provide evidence that the shuffle feature is not producing randomly selected songs.

### *Explore and Describe*

Examine the randomly generated playlists (your group will be given 25) to get an idea of the characteristics of these lists. Write down at least two characteristics about the randomly generated playlist that help you address Mr. Hoffman's concern.

### *Develop Rules*

Use the set of characteristics that your group wrote down to describe randomly generated playlists in the previous section to create a set of one or more rules that flag playlists that **do not appear to have been randomly generated.** (Be sure that each of the characteristics in the previous section is included in a rule.) *These rules should be clearly stated so that another person could easily use them.*

### *Test Rules*

Your group will be given five additional randomly generated playlists on which to test your rules. Let your instructor know that you are ready to receive these playlists. See whether the set of rules your group generated would lead someone to (incorrectly) question whether these playlists are not randomly generated. Based on the performance of your group's set of rules, adapt or change the rules as your group feels necessary.

*Evaluate*

Your group will be provided with Mr. Hoffman's original three playlists. Apply your group's rules to these three playlists to judge whether there is convincing evidence that Mr. Hoffman's iPod Shuffle feature is producing playlists which do not seem to be randomly generated.

*Summarize*

Your group will now write a letter to Mr. Hoffman that includes the following:

- Your group's set of rules, used to judge whether a playlist does not appear to have been randomly generated. In your letter the rules need to be clearly stated so that another person could apply them to a playlist of 20 songs from Mr. Hoffman's music library;
- A response to Mr. Hoffman's claim **that the shuffle feature is not random because it produces playlists in which *some artists are played too often and others are not played enough.***

*Type the letter in a word-processed document and email it to each of your group members and the instructor.*

## *Discussion*

As a group, discuss your responses to each of the following questions.

5. What made it difficult to come up with a rule to determine whether a sequence of data had been randomly generated? Explain.

6. How might your rules ***change*** if there were not an equal number of songs for each artist? Or a longer set of songs per playlist? (Be specific about how your rules might change.)

7. What does your group need to do to improve the process of working as a team? Be specific about how each member of the group will contribute to this improvement.

# *Randomness*

To follow up on the *iPod Shuffle* activity, we will continue the exploration of randomness and human beings' intuitions of randomness. In upcoming class periods, you will explore some of your own intuitions of probabilistic devices. To prepare for this, you should do the following:

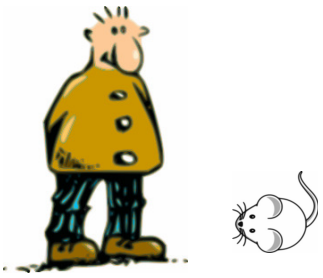
- Watch the YouTube video *Random Sequences: Human vs. Coin* http://www.youtube.com/v/H2lJLXS3AYM

- Read a blog entry about randomness called *What is Randomness?* http://sxxz.blogspot.com/2005/08/what-is-randomness.html

- Watch the YouTube video *Fun Science: Randomness* http://www.youtube.com/watch?v=t8mSHosc0Dk

# *Course Activity: Can You "Beat" Randomness—Part I*

In his best-selling book *The Drunkard's Walk: How Randomness Rules Our Lives*, Leonard Mlodinow describes a probability guessing game that psychologists have played with different subjects such as humans, rats, and other non-human animals: Subjects are shown a sequence of red and green lights or cards, in which the colors appear randomly with different probabilities but otherwise with no pattern. The probability of a specific color stays the same. After subjects watch the colors appear for a while, they are then asked to predict the color that will appear on subsequent lights or cards.

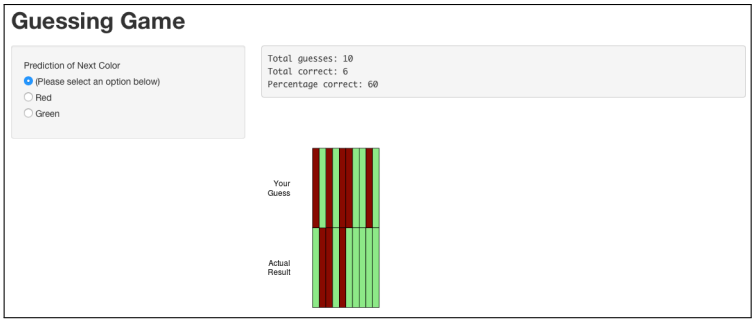


Mlodinow, L. (2008). *The drunkard's walk: How randomness rules our lives*. New York: Pantheon.

We are going to play a game, similar to the one described by Mlodinow.

- You will first watch a sequence of colors at: `https://github.com/zief0002/Statistical-Thinking/blob/master/images/animation.gif`. (Note: Although the sequence is only 50 generated squares, it will repeat, so stop the animation when you are satisfied.)
- After watching the sequence of colors, access the guessing game at `http://statisfactions.shinyapps.io/guessing-game/`
  - Based on the sequence you just watched, predict the color of the next light. To make a prediction, click once on the radio button for "Red" or "Green".
  - Once you make your prediction, the program will display your prediction in the "Your Guess" row. The program will then randomly generate a color and display it in the "Actual Result" row.

– On the top of the window, you can also see the total number of predictions you have made, how many of those predictions were correct (matched the outcome to the actual color), as well as the percentage of your total predictions that were correct.




Screenshot from the online guessing game after making ten predictions. Six of the ten predictions were incorrect.

1. Play the game one time (i.e., make 50 predictions). What percentage of your 50 predictions were correct?

2. Describe the strategy that you used to make your predictions.

3. Play the game again using the strategy you just described. What percentage of your 50 predictions were correct this time?

4. Suppose that while playing the game you observed 10 green lights in a row. Can you predict the color of the next light? Explain.

5. Suppose that while playing the game you observed 20 green lights in a row. Can you predict the color of the next light? Explain.

# *Course Activity: Modeling Random Behavior—Part I*

In the book *Randomness*, author, Deborah J. Bennett, states,

> Everyone has been touched in some way by the laws of chance. From shuffling cards to start a game of bridge to tossing a coin at the start of a football game--most of humanity encounters chance daily. Our vocabulary is full of phrases that involve chance: likely; unlikely; probability; odds; chances; random; etc.

Bennett, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.

Bennett describes the concept of randomness as being deceptively complex because many aspects of it are counterintuitive. In particular, she points out that misconceptions about randomness and probability are dangerous due to the constant use of statistics, probabilities and odds in everyday life.

You will be exploring the following research question:

How good are people at predicting random outcomes of common chance devices such as coins and dice?

## *Intuitions about Coin Flips*

*Questions 1 through 7 are asking for your intuitions. You do not have to calculate exact values. We will explore these questions in more detail later in this activity.*

1. Imagine that you flip a fair coin ten times and record the outcome for each flip. How many heads would you expect to see?

2. What is the smallest number of heads you could obtain?

3. What is the largest number of heads you could obtain?

*Imagine flipping a fair coin ten times and recording the outcome for each flip. Once you obtain the ten flips, you count the number of heads. This process is repeated 100 times. Each time you record the number of heads out of the ten.*

4. Which result(s) for the number of heads would you expect to occur the most often?

5. How often—what percentage of the 100 sets of 10 coins—would you expect to get a result of five heads?

6. How often—what percentage of the 100 sets of 10 coins—would you expect to get a result of all ten heads?

7. Which result, two heads or eight heads, would you expect to see more often? Why?

## *Modeling Coin Flips*

To save time and to gather data quickly, you will use a software package called TinkerPlots™ to model tossing a coin 10 times. You will use the data you generate with the simulation to check your initial intuitions about what would happen if you repeatedly tossed a single coin 10 times.

### *Setting Up the Model*

In simulation studies, a model is used to generate data. A model is a representation for a particular purpose. Models can be used for description, exploration, prediction, or classification. A model might be a physical object or it might be an idea or construct.

- Set up the model for flipping a coin (see instructions in margin).
- After you have set up the model, click the `Run` button.
- A *case table* displaying the outcome for the "coin flip" should have been produced.
- Record the result in the first cell of "Trial 1" in the table below. Run the simulation nine more times and record the outcome for each additional "flip" in the "Trial 1" column.

**Setting Up the Simulation**

- Open TinkerPlots™
- Drag a new `Sampler` from the object toolbar into your blank document.


- Drag a new `Spinner` from the device toolbar at the bottom of the `Sampler` onto the current device.


- There are two outcomes in our spinner, *a* and *b*. Click on each of these and change their names to *Heads* and *Tails*.
- From the *Device options* menu (to the lower-left of the *Spinner*) select *Show Proportion*.

- In modeling a coin flip, *Heads* and *Tails* are equally likely events. Therefore you need to change the proportions so that each is 0.5. Change this by either dragging the dividing line in the spinner or by clicking on the proportion and changing it to the desired value.
- Change the `Draw` value from 2 to 1 and the `Repeat` value from 5 to 1. This simulates tossing a coin one time.
- Change the `Spinner` label from *Attr1* to *Coin*.

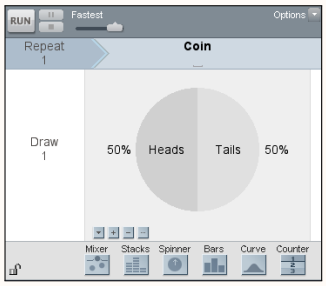



Figure 1: A TinkerPlots™ Sampler Showing the Model for Flipping a Coin

| Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

8. How many *Heads* did you get in the ten "flips"?

In simulation experiments, each time the model is used to produce outcomes, it is referred to as a *trial*. A trial can consist of one or many outcomes depending on the simulation. In this simulation, the trial consists of flipping the coin 10 times. The *result* from each trial is the number of heads. You often need to examine the data from several trials to understand a phenomenon under study.

9. Carry out four more trials of the simulation. Record the outcomes in the table above. Also record the result (e.g., the number of heads) from each trial.

## *Collecting the Results from Many Trials*

In order to answer the questions posed previously, you need to collect the results from each trial of the simulation experiment.

- Enter the results from each of your five trials into a case table (see instructions in margin).

Collecting the results from each trial of a simulation into a case table allows us to organize and display the results from multiple trials. For example, right-clicking on the attribute name (if you are on a Mac press the control key while you click), brings up a menu from which you can sort these results in either ascending (`Sort Rows Ascending`) or descending (`Sort Rows Descending`) order. This will allow us to answer questions such as 'what is the maximum number of heads'.

**Collecting the Results from Each Trial**

- Drag a new `Case Table` from the object toolbar into your blank document.


- Click on `<new>` to change the attribute name. Rename this attribute *Results*.
- Enter the results from each trial (the number of heads) into the results column in the case table.

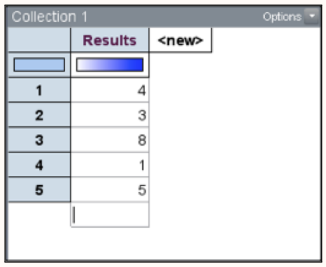



Figure 2: A TinkerPlots™ Case Table Showing the Results for Five Trials

### *Plotting the Results from Many Trials*

Although you can see the result from each trial in the case table, this is not a good way to organize or use the results—especially when there are results from more than just a few trials. For example, consider trying to compare the number of trials in which you got two *Heads* and the number of trials in which you got eight *Heads*. A better way to organize simulation results, after entering them into a case table, is to plot them.

- Plot the results from your five trials (see instructions in margin).

10. Sketch the plot below.

**Plotting the Results from Many Trials**

- Highlight the *Results* attribute from your case table.
- Drag a new `Plot` from the object toolbar into your document.


- All five trial results are now displayed in your plot, but in an unorganized manner.
- Grab one of the circle icons in the plot and drag it to the right. This will help organize the counts. (Note. You may want to resize the window so it is larger.)
- Keep dragging until you are satisfied that you can answer the questions posed earlier.
- Although the data are now more organized along the horizontal axis, the heights are still arbitrary. You can change this by clicking the `Stack Vertical` button in the upper plot toolbar.

**Bin Lines**

Each time you create a plot in vertical bars—called bin lines—will be drawn in by default. When you separate cases by dragging, TinkerPlots™ adds bin lines. These lines show how groups of cases are separated. Labels for each separated group (such as "0–4" and "5–9") are also added to the plot.

To separate cases so that each group only contains a single value, double-click on one of the endpoints and change the value for `Bin width` (i.e., `Bin width = 1`) and click `OK`.

To fully separate the cases so that there are no groups, drag a case icon until the bin lines disappear.

Can you answer the questions posed earlier using your simulation results? You may not feel comfortable answering those questions based on the results from only five trials. You should carry out more trials and add the results to our case table.

- In the `Sampler` from which you originally ran the simulation, change the `Repeat` value from *1* to *10*.
- Click the `Run` button.

This simulates flipping the coin ten times rather than once.

- Add the trial result to your case table of results. Notice that the plot of the results automatically updated when the result was added to the case table.
- Carry out several more trials of the simulation. Add each trial result into your case table. Be sure you have the results from at least ten trials.
- Enter the result from each of your ten trials into your instructor's computer.

11. After each group in the class has entered their data, sketch the plot of results shown on your instructor's compiuter. Be sure to add an appropriate scale and label to your $x$-axis so that you can respond to the questions below.

12. Which result(s) for the number of heads occurred the most often?

13. About what percentage of the trials resulted in five heads?

14. About what percentage of the trials resulted in ten heads?

15. Which result, two heads or eight heads, occurred more often?

## *Intuitions about Dice*

Imagine that you are rolling a six-sided die ten times. In those ten rolls, consider the number of times that you would expect to see the outcome of three.

*Imagine repeating this process 100 times.*

Questions 15 through 18 are asking for your intuitions. You do not have to calculate exact values. We will explore these questions in more detail later in this activity.

16. Which outcome (0 threes to 10 threes) would you expect to occur most often?

17. What percentage of the time would you expect to see an outcome of five threes? Explain.

18. Which outcome, two threes or eight threes, would you expect to see more often? Why?

19. What percentage of the time would you expect to get an outcome of all ten threes?

## *Modeling Dice Rolls*

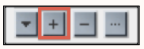
Now you will use TinkerPlots™ to simulate rolling a die ten times. You can again use the data you generate in the simulation to check your initial intuitions about the number of threes that would occur in ten rolls.

### *Setting Up the Model*

- Open a new document in TinkerPlots™. To open a new document, under the `File` menu select `New`.
- Set up the model of tossing ten dice (see instructions in margin).
- After you have set up the model, click the `Run` button to carry out a single trial of the simulation.

**Setting Up the Model: Die Rolling**

- Drag a new `Sampler` from the object toolbar into your blank document.
- This time rather than changing the `Sampler` device to a `Spinner`, you will use the default device of the `Mixer`.
- Click the `Add Element` button (the plus sign) until you have six elements in your mixer.



- Click on each element label and change the labels to numbers one through six, so that each represents one side of a die.
- Change the `Draw` value from *2* to *1* and the `Repeat` value from *5* to *10*. This simulates rolling the die ten times.
- Change the mixer label from *Attr1* to *Dice*.

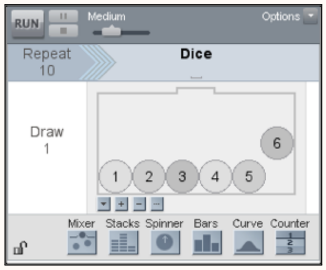



Figure 3: TinkerPlots™ Sampler Showing the Model for Rolling a Die Ten Times

In this simulation, the trial consists of rolling the die ten times. The result from each trial is the number of threes that occur.

- Record the number of *threes* that occurred in the trial in a case table. (If you've forgotten how to get a case table, go back and re-read the directions from earlier in this activity.)
- Carry out nine more trials. Record the results from each trial into your case table.
- Enter the result from each of your ten trials into your instructor's computer.
- Add the results from at least four other groups to your case table so that you have the results from 50 trials.
- Create a plot of the results. (If you've forgotten how to plot the results, go back and re-read the directions from earlier in this activity.)

Use the plot of the results from the 50 trials of the simulation to answer each of the following questions.

20. Sketch a plot of the results. Be sure to label the axis.

21. Which result occurred most often?

22. What percentage of the time did a result of five *threes* occur?

23. Which result, two *threes* or eight *threes*, occurred more often?

24. What percentage of the time did a result of ten *threes* occur?

### *Extensions*

25. Compare both plots of the results with those of classmates from at least two other groups. Are your plots identical? Comment on the similarities and differences.

26. Based on your examination of these plots, sketch the plot you would expect to see if you could simulate the results from 10,000 trials (rather than 50) of the coin flipping experiment.

27. Based on your examination of these plots, sketch the plot you would expect to see if you could simulate the results from 10,000 trials (rather than 50) of the die rolling experiment.

28. Sketch the plot you would expect to see if you could simulate the results from 10,000 trials of the coin flipping experiment, but instead of flipping the coin ten times, you flipped it 20 times.

29. Sketch the plot you would expect to see if you could simulate the results from 10,000 trials of the die rolling experiment, but instead of rolling the die ten times, you rolled it 20 times.

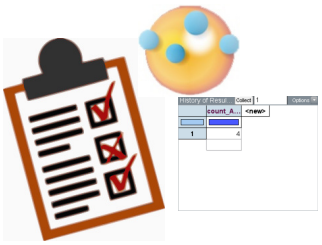
# *Probability Simulation*

To follow up on the *Modeling Random Behavior—Part I* activity, you will get a chance to practice creating models using TinkerPlots™. In upcoming class periods, you will add to these fundamental TinkerPlots™ skills to build increasingly complex models and carry out simulations. One helpful resource for learning TinkerPlots™ are the tutorial videos. These are available both online (`http://www.keycurriculum.com/PreBuilt/tp2/movies/`) and via the `Help` menu in TinkerPlots™.

- Watch the TinkerPlots™ movie *Probability Simulation*. This will help to reinforce some of the ideas that we have been working with in class. `http://www.keycurriculum.com/PreBuilt/tp2/movies/probability-simulation.html`

# *Course Activity: Modeling Random Behavior—Part II*

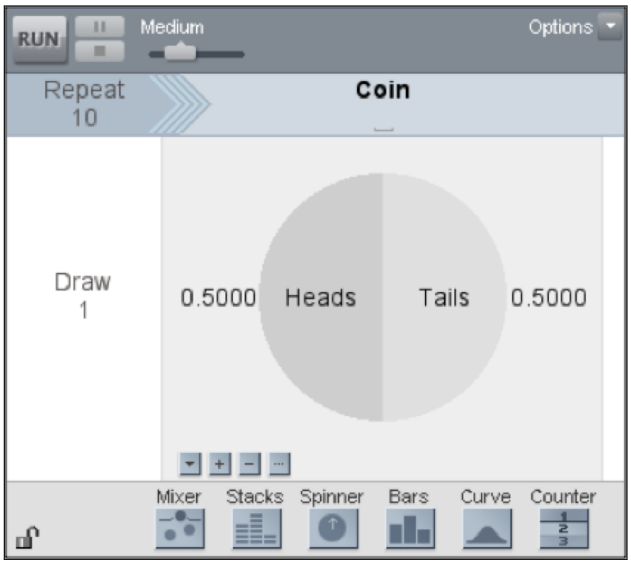
In previous activities and assignments, you have learned how to set up a model to run a simulation experiment using TinkerPlots™. In these simulations, you ran many trials from which you collected a particular outcome (e.g., the number of Heads when flipping a coin ten times). You also learned how to create a case table to collect the results from each trial into, and how to plot those results.



In this activity, you are going to be introduced to the `Collect` functionality in TinkerPlots™. This will automate the collecting of trial results in a simulation. It will also make carrying out several trials easier.

## *Modeling Coin Flips*

In TinkerPlots™, set up a model to simulate tossing a coin ten times. (If you have forgotten how to do this, refer back to the previous class activity.) The figure on the next page might also help to jog your memory.




Model to simulate tossing a coin ten times.

You will use the data you generate with the simulation to check your initial intuitions about what would happen if you repeatedly tossed a single coin ten times.

- After you have set up the model, click the Run button.

### *Automating the Collection of Trial Results*

Rather than having you record the number of Heads that occurred in the 10 flips, you can have TinkerPlots™ do this for you.

> The general idea for having TinkerPlots™ record and collect the trial results is:
> - Plot the outcomes from the trial.
> - Collect the result.

## *Plotting the Outcomes from a Trial*

The first step is to plot the trial outcomes produced from running the model. Remember, plotting in TinkerPlots™ is achieved by highlighting the attribute (column) in a case table that you want to plot, and then dragging a `Plot` from the object toolbar into your document.

- Plot the trial outcomes produced from running the simulation (see instructions in margin).

**Plotting the Trial Outcomes**

- Highlight the attribute *Coin* in the case table.
- Drag a new `Plot` from the object toolbar into your blank document.
- You should now have a plot with the trial's ten outcomes color coded by whether each was *Heads* or *Tails*.
- Drag one of the circle icons in the plot until the outcomes are separated into two groups, *Heads* and *Tails*.
- Stack the outcomes using the `Vertical Stack` tool.

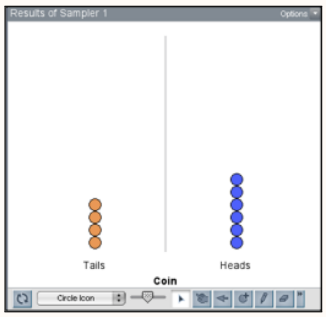



Figure 1: TinkerPlots™ plot showing the stacked, separated trial outcomes after running the model for flipping a coin 10 times.

## *Computing the Trial Result in the Plot*

In simulation experiments, after a trial is carried out, some result is typically recorded. A result is simply a numerical summarization of the outcomes in a particular trial.

For example, in our coin flipping simulation, the result you are interested in is the number of *Heads* that occurred in the ten flips of the coin. The number of *Heads* is a summary measure of the trial outcomes. Consider Figure 1 (in the margin).

The number of *Heads* in this trial is 6. Of course the exact same trial could have been summarized using a different measure. For example, the count of *Tails* (4); the percentage of *Heads* (60%); or the percentage of *Tails* (40%) could have been used instead. The choice to summarize the outcomes using the number of *Heads*, however, allows you to answer the questions posed in the first part of the activity.

The upper plot toolbar (see below) offers several built-in options for summarizing the plotted outcomes of a trial.


The upper plot toolbar.

The two `Counts` buttons—`Count (N)` and `Count (%)`—can be used to summarize the number and percentage of cases in a plot.

- Use `Count (N)` to have TinkerPlots™ count the number of *Heads* and *Tails* that occurred in the trial (see instructions in margin).

Note that this will count the number of cases within each section of a plot. If there are not multiple sections (no bin lines), the number of total cases in the plot will be displayed.

**Summarizing the Outcomes in a Trial**

- Highlight the plot of the trial outcomes.
- Click on the `Count (N)` button in the upper plot toolbar.

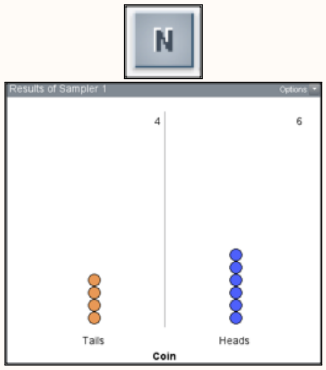



Figure 2: TinkerPlots™ plot after clicking the `Count (N)` button in the upper tool bar.

## *Collecting the Results from Many Trials*

You can also use TinkerPlots™ to automatically collect the summarized result from the trial into a case table.

- Use TinkerPlots™ to automatically collect the result from your simulated trial into a case table (see instructions in margin).

It is important that you right-click on the actual value of the result in the plot you want TinkerPlots™ to collect. For example, in the plot displayed in Figure 3, you would right-click on the value *6*.

The result is then collected in a new case table. This case table, which is called *History of Results*, has a single row with the collected result, in this case six, displayed in a new attribute. The window next to the `Collect` button indicates the number of results that were collected, in this case one. This value can be changed to add the results of additional trials into the case table.

**Collecting the Results from a Trial**

- Right-click the summary result in your plot.
- Select `Collect Statistic`.

Figure 3: After right-clicking on the summary result in a TinkerPlots™ plot, choose `Collect Statistic` from the menu.

- Change the value in the *History of Results* case table to *99* to add the results from an additional 99 trials of the simulation (see instructions in margin).

**Collecting the Results from Additional Trials**

- Change the value in the *History of Results* case table from *1* to *99*.
- Click the `Collect` button.

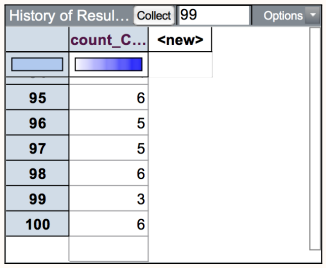



Figure 4: The *History of Results* case table showing the results for 100 trials of the coin flipping simulation.

The result collected from each trial is stored in a row of the *History of Results* case table.

30. Record the result from the $87^{th}$ trial.

- Plot the results from your 100 simulated trials. (If you have forgotten how to do this, refer back to the instructions in the previous activity.)

Use the data you collected from the 100 trials of the simulation to answer each of the following questions.

31. Which result(s) for the number of *Heads* occurred the most often?

32. How often–what percentage of the 100 simulated results—did you get a result of five *Heads*?

33. How often—what percentage of the 100 simulated results—did you get a result of ten *Heads*?

34. Which result, two Heads or eight *Heads*, occurred more often?

## *Modeling Dice Rolls*

Now you will use TinkerPlots™ to simulate rolling a die ten times. You can again use the data you generate in the simulation to check your initial intuitions about the number of *threes* that would occur in ten rolls.

- Open a new document in TinkerPlots™.
- Set up the model of tossing a die ten times.
- Carry out a single trial of the simulation.
- Plot the ten outcomes from the simulated trial.
- Stack and separate the cases into groups.
- Use `Count (N)` to summarize the number of cases in each group.
- Collect the number of *threes* that occurred into a *History of Results* case table.
- Carry out an additional 99 trials.
- Plot the results from your 100 simulated trials.

Use the plot of the results from your 100 simulated trials to answer each of the following questions.

35. Sketch a plot of the results. Be sure to label the axis.

36. Which result occurred most often?

37. What percentage of the time did a result of five *threes* occur?

38. Which result, two *threes* or eight *threes*, occurred more often?

39. What percentage of the time did a result of ten *threes* occur?

### *Extensions*

40. Consider the coin flipping simulation experiment that you ran. Were you able to predict the result for any one trial? Why or why not?

41. What are you able to predict in the coin flipping simulation experiment?

42. Consider the die rolling simulation experiment that you ran. Were you able to predict the result for any one trial? Why or why not?

43. What are you able to predict in the die rolling simulation experiment?

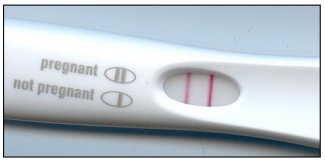
44. Describe how this is similar to what you learned in the *iPod Shuffle* activity.

# *Pregnancy Tests Readiness*

TBA

# *Course Activity: Pregnancy Tests*

Pregnancy tests have evolved greatly over the years. Many of the home pregnancy tests make strong claims. For example, the First Response Gold® Digital Pregnancy Test claims it can give "results as early as five days before the day of your missed period".




An important question would be

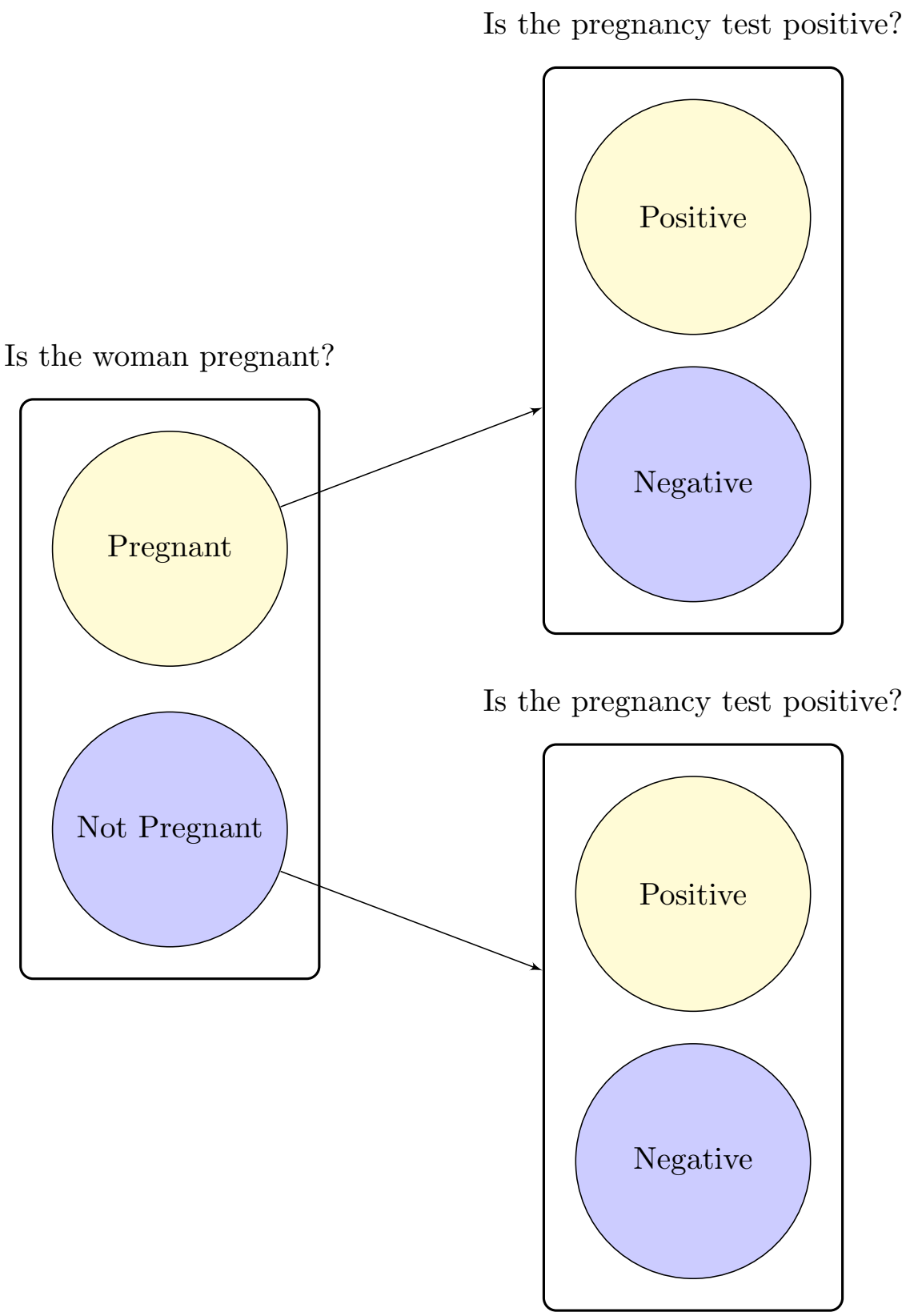
How accurate are the test results from the First Response Gold® Digital Pregnancy Test?

To answer this question, researchers conducted a clinical trial with 215 women who were trying to become pregnant. The women took the First Response Gold® Digital Pregnancy Test daily starting five days before the day of their expected period.

Cole, L. A. (2011). The utility of six over-the-counter (home) pregnancy tests. *Clinical Chemistry & Laboratory Medicine*, *49*(8), 1317–1322.

***Use the diagram on the next page to help respond to the following questions.***

1. To the right of each test result (positive or negative) write "correct" or "incorrect" based on whether the pregnancy test would give a correct or incorrect result.


Is the pregnancy test positive?
Positive
Negative
Is the woman pregnant?
Pregnant
Not Pregnant
Is the pregnancy test positive?
Positive
Negative

2. The researchers found that for women who were pregnant, the First Response Gold® Digital Pregnancy Test correctly gave a positive test result 42% of the time. For women who were NOT pregnant, the First Response Gold® Digital Pregnancy Test correctly gave a negative result 95% of the time. Write the percentages for each of the four test results in the corresponding circle.

The percentage of time a test gives the correct result, both positive and negative, is referred to as the *accuracy* of the test.

3. Suppose that *every woman* in the Twin Cities, age 18–50 took this pregnancy test. Estimate the accuracy of the pregnancy test (percentage of the pregnancy tests taken that are correct) for this group of women. Explain.

4. Suppose that **only** the women in the Twin Cities, age 18–50, **who were trying to become pregnant** took this pregnancy test. How would the accuracy of the pregnancy test (percentage of correct tests) for this group of women compare to the accuracy you estimated in the previous question? Explain.

## *Modeling the Accuracy of the First Response Gold® Digital Pregnancy Test*

We will use TinkerPlots™ to model the accuracy of the First Response Gold® Digital Pregnancy Test for both groups of

women (every woman in the Twin Cities, age 18–50, and only those who are trying to become pregnant). First, we will model the results of the test for every woman in the Twin Cities, age 18–50, henceforth referred to as the "low pregnancy" group, and then we will model the results of the test for women in the Twin Cities, age 18–50 who were trying to become pregnant, henceforth referred to as the "high pregnancy" group.

5. Why do you think we named these groups the "low" and "high" pregnancy groups?

### *Modeling the Accuracy for the "Low Pregnancy" Group*

Suppose we want to see how accurate the test is for for every woman in the Twin Cities, age 18–50. Among these women, we believe that only 6% are pregnant.

- Create a model in TinkerPlots™ using a spinner that represents selecting 100 women from this population (each woman has a 6% chance of being pregnant). If you have forgotten how to do this, refer back to the coin model in *Modeling Random Behavior, Part I*.
- Label this attribute "Pregnancy", and label the two outcomes "Pregnant" and "Not Pregnant".

Unlike previous simulations in TinkerPlots™, we cannot use a single model to answer our research question. The accuracy of the test depends not only on whether the woman is pregnant, but also on the result of the test. Recall that the probability of a correct test result differed whether the woman was actually

pregnant or not. We need to create separate models for the test results that depend on whether or not the woman was pregnant. In computing terms, this idea is called *branching*.

- Increase the size of the sampler window to give more room for the additional sampling devices you will be adding.
- Add a linked spinner to the right of the original spinner to represent the results of the test. (see sidebar).
- Add a second linked spinner on the bottom right to represent the accuracy of the test for non-pregnant women (see sidebar).
- Adjust the percentages in the second and third linked spinners to match the accuracy of the test when pregnant and not pregnant.

**Branching using Multiple Sampling Devices**

- Drag another spinner to the right of the existing "Pregnancy" Spinner in the sampling window. Note you will see a black rectangle appear that will show you where the spinner will be placed.
- Drag a second spinner (so you have three total) to the right of the existing "Pregnancy" Spinner, but place it below the last spinner you dragged in. Again, you will see a black rectangle appear that will show you where the spinner will be placed.
- Label the new attribute "Result".
- Add two elements to each of the new spinners and label one "Positive" and one "Negative". (Each spinner should have a "Positive" and "Negative" element.)
- Set up these two new spinners to reflect the appropriate probabilities based on the test's accuracy rates you indicated on the diagram from before. (Note: Pay attention to the labels, TinkerPlots™ uses alphabetical order.)

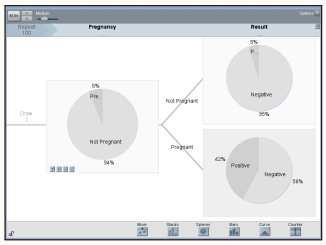



Figure 1: Sampler Options Window in TinkerPlots™

## *Simulate Selecting 100 Women at Random*

Run the simulation with the simulation speed initially set at "Medium" so that you can understand how the model operates. (Once you have viewed 10 or so outcomes you can change the speed to "Fastest".) Obtain 100 results from the simulation.

6. What does each case (row in the results table) represent?

7. Find the 100th case in your simulation results. Does this case represent a pregnant or non-pregnant woman? Were her pregnancy test results positive or negative?

8. Was the test accurate for determining whether the 100th case was actually pregnant?

9. Which of the two "Results" spinners did TinkerPlots™ use most frequently during this simulation?

To examine the accuracy of the pregnancy test for these 100 women, plot the "join" attribute form the table of results. Add the percentages of each of the four groups to the plot.

10. Sketch and label the plot of the "join" attribute.

11. Based on your simulated percentages, what was the accuracy of the pregnancy test. Remember that accuracy refers to the percentage of correct results (positive and negative).

### *Modeling the Accuracy for the "High Pregnancy" Group*

Now, suppose we want to see how accurate the test is for a group of 100 women who are trying to become pregnant. Among this "high pregnancy" group of women, we will assume that 65% are pregnant.

- Set up another sampler that will allow you to simulate from this group. Note that everything in this sampler will be identical to the previous sampler except for the percentage of women who are pregnant.
- Carry out a simulation that samples 100 cases from this group.

# *Introduction to Statistical Hypothesis Testing*

In the previous course activities and homework, you have been exploring events, phenomena, and processes by using a chance model to simulate the particular outcomes or results that occur under such a model. This is the same kind of process that researchers and statisticians engage in when they test statistical hypotheses. Statistical hypothesis testing is a method that can be used to evaluate the likelihood of a specified model given an observed set of data. To illustrate the ideas of statistical hypothesis testing, consider how you might go about testing a coin for "fairness".

You might have suggested something along the lines of "flip the coin many times and keep track of the number of heads and tails". Suppose you tossed the coin 100 times, which resulted in 53 heads and 47 tails. Would you say the coin is "unfair"? What if you had obtained 65 heads and 35 tails instead? Now would you say the coin is "unfair"? How about if you had gotten 84 heads and only 16 tails?

The first result of 53 heads and 47 tails probably did not seem that far fetched to you, and you probably would feel at ease saying that the coin that produced such a result is most likely "fair". On the other hand, the results of 65 heads and 35 tails—and especially 84 heads and 16 tails–likely made you feel uncomfortable about declaring the coin "fair". Why is this? It is because you had a model in mind that you could use to evaluate the observed results.

Most people, when evaluating whether a coin was "fair", compare the observed results to the results that would be

expected if a "fair" coin was flipped 100 times. Such a coin is expected to produce a 50:50 split between heads and tails. If the observed results are close to what is predicted under the assumed "fair" coin model, the model is not disputed. For example, the result of 53 heads from 100 flips is very close to the 50:50 split of heads and tails, and it is probably safe to say the coin flipped was "fair". The other two sets of results, however, did not conform to what was expected if a "fair" coin was flipped 100 times. This might lead you to reject the model of a "fair" coin.

### *Models and Hypotheses*

The key to drawing an inference about whether or not the coin flipped in the previous example was "fair", was the assumption or positing of a particular model—in this case the "fair" coin model. By assuming a particular model, we can then use the model to predict the outcomes or results that are expected under the model. Our observed results are then evaluated within the set of expected outcomes produced from the model. If the observed results do not conform to what is expected, they act as evidence to dispute the model that was assumed.

This may sound like a straight-forward process, but it can actually be quite complex—especially as you are reading research articles and trying to interpret the findings. First off, the assumed model is often not provided, nor described, explicitly within most research articles. It is often left to the reader to figure out what the assumed model was. At first, this may be quite difficult, but like most tasks, as you gain experience in this course and as you read more research, you find that there are a common set of models that are often assumed by researchers.

The models that are assumed are typically related to the research question being asked. Often researchers explicitly state *hypotheses* about their research questions. Hypotheses are simply statements of possible explanations for an observed set of data. For example, one possible explanation for the observed

set of coin flips is that *the coin being flipped is a "fair" coin*.

One complication that you may encounter is that many statisticians and researchers write their hypotheses mathematically. The advantage to writing a hypothesis mathematically is that it explicitly states the assumed model. Consider the stated hypothesis that *the coin being flipped is a "fair" coin*. If, as earlier, we assume this means that a "fair" coin leads to a 50:50 split of heads and tails, we might express the hypothesis more mathematically as:

The probability of heads produced by flipping the coin is 0.5.

Notice that by expressing the hypothesis mathematically, we have stated the assumed model within the hypothesis. Although the more descriptive hypothesis that *the coin being flipped is a "fair" coin* might lead us to use the 50:50 model, by explicitly stating it, there is no question about what we meant by a "fair" coin. It is also common to see hypotheses written using symbolic notation. The symbolic notation acts as a shorthand to quickly state the same hypothesis. For example, the exact same hypothesis can be stated as,

$$\mathrm{H}_0 : \pi = 0.5.$$

The symbol $H_0$ indicates that the hypothesis that follows the colon specifies the model to be assumed. The model that is specified is sometimes reffered to as the *null model*, and thus the entire hypothesis is sometimes referred to as the *null hypothesis* s. The greek letter $\pi$ stands for probability or proportion. If this all seems like gibberish to you right now, don't worry about it. You can always write hypotheses descriptively without resorting to the symbolic notation. We only include this here to point out that all three manners of stating the hypotheses are equivalent and you will likely see different expressions of hypotheses as you read research articles or take other courses.

## *Connections*

In the upcoming course activities, you will explore this process of testing statistical hypotheses. You will be introduced to several common models that are assumed by researchers and statisticians. You will also use TinkerPlots™ to generate simulated data that would be expected under these models. Many of these models are directly related to the chance models that you have explored in the course to this point. For example, you should already be able to use TinkerPlots™ to produce results that would be expected from 100 flips of a "fair" coin. Aside from learning about some of the more common models used in research, you will also learn how to quantify the evidence an observed result provides against the assumed model, and how to use that evidence to evaluate a particular hypothesis and in turn answer a research question.

# *Course Activity: Matching Dogs to Owners*

Do dogs look like their owners (or vice versa)? In a study published in 2004 in *Psychological Science*, researchers Roy and Christenfeld at the University of California, San Diego conducted a study to answer this question. They had observers attempt to match dogs with their owners and also explored whether any ability to make such matches is due to people selecting dogs who resemble them. Forty-five dogs and their owners were photographed separately, and judges were shown one owner, that owner's dog, and one other dog, with the task of picking out the true match. The results suggest that when people pick a pet, they seek one that, at some level, resembles them.

M. M. Roy & N. J. S. Christenfeld. (2004). Do dogs resemble their owners? *Psychological Science, 15*(5), 361–363.

In this activity, you will be exploring the following research question:

Are humans able to match dogs to owners better than blind luck?

You will be asked to look at photos of six dogs and their owners. Your task is to try to match the dogs to their correct owners. After you make your matches, you will be shown which dogs belong to which owners, and you will be asked how many you matched correctly.

1. Write down your guesses and the correct answers.

| Owner | Dog | Correct? |
|---|---|---|
| 1. | | |
| 2. | | |
| 3. | | |
| 4. | | |
| 5. | | |
| 6. | | |

***Discuss the following question.***

2. How many of the six dogs did you correctly match with their owners?

When Albert Hoffman tried this, he matched four out of the six dogs to their correct owners. He was pretty pleased with his result and suggested that dogs do look like their owners.

3. Is it possible that someone could have four correct matches just by random chance—even if that person really had no ability whatsoever to match dogs with their owners? Explain.

4. For each of the potential number of correct matches listed below state whether you think that outcome is likely or unlikely if the person matching dogs to owners was completely guessing.

| |
|---|
| 0 |
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |

5. Explain why you think that some particular outcomes are more unlikely than others.

6. Based on your list in Question 4, does Albert Hoffman's result (four correct matches) convince you that he is able to match dogs to their owners better than blind luck? Or, do you think he was essentially guessing and had a lucky day? Explain your answer.

### *The Random Chance Model*

Consider the argument that dogs do not tend to look like their owners. If that argument is true, then you would expect that any correct matches of dogs to their owners was only because of random chance—not due to human skill in matching.

The good news is that this 'just by chance' process (or random chance) can be modeled using the same chance devices that you have been using in the course thus far. Under the assumption of 'just by chance', the process of matching dogs to owners can be modeled by randomly assigning dogs to owners. After the random assignment, it can be determined how many correct matches were obtained. This process can be repeated a large number of times to simulate the number of correct matches expected just because of random chance.

*Matching Dogs and Owners Under the Random Chance Model*

You will conduct a simulation study based on the 'just by chance' model. Write a brief description of a simulation study that you can perform that randomly pairs the six dogs and owners. Be sure that you explicitly identify each of the following in your description:

7. Write a brief description of the **model** that you will use to generate outcomes (what are the potential outcomes; sampling with or without replacement; probabilities of the potential outcomes; etc.).

Defining a single **trial** is also an important part of describing a simulation study. In this study, a trial represents each of the six owners that have been randomly assigned one of the six dogs. The trial ends when all owners and dogs have been paired.

8. Based on the description of a trial given above, identify the specific **result** that will be collected from each trial of the simulation.

## *Using TinkerPlots™ to Match Dogs and Owners Under the Random Chance Model*

To model the matching of dogs to owners under the random chance (i.e., 'just by chance') model, you need to produce simulated data from a model that randomly assigns six dogs to six owners.

- Use a `Mixer` to model the random assignment of dogs. Label the elements *1*, *2*, *3*, *4*, *5*, and *6*.

## *Random Assignment: Sampling without Replacement*

One aspect of the modeling process that does not quite map to the reality of people matching dogs to owners just by chance is that the TinkerPlots™ models used to this point may match the same dog to multiple owners. This is because sampling devices such as `Mixers` and `Spinners`, by default, randomly sample from the possible outcomes *with replacement*.

- Change the `Mixer` to sample *without replacement* (see instructions in the margin).
- Run the model three times (i.e., three trials) and fill in the table on the following page with the randomly assigned 'dogs' for each trial.

**Sampling Without Replacement**

- Click on the `Device Options` button below the sampling device menu.
- Select `Replacement > Without Replacement`.

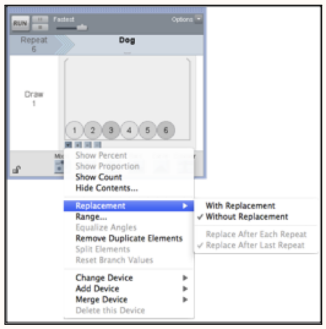



Figure 1: The `Replacement` option under the `Device Options` menu.

| | Dog | | |
|---|---|---|---|
| **Owner** | **Trial 1** | **Trial 2** | **Trial 3** |
| 1. | | | |
| 2. | | | |
| 3. | | | |
| 4. | | | |
| 5. | | | |
| 6. | | | |

9. How many dogs were correctly matched to their owners in each trial?

## *Creating Attributes Using Formulas*

One method for entering values into a newly created attribute, is to manually enter each value of the attribute into the case table. You used this method when you were first introduced to TinkerPlots™. Another method for entering values into a newly created attribute is to use a formula to mathematically manipulate attributes that already exist in the case table. (This is akin to using a formula to create a new variable in a spreadsheet program like *Excel*.)

TinkerPlots™ has many built-in formulas in the formula editor. These include computations that are commonly used in statistical work. There are also formulas included for working with text, lists, and many mathematical areas.

When you open the formula editor, the cursor is in the formula pane. To enter a formula, type in the formula pane, click buttons on the keypad, or use the attribute and function list included. Typically you will use a combination of these methods.

- Use the formula editor to create the attribute *Owner* (see instructions in the margin).
- Run another trial of the simulation by clicking the `Run` button in the `Sampler` to verify that the formula is working.

**Creating the *Owner* Attribute**

- Create a new attribute called *Owner* in the case table of outcomes from the sampler.
- Right-click the *Owner* attribute and select `Edit Formula`.
- Click the arrow next to `Special`.
- Double-click `caseIndex`.
- Click the `OK` button.

The `caseIndex` function simply added the row number for each row in the case table. Now, the values in the *Dog* attribute and the value in the *Owner* attribute can be compared. This could be accomplished by creating another attribute called *Match*. The *Match* attribute will indicate whether each of the six owners matched the dog that was randomly generated. For each row, if the corresponding outcome in both the *Owner* and *Dog* attributes was the same, then you could enter a *1*, for example, in that row's *Match* attribute. If the values of the *Owner* and *Dog* attributes were not the same, you could enter a *0* in that row's *Match* attribute. After plotting and separating the *Match* attribute, the number of ones and zeros could be

counted to determine how many dogs and owners in the trial were correctly matched.

This manual entry of the *Match* attribute, however, would be quite time consuming because you would have to re-enter the attribute values to indicate a match or non-match after each trial was generated. It is better to again use a formula and have TinkerPlots™ automate the computation of whether or not the values for the *Owner* and *Dog* attributes match. This can be accomplished using the `matchCount` function.

- Use the formula editor to create the attribute *Match* (see instructions in the margin).
- Generate another trial of the simulation by clicking the `Run` button in the `Sampler` to verify that the formula is correctly identifying matches and non-matches.

**Creating the *Match* Attribute**

- Create a new attribute called *Match* in the case table of outcomes from the sampler.
- Right-click the attribute named *Match* and select `Edit Formula` from the drop-down menu.
- Click the arrow next to `Function` to display the types of built in functions.
- Click the arrow next to `List` to display the built-in functions to work with lists.
- Double-click the `matchCount` function.
- Click the arrow next to `Attributes` to display the attribute names in the case table.
- Double-click the attribute *Dog*.
- Type a comma
- Double-click the attribute *Owner*.


- Click `OK`.

**Working with the Formula Editor**

There are two methods to work with attributes and functions in the formula pane. You can select attributes and functions from the lists as you have done when selecting `matchCount` and `caseIndex`. You can also type the name of the function directly into the formula pane followed by a set of parentheses. When typing the names of formulas and attributes, if TinkerPlots™ recognizes the formula, the formula name will turn blue. If TinkerPlots™ recognizes an attribute name, the name will turn purple.

### *Plotting and Collecting Results*

10. Create a plot of a trial of the simulation by plotting the outcomes of the *Match* attribute. Sketch this plot below.

- Use TinkerPlots™ to summarize the number of correct matches in the trial by using `Count (N)`.
- Collect the number of correct matches.
- Carry out 99 more trials (100 total trials) of the simulation.
- Plot the results

11. Sketch the plot of the simulation results below.

### *Evaluating the Observed Result*

The plot depicts the number of matches one gets if the six dogs and owners are paired by random chance. As can be seen in the plot, some results are more likely than other results. Remember that these are the results that are *expected* under the model of 'just by chance', also known as the *null model*. The plot is also referred to as the *null distribution* because it shows the variation in the results under the null model.

Any observed result can be examined in relation to the null distribution. The observed result acts as evidence. If the observed result is one that is unlikely under the null model (i.e., an unlikely result in the null distribution) then it represents evidence against the null model. If, on the other hand, the observed result is a likely outcome in the null distribution, then it does not provide evidence against the null model.

12. Based on the results expected under the null model (i.e., the null distribution), for each outcome listed below, state whether that outcome is likely or unlikely.

| |
|---|
| 0 |
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |

13. Based on the likeliness for each of the results expected under the null model that you just recorded, do you think that Albert Hoffman's result of four correct matches out of six is likely or unlikely? Explain.

14. Does Albert Hoffman's result provide evidence that supports or does not support the 'just by chance' model? Explain.

***Now suppose Albert Hoffman's friend, Friedrich Sertürner, tried matching the dogs to their owners and correctly matched two out of six dogs to their owners.***

15. Do you think Friedrich's result is likely or unlikely under the 'just by chance' model? Explain.

16. Does Friedrich's result provide evidence that supports or does not support the 'just by chance' model? Explain.

17. Do you think **your** result from the initial dog–owner matching exercise is likely or unlikely under the 'just by chance' model? Explain.

18. Based on your answer to the last question, does the evidence support or not support the 'just by chance' model?

19. Given your answer to the last question, what does the evidence suggest about whether you have an ability to match dogs to their owners? Explain.

20. Suppose someone matches all six of the dogs to their owners correctly. Is such a result likely under the 'just by chance' model? Explain.

21. Suppose someone were to match zero dogs to their owners correctly. Is this a likely result under the 'just by chance' model? Explain.

# Modeling the Matching Dogs to Owners Problem

**SPECIFY A MODEL**

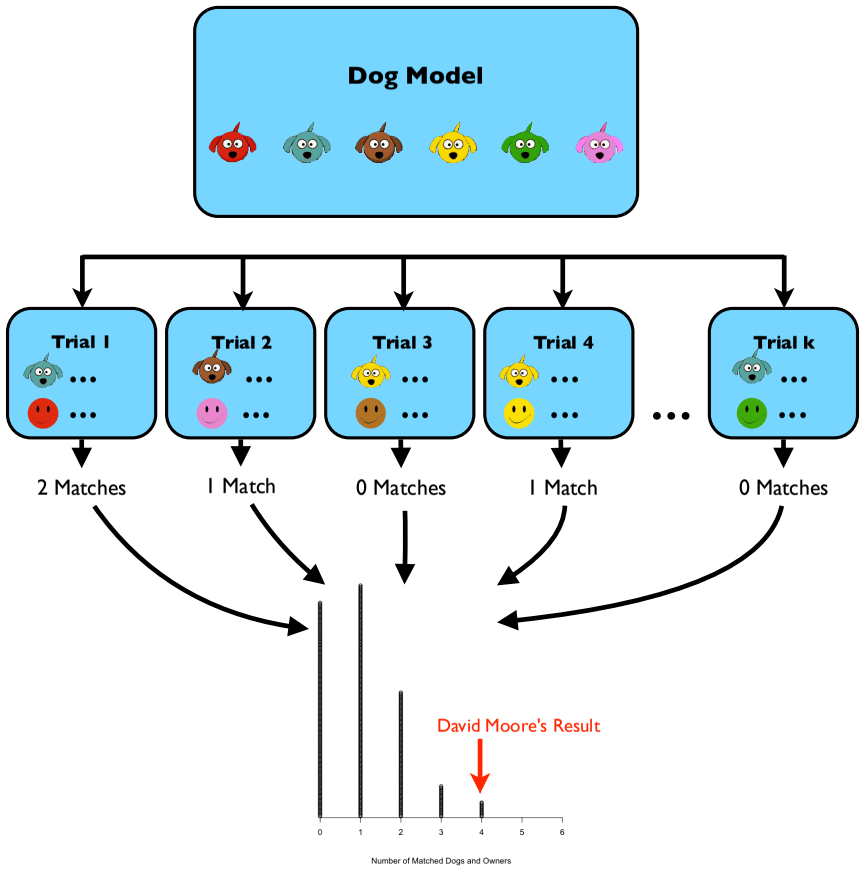
The dog model is such that a sequence of six dogs are chosen *without* replacement. Each is equally likely to begin with.

**RANDOMIZE & REPEAT**

A trial ends when a six dogs and have been randomly distributed. For each trial, compute the number of dogs that match their owners (e.g., the row number). Generate many trials.

**EVALUATE**

Compile all of the trial results into a single distribution, Evaluate the initial question by determining whether the observed result of four out of six matched is likely or unlikely under the blind-guessing (or chance) model.




*Level 1*:
**Model**

*Level 2*:
**Samples and Numerical Summary Measures**

*Level 3*:
**Distribution of the numerical summary measures**

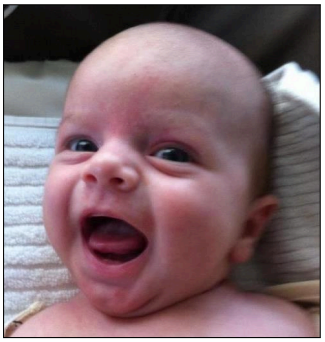
# *Helper or Hinderer Readiness*

During the next class activity, you will examine results from an experiment published in *Nature* on the social evaluation of infants. To prepare for this, we would like you to read the abstract of the article so that you have an understanding of the study that was carried out.

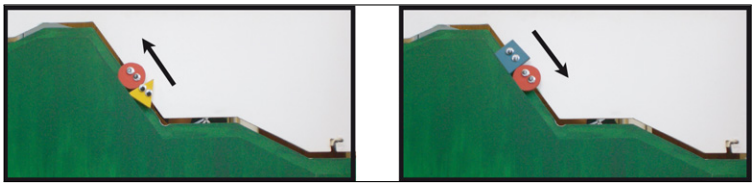
- Read the abstract of the article *Social Evaluation by Preverbal Infants*. The abstract is available at `http://www.nature.com/nature/journal/v450/n7169/full/nature06288.html`.

# *Course Activity: Helper or Hinderer*

Most college students recognize the difference between naughty and nice, right? What about children less than a year old—do they recognize the difference and show a preference for nice over naughty? In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying for the foundation for social interaction. In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were alternately shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character (*helper*) and one where the climber was pushed back down the hill by another character (*hinderer*).



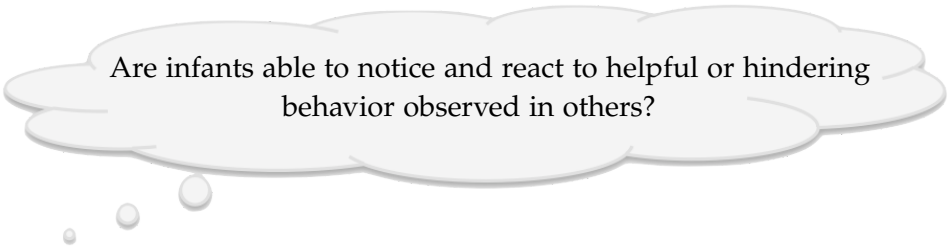
J. K. Hamlin, K. Wynn, & P. Bloom. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559.



Helping and hindering habituation events. On each trial, the climber (red circle) attempts to climb the hill twice, each time falling back to the bottom of the hill. On the third attempt, the climber is either bumped up the hill by the helper (left panel) or bumped down the hill by the hinderer (right panel). http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html

The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The researchers found that 14 of the 16 infants chose the helper over the hinderer.

In this activity, you will be exploring the following research question:




***Discuss the Following Questions***

1. What proportion of these infants chose the helper toy?

2. What does that suggest about the answer to the research question? Explain

Suppose for the moment that the researchers' conjecture is wrong, and infants *do not* really show any preference for either type of toy. In other words, infants just randomly pick one toy or the other, without any regard for whether it was the helper toy or the hinderer. This, remember, is the model based on random chance—the 'just by chance' model.

3. If this is really the case (that infants show no preference between the helper and hinderer), is it possible that 14 out of 16 infants could have chosen the helper toy just by chance?

4. Would the observed result (14 of 16 choosing the helper) be a likely result if infants had no real preference, or would it be an unlikely result? How strong do you believe the evidence is against the 'just by chance' model if 14 of 16 infants were found to choose the helper?

### *Selecting the Helper or Hinderer Under the Random Chance Model*

Similar to the process you followed in the *Matching Dogs to Owners* activity, the key to answering the research question in this activity is to determine the likelihood of the observed result (14 of 16 infants choosing the helper) under the assumption that infants have no preference for either the helper or the hinderer. The 'no preference' model is again the 'just by chance' model—infants randomly select either the helper or hinderer. To find out this likelihood, you will model the process of 16 hypothetical infants making their selections using random chance. Then, you can count how many of these "infants" choose the helper toy. This process can be repeated many times to obtain a distribution of results that would be expected under the 'no preference' or 'just by chance' model. The observed result of 14 of 16 infants choosing the helper can then be evaluated in light of this distribution to determine how likely it would be to obtain such a result (or a more extreme result) under the assumption of random chance. As such, the observed result can provide evidence to help answer the research question.

5. Write a brief description of the **model** that you will use to generate outcomes (what are the potential outcomes; sampling with or without replacement; probabilities of the potential outcomes; etc.).

In this study, a trial represents each of the 16 infants choosing a toy. The trial ends when 16 toys have been chosen randomly.

6. Based on the description of a trial given above, identify the specific **result** that will be collected from each trial of the simulation.

## *Carrying Out the Simulation*

- Based on your description of the model, set up the model in TinkerPlots™.
- Carry out a single trial of the simulation in TinkerPlots™.
- Plot the outcomes from the trial.
- Collect the appropriate summary measure.
- Carry out 99 more trials (100 trials total) of the simulation in TinkerPlots™.

### *Evaluating the Observed Result*

- Plot the results from the simulation.

7. Sketch a plot of the results below.

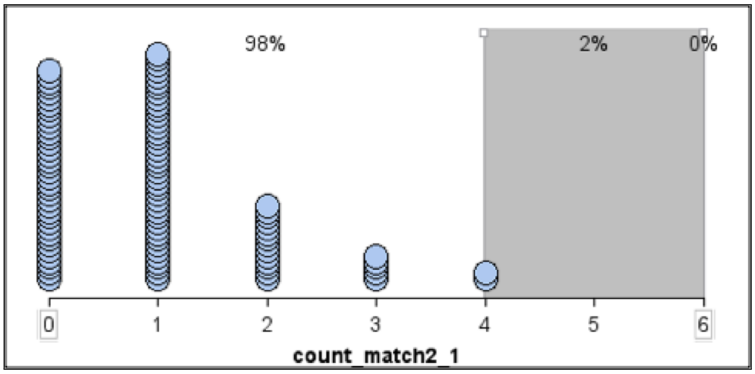
8. Is the observed result from the original experiment likely or unlikely under the hypothesized model? Explain.

9. What does this suggest about the hypothesized model? Explain.

### *Evidence Against the Model*

Remember that the observed result is the evidence you have to support a hypothesized model. Sometimes it constitutes a great deal of evidence for the hypothesized model—when the observed result is very likely—while other times it constitutes very little evidence for the hypothesized model—when the observed result is unlikely under the hypothesized model. To help other people understand how likely or unlikely the observed result is under a particular model, it is typical to quantify the level of evidence we have to support that model.

The level of evidence to support a model is quantified by answering the question: What proportion of the simulated results indicate at least as much evidence as the observed result. For example, consider the results from the matching dogs to owners simulation shown in Figure 1. The grey area shown in the plot indicates the level of evidence to support the hypothesized model given Albert Hoffman's result of four correct matches.




Simulation results obtained based on the hypothesized model. The grey area indicates the level of evidence to support the hypothesized model.

The results of five matches and six matches are included because if someone considers the observed result of four matches as evidence to support the hypothesized model, then the results of five matches must also be evidence to support the hypothesized model, as must the result of six matches.

The level of evidence to support the model can also be quantified. In this example, the level of evidence to support the hypothesized model based on the observed result of four matches is 0.02. This tells us that the observed result of four matches is in the most extreme two percent of the simulated results. This is quite an unlikely result given the hypothesized model and therefore provides evidence *against* the hypothesized model.

10. Quantify the level of evidence to support the model for the observed result of 14 out of 16 infants choosing the helper toy based on the hypothesized model of no difference (see instructions in margin).

11. Given the level of evidence you just computed, what does

**Using the `Divider` Tool to Quantify the Level of Evidence to Support a Model**

- Highlight the plot of the results,
- Click on the `divider` tool in the upper plot toolbar.


- Move the divider by dragging the endlines so that the grey area extends from the observed result to the result that provides the most evidence against the hypothesized model. In this example, the grey area would extend from 14 to 16.
- Click the `Counts (%)` button.

this suggest about the hypothesized model? Explain.

12. What does this suggest about the research question? Is it likely that infants are making their selections based only on random chance? Explain.

# Helper or Hinderer Simulation

**SPECIFY A MODEL**

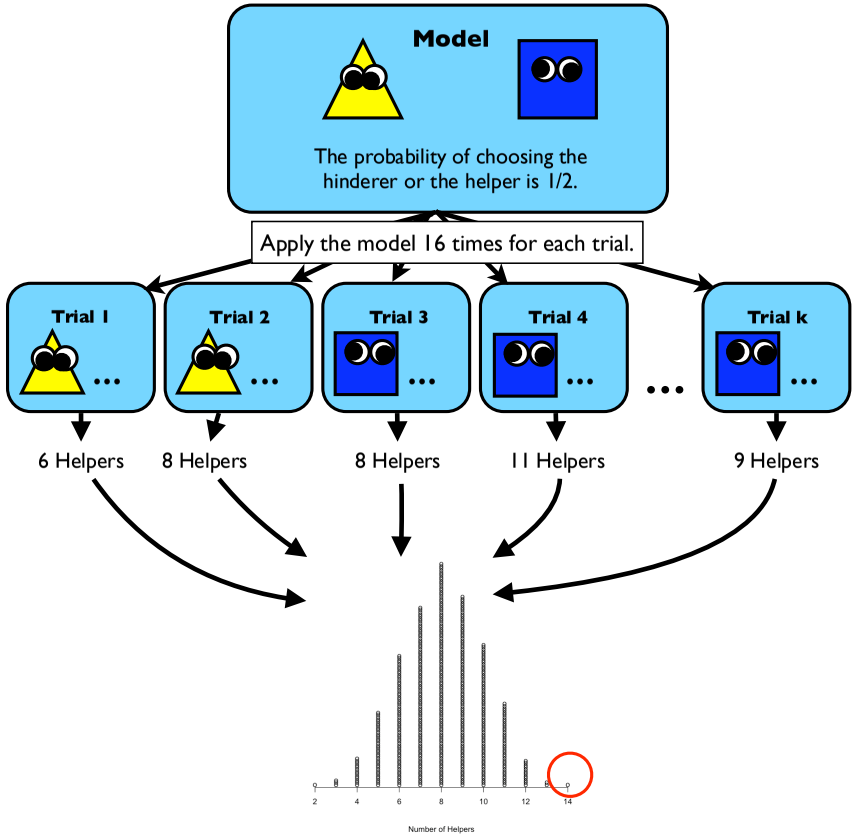
The model is such that there are two outcomes–helper or hinderer. Each is equally likely.

**RANDOMIZE & REPEAT**

A trial ends when 16 outcomes have been randomly generated–the model is applied once for each of the 16 subjects. For each trial, count the number of "infants" that chose the helper. Generate many trials.

**EVALUATE**

Compile all of the numerical summary measures into a single distribution, Evaluate the initial questions by finding the proportion of trials in which there were 14 or more helpers (14, 15, or 16) using this distribution.




*Level 1*:
**Model**

*Level 2*:
**Samples and Numerical Summary Measures**

*Level 3*:
**Distribution of the numerical summary measures**

# *Learning Goals: Unit 1*

The activities, homework, and reading that you have completed for the first part of this course have introduced you to several fundamental ideas in the discipline of statistics including randomness, modeling, and simulation. You have also been introduced to the key elements that are used by researchers and scientists for examining and testing conjectures and hypotheses. In addition, you have learned how to use TinkerPlots™ to model probabilistic events and generate simulated data under a variety of conditions.

These ideas and skills are crucial for your success in the remainder of this course, as they will be extended and built on in the upcoming material. With that in mind, presented here is a list of some key concepts and skills that you should have a good handle on.

## *Literacy/Understanding (Terms and Concepts)*

- Understand that human intuitions about randomness/probability may be faulty
- Understand that randomness/probability cannot be outguessed in the short term but patterns can be observed over the long term.
- Understand that simulation can be used to investigate probabilistic outcomes and model chance events
- Understand that simulation can be used to determine

whether a particular result is likely to have happened *just by chance*

- Understand the *Model–Simulate–Evaluate* framework for the simulation process (see the visuals)
- Understand that different chance models lead to different simulation results (coins vs. dice)
- Understand that there are predictable patterns/characteristics of simulation results based on repeatedly sampling/generating random data (e.g., a bell shape from a graph of sample statistics)

### *Selecting/Using Models*

- Set up a model to generate outcomes from random devices (see for example, *Modeling Random Behavior*)
- Set up a model real-life phenomena (see for example, *One Son*, *Matching Dogs*, and *Helper or Hinderer*)
- Generate simulated data in which each trial depends on a stopping rule (see for example, *One Son*)
- Translate between elements of the simulation process and the real world

### *TinkerPlots™ Skills*

You should also be able to do the following using TinkerPlots™.

- Create a new case table and enter data into the case table
- Use a formula to create a new attribute in a case table
- Plot the values from an attribute in a case table
- Separate plotted values
- Create a new sampler and use different devices to generate random outcomes (e.g., `Spinner`, `Mixer`, `Counter`)

- Numerically summarize (e.g., compute a measure) the randomly generated outcomes from the trial (e.g., `Case Count`, `Case Percentage`, `Mean`, `Median`)
- Collect the numerical summary measure from many trials

In the next activity, *Unit 1 Wrap-Up & Review*, you will have a chance to assess yourself on whether or not you have mastered these ideas through a variety of practice and extension problems. As a pre-cursor to this activity, you may want to review the readings and activities in Unit 1.

# *Course Activity: Unit 1 Wrap-up & Review*

## *Terminology for Unit 1*

1. At this point, you should be familiar with the following terms. Write down what each term represents as well as any notes that may help you remember.

   (a) Model

   (b) Trial

   (c) Observed result

   (d) Random chance ('Just by Chance') model

   (e) Null model

### *Modeling Random Behavior*

2. Suppose that you receive about ten text messages per day and about half of all text messages come from your mother. Describe how you could use each of these devices (coins and dice) to simulate the number of text messages that you receive from your mother per day.

   (a) Coins

   (b) Dice

### *Matching Dogs to Owners*

3. Now consider analyzing a study with eight dogs and owners. Use TinkerPlots™ to simulate 100 trials under the random chance model. How many correct matches would a person have to make in order to convince you that there is evidence against the random chance model? Explain.

### *Helper or Hinderer*

Suppose that the observed result from the Helper or Hinderer research study had been ten of the 16 infants choosing the helper toy (rather than 14 of 16).

4. Explain why you can obtain the strength of evidence of this new observed result using the same simulation analysis that you already conducted.

5. Based on your simulation analysis, quantify the strength of evidence for this new observed result.

6. What conclusion would you draw about the null model? Explain your reasoning process behind this conclusion.

7. If the new observed result had been 13 of 16 choosing the helper toy, quantify the strength of evidence and draw a conclusion about the null model.

Suppose that the study had involved only eight infants rather than 16, and that seven of the eight infants had chosen the helper toy.

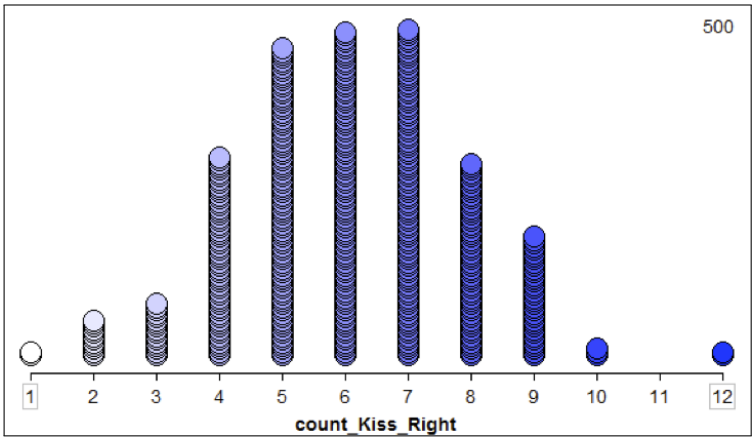
8. Explain how you would modify the simulation analysis for this new study.

9. Would you expect the result from this new study to constitute more, less, or the same amount of evidence that infants really prefer the helper toy, as compared to the original study result (14 out of 16 prefer the helper toy)? Explain your answer.

10. Carry out the simulation analysis and comment on whether your expectation was correct based on the results

## *Further Practice*

A recent study investigating the question, of whether more than half of all kissing couples lean their heads to the right, found that eight of 12 kissing couples leaned their heads to the right. In order to answer the research question, the null model of 'no preference' was used to simulate data for 500 trials. In each trial, the number of 'couples' who lean their heads to the right was collected. The plot of these results is shown below.




Simulation results based on 500 trials.

11. How might you have set up a model in TinkerPlots™ to obtain these results?

12. Use the plot above to answer the research question. Be sure to provide a quantification of the strength of evidence to support your answer.

# *Unit II: Comparing Groups*

> *The nature of doing science, be it natural or social, inevitably calls for comparison. Statistical methods are at the heart of such comparison, for they not only help us gain understanding of the world around us but often define how our research is to be carried out.*
>
> —T. F. Liao (2002)

Drawing inferences and conclusions about the differences among groups is an almost daily occurrence in the lives of most people. In any given hour of any given day, television, radio and social media abound with comparisons. For example, data scientists at *OKCupid*, an online dating site, examined whether frequent tweeters (users of Twitter) have shorter real-life relationships than others.

The website OKTrends includes an answer to this question, as well as many others.

Group comparisons are at the heart of many other interesting questions addressed by psychologists, physicians, scientists, teachers, and engineers. Aside from questions of differences, group comparisons address questions about efficacy, such as "Is a particular curriculum effective in improving students' achievement?" or "Is having a Facebook page for a business effective in improving sales?"

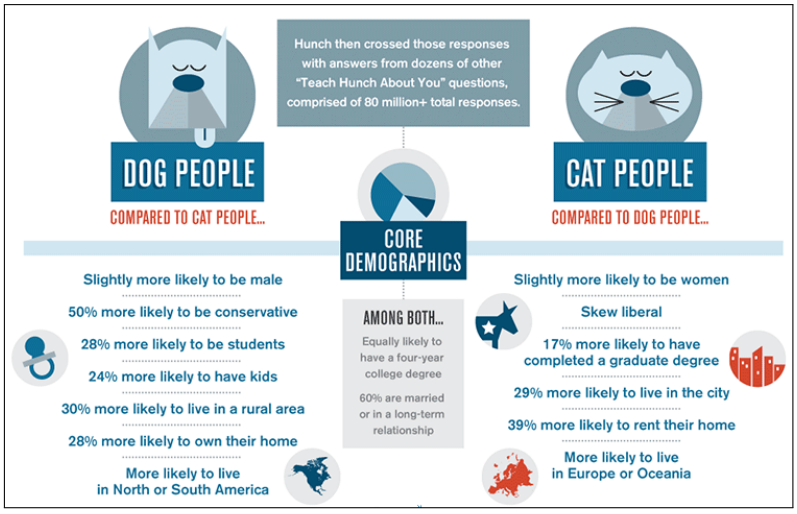
They also address questions about magnitude such as "How much lower are attendance rates for a particular population of students?" or "How much more likely is it that an iPhone user will live in the city (rather than the suburbs or the country) than an Android user?"

This question was asked in a survey by *Hunch*, a personal recommendation service. They created an infographic that highlighted several comparisons of the average demographic characteristics between the two users.

## *Comparisons in the Media*

The media often proffer group comparisons. Consider the infographic displayed in the figure below. It shows a comparison, again based on survey data from *Hunch*, between cat and dog people.

The entire infographic can be seen at Hunch.com.



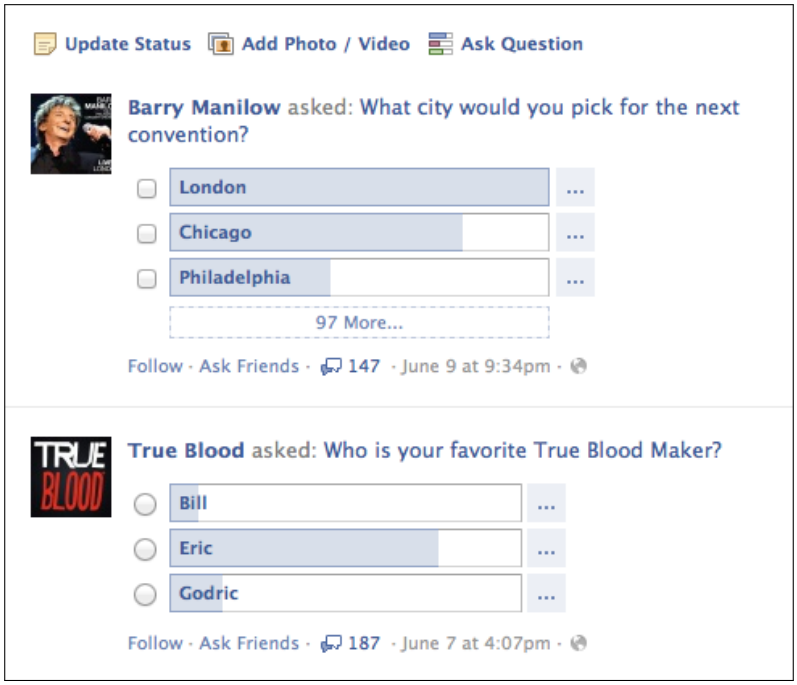

*Hunch* infographic comparing dog and cat people.

The comparison between cat and dog people is primarily conveyed via text. The graphical parts of the infographic are mostly aesthetic.

The figure presented on the next page shows a screenshot from a Facebook application called *Facebook Questions*. The application allows any Facebook user who has installed it to poll other Facebook users. It displays the poll results as a bar graph. Hovering over a particular bar will display the frequency or count of users voting for a particular response.

In contrast to the comparison in the *Hunch* infographic, the comparison here is conveyed through the graph rather than the text. The length of each bar provides an indication of the frequency of voters who responded to each of the poll options relative to the other options. For example, the bar labeled 'Eric' in the True Blood poll is roughly five times longer than the bar labeled 'Goderic'. This corresponds to the five-fold difference in the frequency of votes (3,571 votes for Goderic and 18,102 votes for Eric) between the two candidates. The bar labeled

'Bill' is roughly $\frac{1}{2}$ the length of the bar labeled 'Goderic' and about $\frac{1}{9}$ as long as the bar labeled 'Eric', indicative of the 1,933 votes that had been cast for Bill.




Poll questions and bar graphs displaying the voting results for two questions submitted using the *Facebook Questions* application.

Examining the two graphical displays presented, reflect on the following questions.

- Which of these two displays more clearly conveys the comparisons being made? Why?
- How would you improve on each display? Explain.

## *Statistical Comparisons*

Statisticians have made many contributions to the methodology used by researchers and scientists in making group comparisons. One well-known statistician, Roger Kirk, has suggested that research questions regarding group differences should address three important questions: (a) Is an observed difference real or should it be attributed to chance? (b) If the difference is real, how large is it? and (c) Is the difference large enough to be useful?

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213–218.

Each of these three questions can be addressed through statistical inference. Recall that statistical inference is the process through which we can deal with uncertainty. Consider the following research questions posed by the *Sunlight Foundation*: Has the complexity of speeches given in the United States congress changed over the last 10 years? Is the complexity of speeches given by Democrats different than that of speeches given by Republicans?

### *Operationalization*

In thinking about answering these questions, there are several decisions that need to be made by the researcher. One major decision that would have to be made is how "complexity" is to be measured. This is referred to as operationalizing, and it often involves defining and quantifying a fuzzy construct to help make it understandable (and studyable). In general, making a comparison between numeric properties such as three apples and five apples is much easier than making comparisons of an ill-defined characteristic such as "complexity". In considering how to operationalize a construct, there are often many possibilities. For example, complexity of a written passage can be operationalized by using a readability index such as the Flesch–Kincaid score. Another way of operationalizing the passage is to count the number of words in the passage that appear in Kaplan's list of *The 100 Most Common SAT Words*.

See Kaplan's entire list of The 100 Most Common SAT Words.

It is important to understand the advantages and limitations in the choice of an operationalization. For example, by using the frequency of the 100 most common SAT words, it is easy to provide a quantification for each passage. It is also easy to understand what that quantification means—a passage with a score of 4 uses two fewer SAT words than a passage that has a score of 6. This will ease interpretations during an analysis. On the other hand, this score is a rather rudimentary measure of complexity.

In comparison, the Flesch–Kincaid score is a less rudimentary measure of sentence complexity defined as,

$$\begin{aligned} FK =& 0.39 \times (\text{average words per sentence}) + \\ & 11.8 \times (\text{average syllables per word}) - 15.59. \end{aligned}$$

The score indicates that the text would be at the limit of comprehension for a person with the equivalent of that number of years of education. In more familiar words, the score gives the reading grade level of the passage.

For example, a passage with a Flesch-Kincaid score of 12 indicates that a person with 12 years of education would answer 50 per cent of the questions correctly.

At its core, the Flesch–Kincaid score equates higher grade levels with longer words and longer sentences. It does not, however, indicate anything about the clarity or correctness of a passage of text. If these attributes are important in operationalizing passage complexity, other measures should be used in the quantification. Another strength of the Flesch–Kincaid score is that it has been adopted as the Department of Defense standard in determining the age-appropriateness of reading material. Because of this, the metric is one of the best known and frequently used metrics of readability.

Many famous passages have been scored using the Flesch–Kincaid measure such as the U.S. Constitution (17.8 grade level), the Declaration of Independence (15.1 grade level), the Gettysburg Address (11.2 grade level), Martin Luther King's "I Have a Dream" speech (9.4 grade level), and U2 frontman Bono's 2004 commencement speech at the University of Pennsylvania (5.9 grade level).

The operationalization choice is subjective to the researcher and inevitably has a great deal of impact on the generalizations that the researcher can make. Does the operationalization accurately reflect the construct that the researcher wishes to say something about? For example, say that after making comparisons of congressional speeches using the Flesch–Kincaid score you find that more current speeches have, in general, lower scores than speeches from 10 years ago. This might suggest that the complexity of today's congressional speeches is less than it was 10 years ago. Another researcher might choose a different method of operationalizing complexity and come to a different conclusion. What does the choice of operationalization reflect about the "truth" of the change (or non-change) of complexity in congressional speeches? In science, these types of questions are related to the validity of the inferences that can be made.

Many books on research methods deal extensively with ideas of validity. The Research Methods Knowledge Base provides more information for the interested reader.

## *Summarization*

After deciding upon an operationalization, it may seem that a researcher could just collect her data and answer her research questions. However, there are still a litany of choices that the researcher needs to make before she gets to this point. For example, again consider the two research questions posed previously (Has the complexity of speeches given in the United States congress changed over the last 10 years? Is the complexity of speeches given by Democrats different than that of speeches given by Republicans?).

One of the initial questions is whether the researcher should use raw data or summary data. Raw data would be data that has not been processed or manipulated in any way (i.e., the original data). In this case, it would mean using the transcripts of the speeches given in Congress. Summary data, on the other hand, would refer to data that has been manipulated or summarized in some manner. For example, rather than obtaining a transcript of the actual speeches given, the researcher could instead use data that some other party or individual has already summarized and produced such as the average word per sentence or even the Flesch–Kincaid scores themselves. Most researchers would prefer to work from the raw data and compute these measures themselves, but in many cases the raw data is not available and only summary information is accessible.

As an example, the Minnesota Department of Education is required to provide test results, revenue and expenditure data, and demographic information for each school and district in the state. Because of privacy laws, these data are released to the public at the school-level (e.g., summarized as school averages) rather than at the individual-level (i.e., the raw data used to compute these avaerages). See the MDE website for reports and publically available data for all Minnesota schools and districts.

In the research proposed, It is in fact possible for the researcher to obtain the raw data—transcripts for each of the Congressional speeches given in the last 10 years.  is a verbatim account of the remarks made by senators and representatives while they are on the floor of the Senate and the House of Representatives. Transcripts for each speech could be collected from this or another website. Again, the collection of these transcripts may seem straight-forward (just download them from *The Congressional Record*). However, just a minute or two of examining the website leads to many other questions. For example, after searching on 'Amy Klobuchar', one of the Senators from Minnesota, it was revealed that the search results brought up any mention of Amy Klobuchar—not just

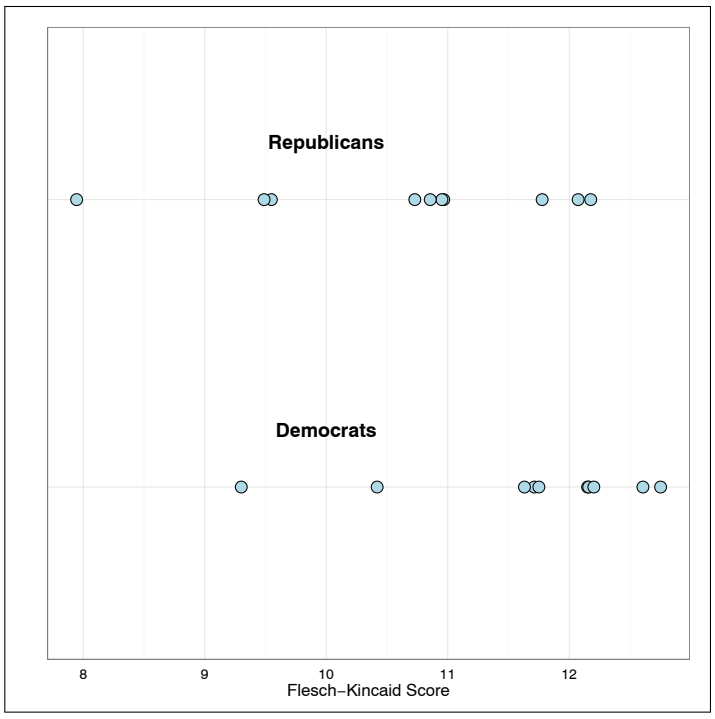
Read more at *The Congressional Record*.

the speeches she gave. Should we use all of these? Or just the speeches? Another decision.

It was also apparent that collecting these data would be very time consuming. There were 219 search hits for Amy Klobuchar in the 112th Congress (2011–2012) alone. Given there are 541 members of Congress in any given year (435 voting Representatives, 6 non-voting Representatives, and 100 Senators), and we would need 10 years worth of data, this is somewhere on the magnitude of 1,082,000 transcripts that we would need to download. (Not to mention that we would also have to compute the Flesch–Kincaid score for each one!). Because of this, many researchers collect only a subset of the data available. This is referred to as sampling and there are well-defined methods for choosing a sample of data to ensure that it adequately represents the entire population. Ideas of sampling will be introduced in Unit 3.

Some of the downloading of documents and processing (e.g., counting syllables, words per sentence, etc.) can be automated using computer languages such as Python or Perl. This coud still take several weeks to write the web scraping program to obtain the data and check the resulting output for errors.

After all of the decisions about operationalization and data collection have been made, and the transcripts have been collected and quantified with a Flesch–Kincaid score we can start to think about analyzing the data to answer the research questions. This leads to more questions and decisions that need to be made. How will you compare the Flesch–Kincaid scores across years? Between Democrats and Republicans? Consider the plot (on the next page) of Flesch–Kincaid scores for ten speeches given by Democrats and ten speeches given by Republicans.




Flesch–Kincaid scores for ten speeches given by Democrats and ten speeches given by Republicans.

Using the plot, reflect on how you might answer the following questions.

- How would you compare the Flesch–Kincaid scores for the speeches given by Democrats and the speeches made by Republicans?
- If you had to summarize the Flesch–Kincaid scores for the speeches given by Democrats by using a single number to quantify all ten speeches, what would this number be? Explain.
- Are you comfortable summarizing the ten scores into a single number? Explain. If not, are there other numbers you would use or add to your original quantification? Explain.
- Using the same methodology, summarize the Flesch–Kincaid scores for the speeches given by Republicans.
- Compare the two quantifications you have just made. Summarize the differences between the Democrats and Republicans based on the two quantifications into a single number

(i.e., how much bigger/smaller is the Democrats score than the Republicans?)

- What does this comparison suggest about the original research question as to whether the complexity of speeches given by Democrats are different than that of speeches given by Republicans? Explain.

### *To Infer or Not to Infer*

It may seem that after all of this work we finally have an answer to our research questions, but it turns out that we may not. This is because we have introduced some error through the process of sampling. Let's say that we had sampled 200 speeches (100 speeches given in 2002 and 100 given in 2012). Further consider that we found the average grade-level for the 100 speeches given in 2002 was 11.5 and the average grade-level for speeches given in 2012 was 10.6. This suggests that Congressional speeches currently are almost an entire grade-level lower than they were 10 years ago.

But, what if we had gotten different speeches in our sample? There are many different samples of 100 speeches from 2002 and 100 speeches from 2012. Each of these different samples might give a slightly different value for the average grade-levels for those years, which in turn affects how different we claim these grade-levels are! It is well-known that different samples give different results (e.g., averages). A very big question that we will address in this unit is 'How different can the results be just because of obtaining a different sample?' By answering this question, we can address and quantify how uncertain our observed results are. For example, we could find, based on different samples, that the difference in grade-level between speeches given in 2002 and 2012 may be as high as 1.3 or as low as 0.5.

The ideas regarding the quantification of uncertainty of results because of sampling relate directly to ideas of statistical inference. Statistical inference is the ability to draw conclusions from sample data. Consider if the range of uncertainty is very

large, such as the difference in grade-level between speeches given in 2002 and 2012 may be as high as 2.1 or as low as -0.2 ( a negative value here indicates that the grade-level for speeches given in 2012 is actually higher than that for speeches given in 2002). This much uncertainty makes it very difficult to address how different the complexity in speeches given in 2002 and 2012, and in fact, whether there is a difference at all.

Quantifying uncertainty is a large part of statistical inference, but it is not the only thing involved in drawing conclusions from sample data. Drawing conclusions from data also implies thinking broadly about the scope of those inferences. For example, does a difference in complexity only refer to a comparison of speeches given in 2002 and 2012? Or does it apply more broadly to the complexity of all Congressional speeches ever given? Or, can we use these data to suggest that speeches in general—Congressional, presidential, commencement, etc.—have changed in complexity over time?

Another piece of drawing conclusions is in considering the 'why'. Why has the complexity of Congressional speeches changed over time? Is this an indictment of Congress? the American public? or neither? Ascribing meaning to why differences occur is a part of the much broader reasons for carrying out research. Unfortunately, the attribution of the reasons or causes of a particular difference are very difficult under most circumstances.

In this unit, you will learn how to quantify the uncertainty associated with the results from a sample and how that uncertainty is expressed in the reporting of research. You will also learn about how particular statistical methods influence the scope of inferences and the attribution of cause that a researcher can make. Understanding these ideas is important in evaluating any given result presented in any research that has used statistical methods.

## *Outline of the Unit*

In this unit, you will begin exploring ideas related to the statistical comparison of groups. After the first activity, which provides you with an overview of the entire process of statistical testing, you will explore these ideas in deeper detail.

After exploring some of the characteristics of distributions that allow for statistical comparisons, you will be introduced to the randomization test. Through learning this method for statistical inference, you will be exposed to the fundamental ideas that are included in all analyses involving statistical inference. You will also learn how to carry out a randomization test using the TinkerPlots™ software.

The ideas of the randomization test will be reinforced throughout the remainder of the unit as will its utility, as you explore data in which the outcomes are both quantitative and categorical. You will also examine data that have been collected under many different conditions, some experimental and others observational.

As you encounter data collected under these various conditions, you will learn how study design and data collection directly affects the scope of inferences that are appropriate. Throughout the unit, you will continue to add to the knowledge you accumulated from Unit 1. This includes both content knowledge and TinkerPlots™ knowledge.

# *America's Most Reliable Airlines*

Based on excerpts from an article by Rebecca Ruiz in the October 1, 2008 *Forbes Magazine*.

According to our analysis of the nation's 10 major airlines, discount carriers actually rank first in reliability.

Southwest, the no-frills discount carrier, handily beat the competition in most of the categories we judged. JetBlue also considered a discount airline despite its plush leather seats and individual television sets, ranked third just behind Continental Airlines. Fourth place went to AirTran, another budget carrier.

Alaska Airlines, American Airlines, and Delta Air Lines were solidly average performers. United Airlines and US Airways landed at the bottom of the list.

To judge reliability in the airline industry, particularly at a time when carriers are responding to oil prices by slashing capacity and raising prices, we looked at six different factors for 10 major airlines.

We collected five years' worth of data relating to on-time arrival, cancellations, complaints and mishandled baggage from the Aviation Consumer Protection Division of the Department of Transportation. Delays and cancellations, the factors most likely to ruin a flier's day, were given double weight.

To better gauge the overall flying experience, we included *J.D. Power and Associates'* consumer satisfaction rankings from 2005 to 2008. These surveys reach more than 9,000 travelers annually and ask participants to rate factors like cost and fees, in-flight services and check-in.

When all of these figures were combined, the discount airlines consistently rose to the top. For each of the years we studied, Southwest's flights were punctual more than 80% of the time; the average was 76.8%. Alaska Airlines gave the most dismal performance, with only 74.6% on-time flights.

In terms of canceled flights, Southwest reigns yet again. The carrier canceled an average of 0.65% of its flights over the five-year period, compared with the worst airline, American, which canceled an average of 2.4%.

AirTran, another budget carrier, had the fewest reports of mishandled baggage—a contentious issue now that airlines are charging as much as $50 to check regular-sized luggage. In 2007, AirTran had about four reports of mishandled baggage per 1,000 customers. The worst-ranking airline, US Airways, had 8.5.

***Be ready to share and discuss your responses to each of the following questions with your group.***

1. What has been your own experience with feeling an airline is reliable or unreliable?
2. What were some of the factors that *Forbes Magazine* considered about each airline as a part of their reliability rankings for the discount air carriers?
3. Are there other things you might want to consider when judging an airline's reliability?
4. Would you consider all factors equally or would you rate some factors higher than others? Explain.

# *Course Activity: Comparing Airlines*

***Share and discuss your responses to each of the following questions with your group.***

1. What has been your own experience with feeling an airline is reliable or unreliable?

2. What were some of the factors that *Forbes Magazine* considered about each airline as a part of their reliability rankings for the discount air carriers?

3. Are there other things you might want to consider when judging an airline's reliability?

4. Would you consider all factors equally or would you rate some factors higher than others? Explain.

## *Group Task*

*Chicago Magazine* wants to write a story about the experiences of travelers who fly from Chicago to Minneapolis. They have heard complaints about arrival delays for each of the airlines that depart from Chicago and fly to Minneapolis. The magazine wants to decide whether the airlines can be considered to be equally unreliable or if any airline is doing a better job in getting their passengers to their destinations on time. The editor-in-chief of the magazine has contacted your group to help provide information for their article. She wants to focus on two regional airlines, Mesa and American Eagle, that fly out of Chicago to Minneapolis. The editor has obtained flight arrival delay time data, based on flights that departed from Chicago and flew to Minneapolis in 2008, for your group to analyze in order to address the following questions:

- Is there a difference in the reliability as measured by arrival time delays for these two regional airlines out of Chicago to Minneapolis? Or are both airlines pretty much the same in terms of their arrival time delays?
- Are any differences you find large enough to influence travelers so that they are advised to choose one airline over the other (all other factors, like cost, being equal)?

### *Explore and Describe*

Examine the arrival time delays for the two airlines using the first set of data. The data can include negative numbers—which indicate the number of minutes that an airline arrived earlier than its scheduled arrival time.

Come up with *at least 3 numerical measures* that can be used to measure and compare airlines' reliability based on these data. Compute these measures for each airline and describe in words what each of your measures would indicate about an airline's reliability. These descriptions should be written so that a traveler can easily understand what they are measuring.

### *Develop Rules*

For each one of your measures, compute the difference between the two airlines and decide how large the difference between the two airlines would need to be to say that one airline is truly more/less reliable than the other.

Use at least TWO of your measures to *develop and state a single rule* that another person can apply to data from two airlines to determine whether or not these airlines truly differ in their reliability.

### *Test Rules*

Using four other sets of data collected from these same two airlines, apply your rule to each set of data. Adapt or modify your group's rule as you need to. This may include small changes or adding to or changing the measures that you used in your initial rule.

*Summarize*

Your group will now write a letter to the editor-in-chief of *Chicago Magazine* that includes the following:

- Your group's set of rules, used to analyze the five sets of data. In your letter the rules need to be clearly stated so that another person could apply them to compare data for two other airlines;
- A response to the editor's initial questions:
  - Is there a difference in overall reliability as measured by arrival time delays for these two regional airlines out of Chicago to Minneapolis? Or are both airlines pretty much the same in terms of their arrival time delays?
  - Are any differences you find large enough to influence travelers so that they are advised to choose one airline over the other (all other factors being equal)?

*Type the letter in a word-processed document and email it to each of your group members and the instructor.*

## *Discussion*

As a group, discuss your responses to each of the following questions.

1. Was it difficult to come up with a rule to determine whether one airline was really different in overall reliability than another airline? Why or why not?

2. How might your rules change if you were comparing more than two airlines for each city?

3. How might your rules change if you were comparing two airlines for more than two cities?

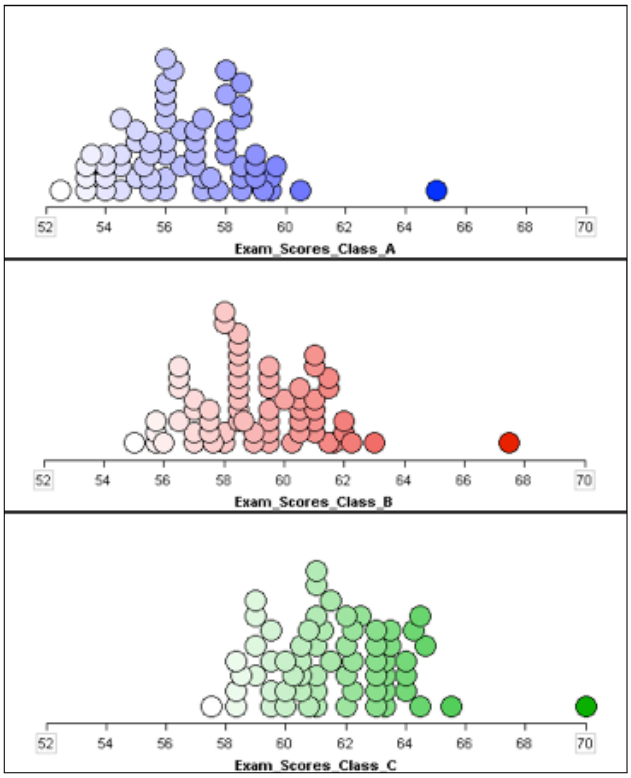
4. What does your group need to do to improve the process of working as a team? Be specific about how each member of the group will contribute to this improvement.

# *Course Activity: Characteristics of Distributions*

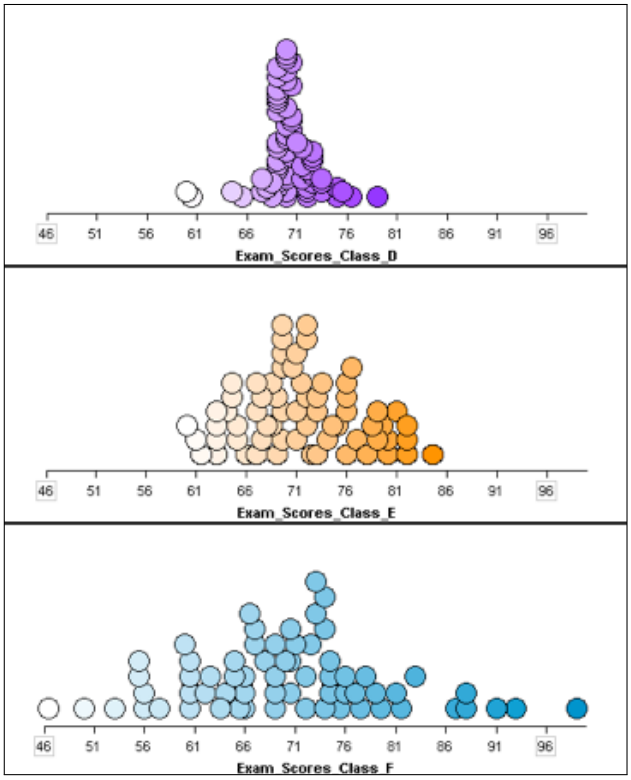
Imagine multiple sections of the same college course, taught by different instructors. Below are a series of plots that depict the distributions of hypothetical exam scores in various sections.

*Nausicaa Distribution* is an Etsy store from which you can order statistical gifts including ten different distribution plushies. http://www.etsy.com/shop/NausicaaDistribution?section_id=6067670

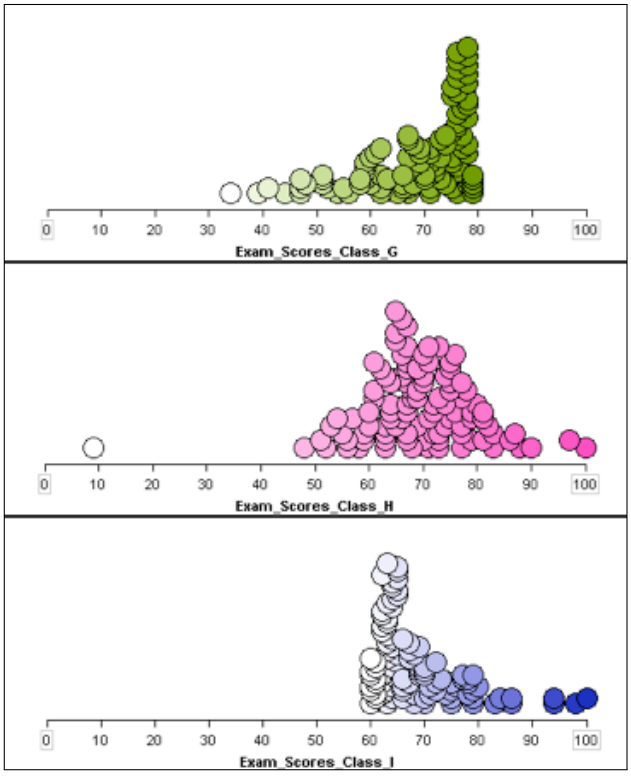
1. For classes *A*, *B* and *C*, what is the main characteristic that distinguishes these distributions from each other? What are potential factors that might explain the differences?



2. What is the main characteristic that distinguishes the distributions of exam scores in class *D*, *E*, and *F*? What are potential factors that might explain the differences?

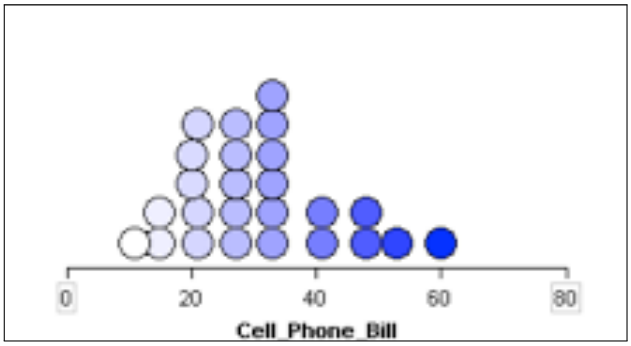


3. What is the main characteristic that distinguishes the distributions of exam scores in class *G*, *H*, and *I*? What are potential factors that might explain the differences?



## *Cell Phone Bills*

Consider a survey study conducted on a random sample of 25 University of Minnesota students. One survey item asked students to self-report the amount of his or her last cell phone bill (in dollars). The plot of the bill amounts is shown below.




4. If you wanted to tell someone the amount of a "typical" cell phone bill for these students, what would you say?

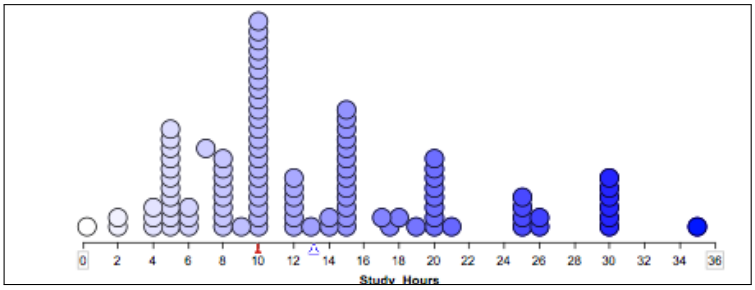
5. How would you describe (quantify) the *overall* variation in these 25 cell phone bills? How far from the typical amount that you identified in Question 4 are *most* of the students bills?

6. What is a potential factor(s) that might explain the variation in these bills?

7. Using the typical cell phone bill you identified previously as a reference point, consider the *overall* variation in the distribution on both sides of this point. Is the variation roughly the same on the left- and right-hand side of this point? Is there more or less variation on either side of this value?

## *Number of Hours Studied*

The plot below contains responses from 100 EPsy 3264 students who responded to the survey question, "How many hours per week do you typically study?" These students' responses are a random sample from all responses obtained from all in-class sections taught from 2004–2010. Examine the plot of these data.




8. What does each dot (i.e., case) in the distribution represent?

9. Summarize the characteristics of the distribution. Be sure to identify the typical amount of time spent studying and the variation in the amount of studying. When describing the variation, you should quantify the "average" amount of deviation from the typical value and also indicate the shape.

10. What is a potential factor(s) that might explain the variation in these data?

# *Describing Distributions*

One of the important steps in any statistical analysis is that of summarizing data. It is good practice to examine both a graphical and a numerical summarization of your data. These summarizations are often part of the evidence that researchers use to support any conclusions drawn from the data. They also allow researchers to discover structure that might have otherwise been overlooked in the raw data that was actually collected. Lastly, both graphical and numerical summaries of the data often point to other analyses that may be undertaken with the data.

Once raw data has been collected in a study, it can be overwhelming to pull any kind of meaning out of it. For example, it is not uncommon for Google to be dealing with millions of cases. How can Google—or any researcher for that matter—go from all of that raw data to something that can help them answer their research questions?

Rather than examining all of those cases individually, researchers examine the data collectively, often by plotting it. This is what is meant by a graphical summary of the data; it is quite literally, a picture of the distribution. For example, the plots of the exam scores you saw in the *Characteristics of Distributions* activity provided graphical summaries of the raw exam scores for each of the nine classes.

There are many, many different types of plots that have been created to graphically summarize data. Each can provide a slightly different representation of the data. Metaphorically, you can imagine each of these different plot types as a differ-

ent photo taken of the exact same person. Some may be color, others black and white. Some may be taken from different perspectives, angles or distances. While all photographs "summarize" the same person, you may notice characteristics of that person in some photos that are not evident in others. Many of the photos, however, will show the same thing.

## *Shape*

The dotplot that TinkerPlots™ provides is a very useful plot. It allows us to summarize the *shape* of the distribution very easily. Shape is used to describe a distribution's symmetry. As you might expect, *symmetric* distributions are shaped the same on either side of the center. (Another way of thinking about this is that if you folded the distribution at the center, the folded half of the distribution would align pretty well on top of the other half.) For example,"bell-shaped" or "normal" distributions are symmetric.

TinkerPlots™ also provides other types of plots than the dotplot, including the box plot (sometimes called the box-and-whiskers plot) and the hat plot (a variation of the box plot).

When a distribution is asymmetric, it is referred to as a *skewed* distribution. The distribution shown in Figure 1 is a skewed distribution. In this distribution, there appears to be a longer tail on the right side of a distribution. Because the tail is on the right side of the distribution, statisticians would say it is "skewed to the right" or "positively skewed". In a similar way, a distribution that tails to the left is "skewed to the left" or "negatively skewed".

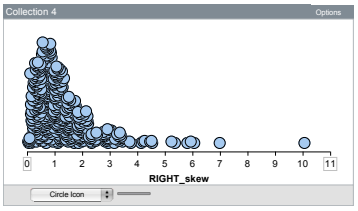



Figure 1: A distribution skewed to the right.

## *Location*

Aside from the overall shape of the distribution, it is also useful to summarize the *location* of the distribution. The location of the distribution provides a summarization of a so-called "typical" value for the data. A "typical" value can be estimated from the plot of the distribution. You can also use more formally calculated summaries of the location such as the mean,

median, or mode. These values are easily calculated using TinkerPlots™.

When looking at a plot of a distribution, data analysts often consider the number of modes or "humps" that are seen in a plot of the distribution. Here, the concept of mode is slightly different—although related—to the concept of mode that you may have learned in previous mathematics or statistics courses. The mode of a distribution gives a general sense of the values or measurements that occur frequently. This may be a single number, but many times is not. For example, the first hump of the distribution shown in Figure 2 suggests that values around 9 are very common. The actual value of 9, however, may only show up once or twice in the data.

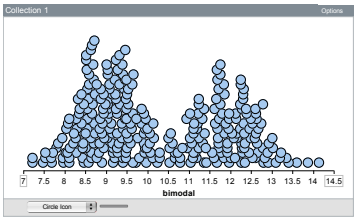


Figure 2: A bimodal distribution showing two modes.

A distribution can be unimodal (one mode), bimodal (two modes), multimodal (many modes), or uniform (no modes). The distribution shown in Figure 2 is bimodal—notice there are two humps. Uniform distributions have roughly the same frequency for all possible values (they look essentially flat) and thus have no modes.

## *Variation*

The other characteristic of a distribution that should be summarized is the *variation*. Summarizing the variation gives an indication of how variable the data are. One method of numerically summarizing the variability in the data is to quantify how close the observations are relative to the "typical" value *on average*. Are they for the most part close to the "typical" value? Far from the "typical" value? How close?

It turns out, that the shape of the distribution also helps describe the variation in the data. For example, "bell-shaped" distributions have most observations close to the typical value, and more extreme observations show up both below and above the typical value (the variation is the same on both sides of the "typical" value). Whereas skewed distributions have many observations near the typical value, but extreme values only

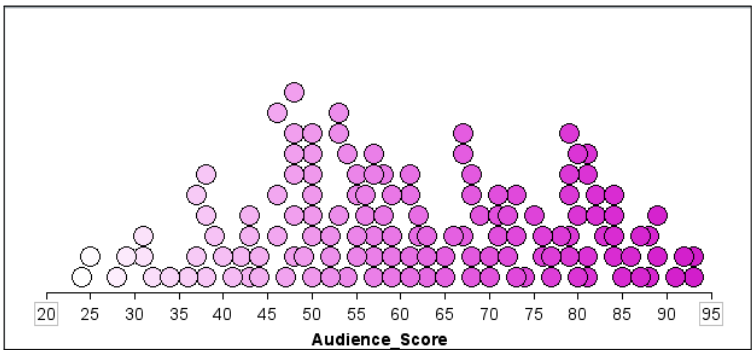
deviate from this value in one direction (there is more variation in the data on one side of the "typical" value than the other).

One thing that affects the variation, and should be described is whether there are observations that stand out from the other observations. Often these observations have extremely large or small values relative to the other observations. These observations are referred to as *outliers*, or extreme cases. For example, in the distribution shown in Figure 1, the observation that has a value near 10 would likely be considered an outlier.

## *Putting It All Together*

Rotten Tomatoes is a website which aggregates the movie critics' reviews of films. The website marks each review as either positive or negative and then gives a score based on the percentage of positive reviews. In addition, the general public can also give a positive or negative review to a film. Again, these reviews are tabulated and the score given each film represents the percentage of positive reviews.

The following is a graphical summary of the scores for 134 movies released in 2009 based on the general public's reviews.




The scores for a sample of movies released in 2009 based on the general public's reviews. The scores represent the percentage of positive reviews for each movie

A written description of the distribution might read as follows,

> The distribution of scores for this sample of movies is fairly symmetric. The median score for these movies is near 60 indicatng that a typical movie released in 2009 is positively reviwed by about 60% of the public. The distribution also indicates that there is a lot of variation in the movies' scores. Most of the movies in the sample have a score between 35 and 85, suggesting large differences in the public's opinion of the quality of movies.

# *Course Activity: Memorization*

Many times during the semester, you may feel like your brain just can't hold all of the information you are learning in classes. Are there ways to improve our memories so that we can comprehend even more information? Research in cognitive psychology has suggested that the answer to that question is a resounding "yes". This literature has suggested several strategies to improve memory, enhance recall and increase retention of information.

One of the strategies identified by cognitive psychologists is that of chunking. Chunking refers to the process of taking individual units of information and grouping them into larger units (chunks). One common example of chunking occurs when we write and recall phone numbers. For example, a phone number sequence of 8-6-7-5-3-0-9 would be chunked into 867-5309.

In this activity, you will be exploring the following research question:

Does chunking by familiar letters improve memory?

To examine this research question, you will use the data collected from the memory experiment your class just partook in.

## *Examining the Observed Data*

The first part of any analysis is to examine the observed data. These are the data that are observed in the research study. Before you can examine the data, however, you will need to enter the data collected in the study into a TinkerPlots™ case table. See instructions in margin for how to set up the table.

**Setting up the Case Table**

- Drag a `Table` from the object toolbar into your document.
- Create a new attribute called *Score* in the first column of the case table.
- Create another new attribute called *Condition* in the second column of the case table.

Each row in the table will comprise a subject in the research study. Each column will comprise an attribute of the subject. For our purposes, you will need to enter data for two attributes. The first attribute will indicate the subject's score from the memory experiment. This is called the *response variable* since it contains data on the subjects' responses to the experiment. The second attribute will indicate the treatment condition that the subject was assigned to. This is called a *treatment variable*. In this research study the two treatment conditions are the experimental condition (familiar chunking) and the control condition (unfamiliar chunking).

- Enter the observed data from your class experiment into a TinkerPlots™ case table.
- Plot the observed data (see instructions in margin).

**Plotting the Observed Data**

- Drag a `Plot` from the object toolbar into your document.
- Drag the response attribute from the case table containing the trial results to the *x*-axis of the plot.
- Drag a case icon to the right until the cases in the plot are fully separated (e.g., no vertical bin lines). Can also double-click on one of the endpoints and change `Bin width = 0`.
- Drag the treatment attribute from the case table containing the trial results to the *y*-axis of the plot.
- Click the `Vertical Stack` button in the upper plot toolbar to organize the icons.

1. Sketch the plot below.

### *Summarizing the Difference Between the Two Conditions*

In order to answer the research question, you need to summarize the difference between the two treatment conditions into a single number. You typically do this by finding the mean for each treatment condition, and then computing the difference between the means. The difference in means satisfies the need for a single number summary. It also has another very nice quality, and that is the difference in means is interpretable. The difference in means indicates *how much better* the typical subject in the experimental condition does than a typical subject in the control condition.

2. Find the difference in means for the observed data by subtracting the mean score for the control condition from the mean score for the experimental condition.

3. Interpret this difference using the context of the memory study.

4. If there is *not* an effect of chunking on memory, what would you expect the difference in means to be? Explain.

5. Does the difference you found in the observed data suggest there is an effect of chunking on memory or not? Explain.

## *Considering Chance Variation as an Explanation for the Difference in Means*

Before you conclude that chunking has an effect on memory, consider another alternative—*the difference in means you saw in the observed data is solely attributable to chance variation.* Think back to the activities and homework you did previously. Would you conclude that a coin is "unfair" because you flipped it ten times and got six heads? Probably not.

If you repeatedly flip a coin ten times, sometimes you get five heads. Other times you get four heads, or three heads, or seven heads, etc. This variation in the results is not because you flipped a different coin or used a different model, but rather because the random process introduces variation into the results.

Concluding that there is an effect of chunking on memory just because the difference in means from the observed data is not zero is akin to concluding that a coin is not fair just because you did not get exactly 50% heads!

How did you know whether a particular result—say 80 heads in 100 flips—indicated a coin was unfair? You modeled flipping a fair coin 100 times. By examining the variation in the results of repeatedly carrying out this experiment, you could investigate whether the observed result of 80 heads was likely or not under the assumption of flipping a fair coin.

You will use the same ideas to examine the result obtained from our observed memory data.

## *The Null Model*

To examine whether a result obtained in the observed data is solely due to chance, one approach is to imagine the *scenario under which the chunking had no effect* whatsoever. Under this assumption or scenario, evidence would be collected to de-

termine if the difference in means that was observed in the data is too large to probabilistically believe that there is no effect of chunking. This statement or assumption of no effect of chunking is called the *null model* and is written as

$H_0$ : There is no difference between the mean number of letters remembered between the chunking and non − chunking groups.

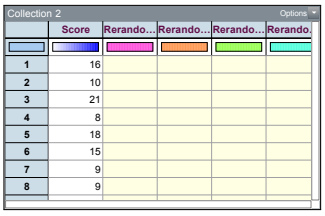
If chunking is truly ineffective, then each subject's score on the memory test is only a function of that person and not a function of anything systematic, such as the chunking The implication of this is that, had a subject been assigned to the other condition (through a different random assignment), her score on the memory test would have been identical since, in a sense, both conditions are doing nothing in terms of affecting the memory test scores.

### *Re-randomization: Inspecting Other Possible Random Assignments of the Subjects*

A researcher can take advantage of the idea that each subject's score on the memory test would be identical whether she was assigned to treatment or control and examine other possible random assignments of the subjects to conditions that could have occurred. To do this,

- In TinkerPlots™ drag a new case table into your document.
- Copy and paste the subjects' observed scores on the memory test (i.e., the response attribute) from the previous case table into this new case table. Call this attribute *Scores*.
- Create a new attribute in the case table called *Rerandomized Treatment 1*.
- Add three more attributes to the case table, labeling them *Rerandomized Treatment 2*, *Rerandomized Treatment 3*, and *Rerandomized Treatment 4*.

The new case table should now have an attribute of the original observed memory test scores and an empty attribute where you will input the "new" random assignment of conditions.



Screenshot of the newly created case table.

## *Physical Simulation of the Re-Randomization*

To aid you in creating these 'new" random assignment of conditions, fill in the following:

In the original experiment _____ subjects were randomly assigned to the experimental (familiar chunking) condition and _____ subjects were assigned to the control (unfamiliar chunking) condition.

- You will be given several index cards with either an *E* (for experimental) or a *C* (for control). Each index card represents a subject. Count the index cards to be sure that you have the same number of *E* cards as subjects originally assigned to the experimental condition and the same number of *C* cards as subjects originally assigned to the control condition.
- Shuffle the index cards together several times.
- Deal the shuffled index cards out one at a time. Record the condition on the first index card in the *Rerandomized Treatment 1* attribute so it corresponds to the first subject. Continue recording the condition on each subsequent index card in turn for each subsequent subject.
- Plot the re-randomized data. (If you have forgotten how to plot the data, go back and re-read the directions from earlier in this activity.)

The data in the case table, and in the plot you just created, represent another way that the subjects could have been randomly assigned to the two conditions. This random assignment likely has different subjects in the control and experimental conditions than the observed data. Because of this, the mean memory score for the two conditions will also likely differ from

the observed data. This, in turn, implies that the difference in means will also be different.

6. Compute and record the difference in means for the re-randomized data. Be sure that the order you use when subtracting is consistent with the order you subtracted to obtain the original observed result. Note that you may obtain a negative number here.

- Repeat the re-randomization process three more times. Each time, compute and record the difference between the means of the two groups. (Remember to subtract the mean score for the control condition from the mean score for the experimental condition.)
- Input each of the four differences you obtained from the four re-randomizations into your instructor's computer.

### *Examining the Distribution of the Difference in Means*

- Open a new TinkerPlots™ document.
- Drag a new case table into the document.
- Create one new attribute called *Difference in Means*.
- Enter the difference in means from all groups (recorded in the instructor's computer) into the *Difference in Means* attribute.
- Plot the *Difference in Means* attribute.

7. Sketch the plot of the difference in means below.

8. Does it look like it centers around zero? Explain why the distribution is centered at zero. (Hint: Think back to what the null model was.)

9. Quantify the strength of evidence for the observed result (i.e., How far in the tail of the distribution is the observed result? The furthest 5%? 1%?).

When reporting the results from statistical analyses, the strength of evidence is referred to as a ***p*-value**.

10. What does this suggest about whether the difference in means that was observed is solely due to chance? Explain.

11. Use your previous answer to answer the research question.

**Recap**

If there really were no effect of the grouping of letters, is it possible that random chance alone could have resulted in such an extreme observed difference between the two conditions? Once again, the answer is yes, this is indeed possible. Also once again, the key question is *how likely it would be for random chance alone to produce experimental data that favor the familiar chunking condition by as much as the actual experimental data do*. You will aim to answer that question using the following simulation analysis strategy:

- **Model:** Assume that there is no effect of the grouping of letters on the scores (the null model).
- **Simulate:** Replicate the random assignment of these subjects and their memory scores between the two conditions. You will repeat this random assignment a large number of times. Each time you will calculate a measure of how different the conditions are, in order to get a sense for what is expected and what is surprising.
- **Evaluate:** If the actual result obtained is in the tail of the null model's distribution, you will reject that null model.

# *What the p-value?*

For the next class activity, you will learn how to conduct a randomization test using TinkerPlots™. To introduce you to the context of the data you will be using, we would like you to read an article describing a study that was reported in *Nature*. We would also like you to read a web article that will help you further understand ideas related to *p*-values.

- Read the article, *Visual Discrimination Learning Requires Sleep after Training*. The article is available at `http://www.nature.com/neuro/journal/v3/n12/pdf/nn1200_1237.pdf`.

- Read the web article *Mission Improbable: A Concise and Precise Definition of P-Value* `http://news.sciencemag.org/sciencenow/2009/10/30-01.html`

# *Course Activity: Sleep Deprivation*

Sleep deprivation has been shown to have harmful effects such as fatigue, daytime sleepiness, clumsiness and weight loss or weight gain. Researchers have also established that sleep deprivation has a harmful effect on learning. But do these effects linger for several days, or can a person "make up" for sleep deprivation by getting a full night's sleep in subsequent nights?

A recent study (Stickgold, James, and Hobson, 2000) investigated this question by randomly assigning 21 subjects (volunteers between the ages of 18 and 25) to one of two groups: one group was deprived of sleep on the night following training and pre-testing with a visual discrimination task, and the other group was permitted unrestricted sleep on that first night. Both groups were then allowed as much sleep as they wanted on the following two nights. All subjects were then re-tested on the third day.

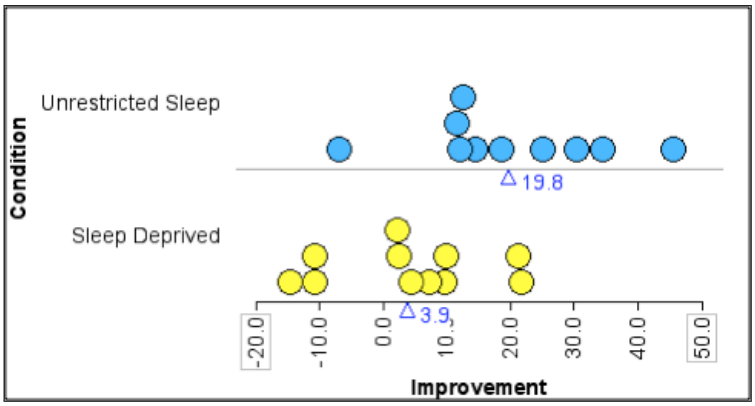
In this activity, you will be exploring the following research question:

Does the effect of sleep deprivation last, or can a person "make up" for sleep deprivation by getting a full night's sleep in subsequent nights?

Subjects' performance on the test was recorded as the minimum time (in milliseconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen. The sorted data and plots presented here are the improvements in those reporting times between the pre-test and post-test (a negative value indicates a decrease in performance):

| ***Sleep deprivation*** ($n = 11$) | ***Unrestricted sleep*** ($n = 10$) |
|---|---|
| −14.7 | −7.0 |
| −10.7 | 11.6 |
| −10.7 | 12.1 |
| 2.2 | 12.6 |
| 2.4 | 14.5 |
| 4.5 | 18.6 |
| 7.2 | 25.2 |
| 9.6 | 30.5 |
| 10.0 | 34.5 |
| 21.3 | 45.6 |
| 21.8 | |

Data for the sleep deprivation study.




Plot of the observed data. The triangle under each plot indicates the mean improvement score for the respective group.

***Discuss the following questions.***

1. Does it appear that subjects who got unrestricted sleep on the first night tended to have higher improvement scores than subjects who were sleep deprived on the first night? Explain briefly.

2. Is the mean improvement higher for those who got unrestricted sleep? Calculate the difference in the means of the improvement scores. Does this appear to be a large difference?

3. Is it possible that there is really no harmful effect of sleep deprivation, and random chance alone produced the observed differences between these two groups?

In this study, the random chance is introduced not through the sampling process (like in Unit 1), but rather in the random assignment to groups. While it is possible that sleep deprivation has a harmful effect, it is also possible that it does not have a harmful effect and the researchers were just unlucky and happened to "assign" the subjects who were going to have more improvement in their test scores (regardless of which group they were going to be assigned to) into the unrestricted sleep group.

Similar to the simulation study in the *Memorization* activity, one way to examine this is to consider what you would likely see if there really is no difference in test score improvement between the two conditions. (This is the null hypothesis!) If that is the case, it is reasonable to assume that these subjects

would improve the same amount regardless of which group they had been assigned to because the effect on test scores would be identical for both groups.

Even if that is the case, however, since there are many possible ways to randomly assign 21 subjects into two groups, it is possible that the random assignment that came up was just unlucky and happened to "assign" the subjects who were going to have more improvement in their test scores into the unrestricted sleep group. If the random assignment had come up differently, it might have appeared as though there were no harmful effects, or even that the effects were beneficial—all because of random chance!

**The key statistical question is:** If there really is no difference between the conditions in their effects on test improvement, how unlikely is it to see a result as extreme or more extreme than the one you observed in the data just because of the random assignment process alone?

### *Modeling the Differences in Improvent Under the Null Model*

You will conduct a randomization test using TinkerPlots™ to find out how likely it would be, assuming there is *no difference* between the conditions in their effects on test improvement, to see a result as extreme or more extreme than the mean difference you observed in the data (15.9) just because of the random assignment process alone.

The underlying idea of the randomization test is that the treatment condition labels are re-randomized—which produces a different random assignment of the subjects to the two conditions that could have occurred. The difference in the mean improvement scores is then computed under this re-randomization. This process of re-randomizing of the data

and computing the difference in the mean improvement scores is repeated many times. The distribution of these differences displays the variation expected just because of the random assignment process alone and can be used to evaluate the result that was originally observed.

## *Randomization Tests in TinkerPlots™*

In order to carry out a randomization test using TinkerPlots™, you need to include multiple devices in the sampler. The first device will include the observed response data for all of the subjects. The second device will contain the treatment condition labels.

## *Modeling a Set of Fixed Responses Under the Null Model*

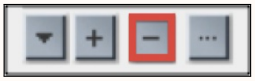
Under the null hypothesis of no difference between the two conditions, these response values are fixed—they will always be the same for the subjects. To produce simulated data that are fixed, you will use a sampling device called a `Counter`. Whereas the devices you have used thus far, spinners and mixers, select elements and values randomly, the counter device selects elements systematically.

- Open the *Sleep-Deprivation.tp* data set.
- Set up a model that will produce the fixed responses for the subjects under the null model (see instructions in margin).
- Run the model.

4. Why did you change the `Repeat` value to *21*?

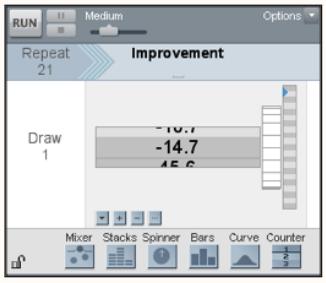
**Setting Up the Model: Fixed Responses**

- Drag a new `Sampler` from the object toolbar into your blank document.
- The default device in the sampler is a `Mixer` with three elements. Remove all three of the elements by repeatedly clicking the `Remove Element` button in the device toolbar at the bottom of the sampler.



- Drag a `Counter` from the device toolbar at the bottom of the sampler onto the current device.


- Right-click the attribute *Improvement* in the case table and select `Copy Attribute`.
- Click the grey area in the sampler window (outside of where the numbers will go) and from the `Edit` menu select `Paste Cases`.
- Change the name of the device from *Attr1* to *Improvement*.
- Change the `Draw` value to *1*, and the `Repeat` value to *21*.



5. Can you predict the outcomes generated in each trial?

## *Modeling the Random Assignment of the Treatment Condition Labels by Linking Multiple Devices*

To model the random assignment of the treatment condition labels that might have occurred, you need to produce simulated data from another model that generates 11 labels of *Sleep Deprived* and ten labels of *Unrestricted Sleep*. To do this you will use the `Stacks` sampling device.

You also need to have the 21 fixed responses appear in the same case table as the 21 group labels. This allows us to easily attach a particular response to a label. To have the outcomes from multiple sampling devices appear in the same case table in TinkerPlots™, you *link* multiple sampling devices in the same sampler.

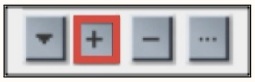
- Add a linked `Stacks` device to the sampler with the *Improvement* counter (see instructions in the margin).

When you add linked devices, note that the value for `Draw` changes automatically to the number of devices included in the sampler. A TinkerPlots™ sampler showing two linked devices modeling the random assignment of responses to conditions is shown below.

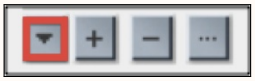

**Linking Multiple Devices**

- Drag a `Stacks` device from the device menu to the right-hand side of the existing *Improvement* counter.


- The sampler should contain two devices linked by a grey line.
- Change the device name from *Attr2* to *Condition*.
- Click the `Add Element` button twice to add two elements to the stacks. These elements will indicate the treatment condition labels.



- Change the label of the first bar from *a* to *Sleep Deprived*. Change the label of the second bar from *b* to *Unrestricted Sleep*.
- Click on the `Device Options` button (below the stacks device) and select `Show Count`.



- Change the count value for the *Sleep Deprived* label to *11* and change the count value for the *Unrestricted Sleep* label to *10*.
- Click on the `Device Options` button (below the stacks device) and from the `Replacement` menu select `Without Replacement`.

- Run the model.

The outcomes from both linked devices are recorded in the case table, each in their own attribute. In addition, an attribute called *Join* is also created that includes the outcomes of all the linked devices separated by a comma.

In this simulation, the trial represents what might have occurred for another random assignment of subjects to conditions.

- Plot the trial data (see instructions in the margin).

6. Sketch the plot below.

**Plotting the Data from a Trial**

- Drag a `Plot` from the object toolbar into your document.
- Drag the response attribute from the case table containing the trial results to the $x$-axis of the plot.
- Drag a case icon to the right until the cases in the plot are fully separated.
- Drag the treatment condition attribute from the case table containing the trial results to the $y$-axis of the plot.
- Click the `Vertical Stack` button in the upper plot toolbar to organize the icons.

As you have done in previous simulations, you will numerically summarize the trial results. From earlier in this activity, you summarized the observed data by computing the difference in the mean improvement scores between the two conditions. In fact, you computed,

$$\bar{X}_{\text{Unrestricted Sleep}} - \bar{X}_{\text{Sleep Deprived}}$$

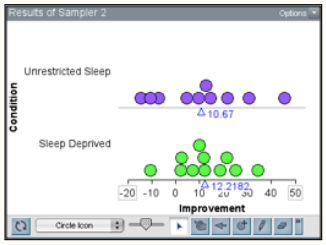
You need to compute the summary of the trial data in the exact same manner.

- Use TinkerPlots™ to summarize the results of the trial by computing the mean improvement scores for each condition in the re-randomized data (see instructions in margin).

**Summarizing Trial Results: Means**

- Click on the `Mean` button in the upper plot toolbar.

- Click on the `Average Options` menu (the upside-down triangle next to the `Mean` button) and select `Show Numeric Value(s)`.



7. Compute the difference in means between the re-randomized groups in the trial.

## *Computing the Difference in Means*

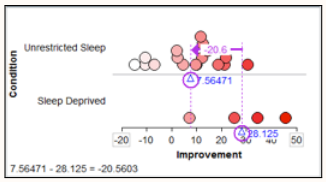
You need to have TinkerPlots™ compute the difference in means for the simulated data so that you can collect this result from many trials. The `Ruler` can be used to compute the difference between two values in a plot.

- Use the `Ruler` to compute the difference in means between the re-randomized groups in the trial (see instructions in margin).
- Check that the difference in means is the same as the difference you computed in Question 8. (If the difference calculated by TinkerPlots™ is correct, but has a reversed sign, you dragged the dotted lines to the wrong group.)

**Computing the Difference in Means**

- Click on the `Ruler` button in the upper plot toolbar.

- Drag the right-most vertical dotted line on top of the mean-triangle for the *Unrestricted Sleep* condition. (Note that a purple circle will appear around the mean triangle.)
- Drag the left-most vertical dotted line on top of the mean-triangle for the *Sleep Deprived* condition. (Note that a purple circle will appear around the mean triangle.)
- The difference between the two means will be calculated in the lower left-hand corner of the plot window.




## *Collecting the Difference in Means*

- Collect the difference in means calculated by the ruler.
- In the *History of Results* case table, collect an additional 99 measures.

## *Evaluating the Observed Result*

- Plot the differences in means from the 100 simulated trials.

8. Sketch the plot below.

9. What are the cases in the plot? (Hint: Ask yourself what each individual dot represents.)

10. Where is the plot of the results centered (at which value)? Explain why this makes sense. (Hint: Think about what the null model is.)

11. Based on the plot, is the actual experimental result found in the observed data likely to have arisen solely from random assignment? Explain.

12. Quantify the strength of evidence for the observed result (i.e., How far in the tail of the distribution is the observed result? The furthest 5%? 1%?).

13. In light of your answers to the previous two questions, would you say that the results that the researchers obtained provide strong evidence that the effect of sleep deprivation is harmful (i.e., that the null model is not correct)? Or can a person "make up" for the lost sleep by getting a full nights rest on subsequent nights? Explain your reasoning based on your simulation results. Include a discussion of the purpose of the simulation process and what information it revealed to help you answer this research question.

### *Example Write-Up for Sleep Deprivation Study*

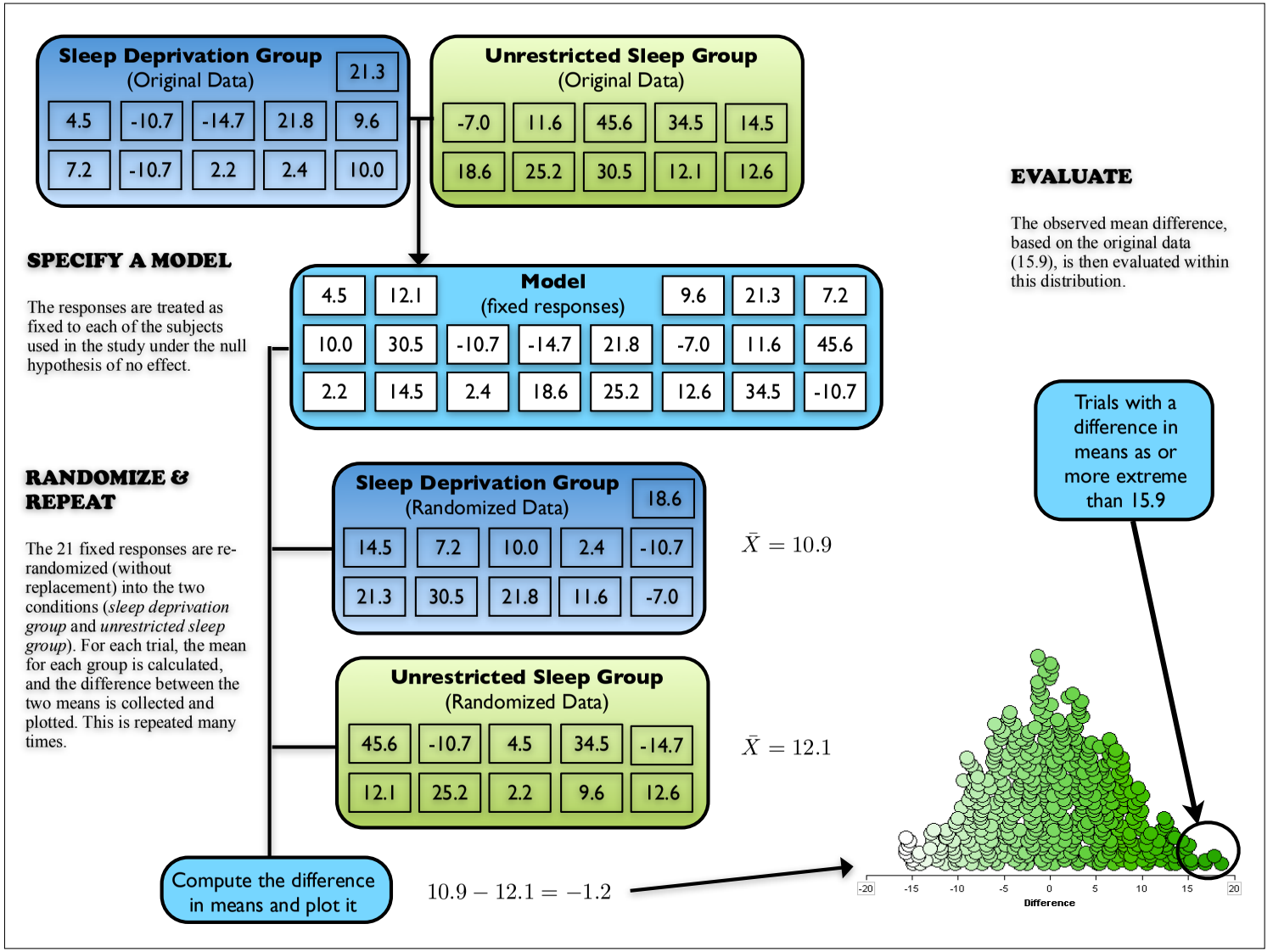
When reporting the results of a simulation study, pertinent information from the analysis that needs to be included is:

- **The *type of test* used in the analysis (including the number of trials):** Randomization test (100 trials)
- **The *null model* assumed in the test:** No difference in the mean improvement scores between the two conditions.
- **The *observed result* based on the data:** Unrestricted sleep ($M = 19.82$ milliseconds), sleep deprived ($M = 3.9$ milliseconds), and the difference in means (15.92)
- **The *p*-value for the test:** $p = 0.02$ (one-tailed)
- **All appropriate *inferences* based on the *p*-value and study design:** Observed results do not support the null model

As an example, we may write-up the analyses of the sleep deprivation study as follows.

> To study whether or not a person can "make up" for lost sleep by getting a full night of sleep on subsequent nights, 21 study participants were randomly assigned to one of two conditions. The subjects assigned to the sleep deprived condition ($n$ = 11) were deprived of sleep on the night following training and pretesting with a visual discrimination task. The subjects assigned to the unrestricted sleep condition ($n$ = 10) were permitted unrestricted sleep on that first night. Both groups were then allowed as much sleep as they wanted on the following two nights. All subjects were then retested on the third day. The data suggests that the group allowed unrestricted sleep ($M$ = 19.82 ms) showed, on average, 15.92 ms more improvement than the sleep-deprived group ($M$ = 3.9 ms).

A randomization test was used to determine whether there was a statistically reliable difference in the improvement in reporting times between participants in these two groups. Under the null model of no difference in the average improvement scores, a $p$-value of 0.02 (one-sided) was computed by re-randomizing the data 100 times. The observed data provides strong evidence against the null model of no difference in improvement scores. The results of this study may suggest that people cannot "make up" for sleep deprivation by getting a full night of sleep on subsequent nights.


Sleep Deprivation Group
(Original Data)
21.3
4.5 -10.7 -14.7 21.8 9.6
7.2 -10.7 2.2 2.4 10.0
Unrestricted Sleep Group
(Original Data)
-7.0 11.6 45.6 34.5 14.5
18.6 25.2 30.5 12.1 12.6
SPECIFY A MODEL
The responses are treated as fixed to each of the subjects used in the study under the null hypothesis of no effect.
Model
(fixed responses)
4.5 12.1 9.6 21.3 7.2
10.0 30.5 -10.7 -14.7 21.8 -7.0 11.6 45.6
2.2 14.5 2.4 18.6 25.2 12.6 34.5 -10.7
RANDOMIZE & REPEAT
The 21 fixed responses are re-randomized (without replacement) into the two conditions (sleep deprivation group and unrestricted sleep group). For each trial, the mean for each group is calculated, and the difference between the two means is collected and plotted. This is repeated many times.
Sleep Deprivation Group
(Randomized Data)
18.6
14.5 7.2 10.0 2.4 -10.7
21.3 30.5 21.8 11.6 -7.0
$\bar{X} = 10.9$
Unrestricted Sleep Group
(Randomized Data)
45.6 -10.7 4.5 34.5 -14.7
12.1 25.2 2.2 9.6 12.6
$\bar{X} = 12.1$
Compute the difference in means and plot it
$10.9 - 12.1 = -1.2$
EVALUATE
The observed mean difference, based on the original data (15.9), is then evaluated within this distribution.
Trials with a difference in means as or more extreme than 15.9
-20 -15 -10 -5 0 5 10 15 20
Difference

# *Course Activity: Strength Shoe®*

The Strength Shoe® is a modified athletic shoe with a 4-cm platform attached to the front half of the sole. Its manufacturer claims that this shoe can increase a person's jumping ability. In this activity you will be examining the following question:


How can you determine whether the manufacturer's claim about the Strength Shoe® is legitimate?

***Discuss the following questions.***

1. If your friend who wears strength shoes can jump much farther than another friend who wears ordinary shoes, would you consider that compelling evidence that strength shoes really do increase jumping ability? Explain.

2. Now suppose that you take a random sample of individuals by randomly selecting them from the population. You observe who does and does not wear strength shoes, and then compare the two group's jumping ability. If, on average, the group who wears strength shoes can jump much farther than the group who wears ordinary shoes, would you consider that compelling evidence that strength shoes really do increase jumping ability? Explain.

Both descriptions above lack compelling evidence to claim that people who wear the Strength Shoe® jump farther.

The evidence from the first situation is based on *anecdotal evidence*. Anecdotal evidence results from situations that come to mind easily and is of little value in scientific research. Much of the practice of statistics involves designing studies and collecting data so people do not have to rely on anecdotal evidence.

The problem with the evidence from the second situation is that you do not know whether or not the two groups might differ in more ways than simply one. For example, subjects who choose to wear the strength shoes could be more athletic to begin with than those who opt to wear the ordinary shoes. If one group is more athletic than the other, it could *confound* the results of the study.

When investigating whether or not one variable causes an effect on another, researchers seek to exert control by creating a comparison group and then assigning subjects to either the treatment group or the comparison group. An experiment is a study in which the experimenter actively imposes the treatment condition on the subjects. Ideally, the groups of subjects are identical in all respects other than the condition, so the researcher can then see the variable's direct effects on the response variable.

A 1993 study published in the *American Journal of Sports Medicine* investigated the Strength Shoe® claim using 12 intercollegiate track and field athletes as study participants. Suppose you also want to investigate this claim, and you recruit 12 of your friends to serve as subjects. You plan to have six people wear a Strength Shoe® and the other six wear an ordinary shoe and then measure each group's jumping ability.

Cook, S. D., Schultz, G., Omey, M. L., Wolf, M. W., & Brunet, M. F. (1993). Development of lower leg strength and flexibility with the strength shoe. *American Journal of Sports Medicine, 21,* 445–448.

***Discuss the following question.***

3. How might you assign subjects to these two groups in an effort to balance out potentially confounding variables?

## *Random Assignment*

*Random assignment* is the preferred method of assigning subjects to treatment conditions in an experiment. One characteristic of random assignment that makes it a good method of assigning subjects to conditions is that under random assignment, each subject has an equal chance (probability) of being assigned to any of the treatment conditions. In addition, there are several other benefits to using random assignment. You will explore the properties and benefits of random assignment in this activity.

The word 'randomization' is a synonym for random assignment.

4. Describe in detail (so another student could replicate the process) how you might implement the process of randomly assigning subjects to either the treatment or comparison condition.

Suppose that your 12 subjects are listed in the following table. You record their gender and height in inches because you suspect that these variables might be related to jumping ability:

| Name | Gender | Height |
|---|---|---|
| Anna | Female | 61 |
| Kyle | Male | 71 |
| Patrick | Male | 70 |
| Audrey | Female | 67 |
| Mary | Female | 66 |
| Peter | Male | 69 |
| Barbie | Female | 63 |
| Matt | Male | 73 |
| Russ | Male | 68 |
| Brad | Male | 70 |
| Michael | Male | 71 |
| Shawn | Male | 67 |

Data for the Strength Shoe® study.

5. Take 12 index cards, and write each subject's name (along with their gender, and height) on a different card. Shuffle the cards and randomly deal out six for the Strength Shoe® group and six for the ordinary shoe group. Record the names assigned to each group in this table, along with their gender and height in the tables on the following page.

| Name | Gender | Height |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

Strength Shoe® Group

| Name | Gender | Height |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

Ordinary Shoe Group

6. Calculate and report the proportion of men in each group. Also subtract these two proportions (taking the Strength Shoe® group's proportion minus the ordinary shoe group's proportion).

   *Strength Shoe® group proportion of men:*

   *Ordinary shoe group proportion of men:*

   *Difference in proportions (strength − ordinary):*

7. Calculate and report the average height in each group. Also subtract these two averages (taking the Strength Shoe® group's average minus the ordinary shoe group's average).

   *Strength Shoe® average height:*

   *Ordinary shoe group average height:*

   *Difference in averages (strength − ordinary):*

8. Are the two groups identical with regard to both of these variables? Are they similar?

- Combine your results with those of your classmates by typing your two differences into the case table on the instructor's computer.

9. Sketch a plot of the differences in proportions of men.

10. Sketch a plot of the differences in average heights.

11. Do both plots appear to be centered around zero? Explain why this indicates that the randomization was effective.

## *Randomization Applet*

Now you will explore properties of randomization further by using a web applet to repeat this randomization process many more times.

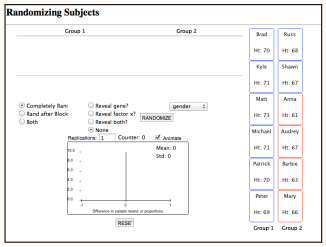
12. Use the *Randomization of Subjects* web applet to randomize (randomly assign) the 12 subjects to groups (see instructions in margin). Report the proportion of men in each group and the difference in proportions for this particular random assignment outcome.

13. Click `Randomize` again. Did you get the exact same assignment of subjects to groups? Is the difference in proportions of men the same as before?

14. Predict what the plot of differences in proportions will look like if you examined the distribution of the difference in proportions of men for many randomizations of the subjects into groups. In particular, where do you think that distribution will be centered? Explain.

***Randomization of Subjects* Web Applet**

- Open the URL `http://www.rossmanchance.com/applets` in your internet browser.
- Click the *Randomization of Subjects* link on the right-hand side of the page.
- Each of the 12 subject's name and height are recorded on a card on the right-hand side of the applet. These cards are also color-coded to indicate the subject's gender (males = blue; females = red).




- Click `Randomize` button. The applet will randomly assign the 12 subjects into the two groups.
- Click `Show Tables` to see the lists of names in the plots in the top part of the window.

15. Run 200 trials of the simulation (see instructions in margin). Where is this distribution centered? Is this what you predicted?

**Running Many Trials of the Randomization**

- Click the `Reset` button.
- Change the number of replications to *200*.
- Uncheck the `Animate` box.
- Click the `Randomize` button. The applet will display the distribution of the 200 differences in proportions of men between the two groups.

16. Does randomization always balance out the gender variable exactly between the two treatment groups in each randomization? Explain.

17. Does it tend to balance out the gender variable in the long run, after many trials? Explain.

18. Comment on the distribution of difference in height (see instructions in margin for changing summary measure information). Be sure to address whether or not randomization tends to balance out the heights between the two groups, and explain the evidence for your conclusion.

**Changing to another Summary Measure**

- Use the pull-down menu to change *gender* to *height*. The applet now displays the distribution of 200 differences in average heights between the two groups.

Now suppose that there are two more variables related to jumping ability that you had not considered or could not measure. These variables that were not measured are called *confounding variables*.

19. Would you expect randomization to balance out these variables between the two treatment groups as well? Explain.

20. Comment on what the plots reveal about whether or not randomization is effective for balancing unrecorded or unseen variables (see instructions in margin for displaying unrecorded or unseen variables).

**Displaying information for the Unrecorded or Unseen Variables**

- Click `Reveal Both`. The applet reports values for a categorical genetic trait and a quantitative $X$ factor.
- Using the pull-down menu, select each of these in turn.

21. Suppose you ended up with a randomization that included only one person with the gene assigned to group A. Would you be convinced something went wrong with the random assignment process? Explain based on your output from the previous problem.

22. Suggest a value for the difference between the two sample means of the *X-variable* that would convince you that something had gone wrong with the random assignment process. That is, how large would that difference need to be for you to be convinced that something other than "chance" created such an extreme difference between the two groups? Explain your choice.

23. Now suppose you conduct this random assignment and find that the Strength Shoe® group jumps substantially farther, on average, than the ordinary shoe group. Would you be comfortable concluding that the Strength Shoe® caused the increased jumping distance? Explain how you would argue that no confounding variable was responsible.

# *Random Assignment*

Experimenters try to assign subjects to groups so that lurking and potentially confounding variables tend to balance out between the two groups. The *Strength Shoe*® activity demonstrated that random assignment generally achieves its goal of creating groups that are similar in all respects except for the treatment imposed. Because of this equivalence, if the response variable turns out to differ substantially between the groups, you can attribute that difference to the difference in treatment conditions. Because of this, using random assignment has the potential to allow researchers to establish a cause-and-effect relationship between two variables.

To further help you understand how random assignment can be used to draw causal inferences, we would like you to read a short research report. We would also like you to read an excerpt from a research methods website that will further explain what it means for random assignment to create "identical" groups.

- Read the research report, *Random Assignment Evaluation Studies: A Guide for Out-of-School Time Program Practitioners*. The report is available at `http://www.childtrends.org/wp-content/uploads/2008/01/Random-Assigment-Evaluations.pdf`.

- Read the web excerpt, *Probabilistic Equivalence*. This excerpt is available at `http://socialresearchmethods.net/kb/expequi.php`.

# *Course Activity: Dolphin Therapy*

Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18–65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups.

Antonioli, C., & Reveley, M. A. (2005). Randomised controlled trial of animal facilitated therapy with dolphins in the treatment of depression. *British Medical Journal, 331*, 1–4.

Both groups engaged in the same amount of swimming and snorkeling each day, but one group did so in the presence of bottlenose dolphins and the other group did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed substantial improvement (reducing their level of depression) by the end of the study.

Is swimming with dolphins therapeutic for patients suffering from clinical depression?

***Discuss the following questions.***

1. For this study, specify the treatment variable and each of the possible treatment levels.

2. For this study, specify the response variable and each of the possible response categories.

The researchers found that ten of 15 subjects in the dolphin therapy group showed substantial improvement, compared to three of 15 subjects in the control group.

3. Organize these data/results (i.e., frequencies) into a $2 \times 2$ table.

| | **No Improvement** | **Improvement** | **Total** |
|---|---|---|---|
| **Control Group** | | | |
| **Dolphin Therapy** | | | |
| **Total** | | | |

4. Compute the percent of the overall sample (all 30 subjects) that improved.

5. Of the 15 subjects assigned to the dolphin therapy condition, what percent improved?

6. Of the 15 subjects assigned to the control condition, what percent improved?

7. Find the difference between the percentage of subjects assigned to dolphin therapy condition that improved and the percentage of subjects assigned to the control condition that improved.

8. Write a few sentences summarizing the results in the sample. This summary should include a summary of what the data suggest about: (1) the *overall improvement* of these depression subjects; (2) the differences between the two treatment groups; and (3) whether or not the data appear to support the claim that dolphin therapy is effective.

As you consider the next steps in the analysis, you should consider the following questions. Although you do not have to answer these questions explicitly, they are questions that have you reflect on the key statistical question that you are seeking to answer.

- The above descriptive analysis tells us what you have learned about the 30 subjects in the study. But can you make any inferences beyond what happened in this study?
- Does the higher improvement rate in the dolphin group provide convincing evidence that the dolphin therapy is effective?
- Is it possible that there is no difference between the two treatments and that the difference observed could have arisen just from the random nature of putting the 30 subjects into groups (i.e., the luck of the draw)?
- You cannot expect the random assignment to always create perfectly equal groups, but is it reasonable to believe the random assignment alone could have led to this large of a difference?

While it is possible that the dolphin therapy is effective, it is also possible that dolphin therapy is not more effective and the researchers were unlucky and happened to "assign" more of the subjects who were going to improve into the dolphin therapy group than the control group.

One way to examine this is to consider what you would likely see if 13 of the 30 people were going to improve (the number of subjects who improved in our sample) regardless of whether they swam with dolphins or not. If that is the case, you would have expected, on average, about six or seven of those subjects to end up in each group (the null model suggests this).

**The key statistical question is:** If there really is no difference between the therapeutic and control conditions in their effects of improvement, how unlikely is it to see a result as extreme or more extreme than the one you observed in the data just because of the random assignment process alone?

## *Modeling the Improvement Under the Null Model*

You will answer this question by using TinkerPlots™ to conduct a randomization test in order to replicate the results you could have gotten just because of the random assignment, *under the condition where dolphin therapy is not effective—the control and dolphin therapy conditions are equally ineffective at improving depression.*

- Open the *Dolphin-Therapy.tp* data set.
- Set up a model that will produce the fixed responses for the subjects under the null model. (If you have forgotten how to do this, refer back to the instructions in the *Sleep Deprivation* activity.).

- Add a linked stacks device to re-randomize the treatment condition labels. (If you have forgotten how to do this, refer back to the instructions in the *Sleep Deprivation* activity.)

- Run the model.

## *Plotting and Collecting the Results*

- Use TinkerPlots™ to plot the results for the trial.

- Collect the results from the trial (see instructions in margin).

**Plotting and Collecting Results**

- Drag a new plot into the workspace and place the *Condition* attribute on the *y*-axis and the *Improvement* attribute on the *x*-axis.
- Display the percentages for each cell.
- In the plot window, right-click on the percentage value for the improved patients in the *Dolphin Therapy* group and select `Collect Statistic`.
- Repeat for the *Control* group, again right-clicking on the percentage value for the improved patients.
- Create a third attribute in your *History of Results* case table and name it *Difference*.
- Right-click *Difference* and select `Edit Formula`. Set up a formula that computes the percent difference between the two conditions.

9. What percentage of subjects assigned to the dolphin therapy condition improved? What percent of subjects assigned to the control therapy condition improved? What is the difference in percents between these two groups?

10. Run another trial of the simulation. What percentage of subjects assigned to the dolphin therapy condition improved in this trial? What percent of subjects assigned to the control therapy condition improved? What is the difference in percents between these two groups?

- Carry out 100 randomized trials of the simulation in TinkerPlots™.

### *Evaluating the Observed Result*

- Plot the differences in percentages for the 100 simulated trials.

11. Sketch the plot below.

12. What are the cases in the plot? (Hint: Ask yourself what each individual dot represents.)

13. Where is the plot of the results centered (at which value)? Explain why this makes sense.

14. Report the approximate *p*-value (i.e., strength of evidence) based on the observed result.

It is useful to provide a qualitative description for the strength of evidence along with the quantification. While there are no hard-and-fast rules for gauging how strong the evidence is against the null model, the following guidelines can be used:

- A $p$-value above 0.10 constitutes little to no evidence against the null model.
- A $p$-value between 0.05 and 0.10 constitutes borderline/weak evidence against the null model.
- A $p$-value between 0.025 and 0.05 constitutes moderate evidence against the null model.
- A $p$-value between 0.001 and 0.025 constitutes substantial/strong evidence against the null model.
- A $p$-value below 0.001 constitutes overwhelming evidence against the null model.

15. Based on the $p$-value, how strong would you consider the evidence against the null model?

16. Based on the $p$-value, provide an answer to the research question.

17. Write a brief memo in which you report *all* of the pertinent results from the analysis (use the example write-up provided in the *Sleep Deprivation* activity as a guide).

# *Course Activity: Latino Achievement*

The Center for Immigration Studies at the United States Census Bureau has reported that despite shifts in the ethnic makeup of the immigrant population, Latin America—and Mexico specifically—remains this country's greatest source of immigrants. Although the average immigrant is approximately 40-years-old, large numbers are children who enroll in U.S. schools upon arrival. Their subsequent educational achievement affects not only their own economic prospects but also those of their families, communities and the nation as a whole.

Katherine Stamps and Stephanie Bohon studied the educational achievement of Latino immigrants by examining a random sample of the 2000 decennial Census data, a subset of which will be used. One interesting research question that has emerged from their research is whether there is a link between where the immigrants originated and their subsequent educational achievement.

Stamps, K., & Bohen, S. A. (2005). Educational attainment in new and established Latino metropolitan destinations. *Social Science Quarterly, 87*(5), 1225–1240.

Do immigrants from Mexico have lower average educational achievement scores than immigrants from other Latin American countries?

The data contains a random sample of 150 Latino immigrants, of which 116 are from Mexico and 34 are from other Latin American countries. The response variable is an educational achievement score, ranging from 1 to 100, in which higher values indicate higher levels of educational achievement.

***Discuss the following questions.***

1. For this study, specify the treatment variable and each of the possible treatment levels.

2. For this study, specify the response variable and each of the possible response categories.

- Open the *Latino-Achievement.tp* data set.
- Plot the achievement scores grouped by "treatment".

3. Sketch the plot.

4. Is the mean educational achievement level higher for immigrants from other Latin American countries? Calculate the difference in the means of the educational achievement levels.

5. Is the difference between the means large enough to convince you that immigrants from other Latin American countries have a higher mean educational achievement level than immigrants from Mexico? Why or why not?

6. What are other possible explanations that you can think of for the difference between the two means?

## *Modeling Achievement Under the Null Model*

You will conduct a randomization test using TinkerPlots™ to find out how likely it would be, assuming there is no difference between the two immigration groups in their average educational achievement level.

7. Describe the null model to be used to simulate data in this investigation.

- Set up a model that will produce the fixed responses (achievement scores) for the subjects under the null model.
- Add a linked stacks device to re-randomize the group labels.
- Run the model.

### *Plotting and Collecting the Results*

- Use TinkerPlots™ to plot the results for the trial.
- Collect the difference in mean achievement level between the two groups using the `Ruler` tool.

8. What is the mean educational achievement level of immigrants from Mexico for this single simulated trial? What is the mean educational achievement level of immigrants from Other Latin American countries for this single simulated trial? What is the difference in means between these two groups?

- Carry out 500 randomized trials of the simulation in TinkerPlots™.

### *Evaluating the Observed Result*

- Plot the differences in means for the 500 simulated trials.

9. Sketch the plot below.

10. What are the cases in the plot? (Hint: Ask yourself what each individual dot represents.)

11. Where is the plot of the results centered (at which value)? Explain why this makes sense.

12. Report the approximate $p$-value (i.e., strength of evidence) based on the observed result.

13. Based on the $p$-value, how strong would you consider the evidence against the null model?

14. Based on the $p$-value, provide an answer to the research question.

15. Can the researchers generalize the results to the population of all Latino immigrants in America? Why or why not?

16. Can the researchers attribute the difference in the average educational achievement level to the immigrant's home country (Mexico vs. Other)? Explain. If they can't, provide an alternative explanation for the differences.

17. Write a brief memo in which you report the pertinent results from the analysis. When reporting the results of a simulation study, pertinent information from the analysis that needs to be included is:

    - The type of test used in the analysis (including the number of trials);
    - The null model assumed in the test;
    - The observed result based on the data;
    - The $p$-value for the test; and
    - All appropriate inferences based on the $p$-value and study design.

# *Random Selection*

The Center for Immigration Studies at the United States Census Bureau has reported that despite shifts in the ethnic makeup of the immigrant population, Latin America—and Mexico specifically—remains this country's greatest source of immigrants. Although the average immigrant is approximately 40 years old, large numbers are children who enroll in U.S. schools upon arrival. Their subsequent educational achievement affects not only their own economic prospects but also those of their families, communities, and the nation as a whole.

Stamps and Bohen studied the educational achievement of Latino immigrants by examining a random sample of the 2000 decennial Census data. One interesting research question that has emerged from their research is whether there is a link between where the immigrants originated and their subsequent educational achievement. Specifically, the question is if there is a difference in the educational achievement of immigrants from Mexico and that of immigrants from other Latin American countries. During the in-class activity, *Latino Achievement*, you used a subset of Stamps and Bohen's data to examine this question.

## *Studies that Use Random Sampling*

One of the biggest differences between the study described in the *Latino Achievement* activity and previous studies you have examined in this unit is that the subjects used in the study

described in the *Latino Achievement* activity were *randomly sampled* from a larger population. In studies that employ random sampling, the random mechanism is in the selection of subjects. In the previous studies introduced in this unit, the random mechanism was in the assignment of the subjects to treatments or conditions. The use of random assignment or random sampling (or both!) directly impacts the inferences and conclusions that can be drawn.

The goal of studies that employ random sampling is also very different than the goal of studies that employ random assignment. With studies that employ random assignment, the goal is to draw cause-and-effect conclusions about a particular treatment. We can draw cause-and-effect conclusions from studies that have employed random assignment, because the possibility of alternative explanations, other than the treatment, can be ruled out (recall the *Strength Shoe*® activity). In studies with random sampling, the goal is not to draw cause-and-effect conclusions about a treatment, but rather to generalize a conclusion that was found in the sample data to the broader population from which that sample was drawn.

Consider the study described in the *Dolphin Therapy* activity. Here, the focus was on whether or not there was an effect of swimming with dolphins (the treatment) on the level of depression. Statistically, we wanted to determine the likelihood of the observed result (i.e., the difference in depression levels between the two conditions) given all the potential ways that the subjects could have been assigned to the two conditions. This, plus the initial random assignment to conditions, allows us to evaluate whether the dolphin therapy is truly effective. It does not allow us to say who the treatment is effective for. Is it all depressed people? All depressed people between the ages of 18-65? All depressed people between the ages of 18-65 who have a clinical diagnosis of mild to moderate depression?

## *Random Sampling*

A random sample is a sample in which the method used to choose the subjects from the broader population of interest is based on random chance. Although there are many types of random sampling, the term is often associated with *simple random sampling*. A simple random sample (or SRS) is a sample in which each subject has the same probability of being sampled as every other subject. One way to obtain a simple random sample is to sample the subjects or observational units *with replacement*. Consider drawing a sample of size three from the following five subjects,

$$\{\text{Jordan, Jonathon, Joey, Donnie, Danny}\}$$

Each of the subjects has a $\frac{1}{5}$ probability of being selected. Now, suppose the first subject selected is Donnie. If we remove Danny from the population (sampling without replacement), now each of the remaining subjects—Jordan, Jonathon, Joey, and Danny—have a $\frac{1}{4}$ chance of being the next subject selected. This does not fit the definition of a simple random sample. How does this change if we sample with replacement? After Danny is initially selected, his name is recorded and he is replaced into the population. Now, the probability of being the second subject selected is again $\frac{1}{5}$, meeting the criterion for a simple random sample.

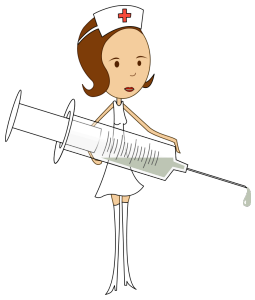
Of course, Danny (or another subject) might be selected multiple times. In theory this is required, but in practice, once a subject has been selected for a sample, she is usually no longer an eligible candidate to be re-sampled. In studies carried out in practice, sampling is without replacement. If the population is very large, the difference in probability for inclusion between subjects is small enough that it does not affect the analysis.

To make this concrete, consider a similar situation as to the one before, selecting three subjects, but this time the population size is 100,000. The probabilities for the three subjects are $\frac{1}{100,000}$, $\frac{1}{99,999}$, and $\frac{1}{99,998}$. These probabilities for all intents and

purposes are equivalent.

# *Course Activity: Murderous Nurse*

For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine.

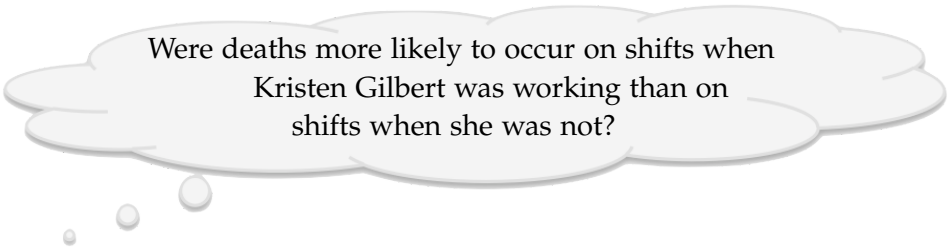


Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU. Here are the data presented during her trial:

Cobb, G. W., & Gehlbach, S. (2006). Statistics in the courtroom: United States vs. Kristen Gilbert. In R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern (Eds.), *Statistics: A guide to the unknown* (4th Edition), pp. 3–18. Duxbury: Belmont, CA.

| | **Gilbert Working On Shift** | **Gilbert Not Working On Shift** | **Total** |
|---|---|---|---|
| **Death Occurred On Shift** | 40 | 34 | 74 |
| **No Death On Shift** | 217 | 1350 | 1567 |
| **Total** | 257 | 1384 | 1641 |

You will use these data to answer the following research question:




### *Discuss the Following Questions*

1. Among all 1,641 shifts, what percentage of shifts had a death occur?

2. Among the 257 shifts when Gilbert was working, what percentage of shifts had a death occur?

3. Among the 1,384 shifts when Gilbert was not working, what percentage of shifts had a death occur?

4. Compute the difference between the percentage of shifts in which a death occurred when Gilbert was working and the percentage of shifts in which a death occurred when Gilbert was not working.

5. For this study, specify the treatment variable and each of the possible treatment levels.

6. For this study, specify the response variable and each of the possible response categories

7. Were shifts that Gilbert was working more likely to have a death occur than on shifts when she was not?

8. Does the difference in percentages convince you that Gilbert was giving lethal injections of epinephrine to patients? Why or why not?

9. What might other possible explanations be for the difference between the two percentages?

10. In what ways is this situation similar to other situations where you tested for a difference between two groups (e.g., the *Sleep Deprivation Study* activity, the *Dolphin Therapy Study* activity, the *Latino Achievement* activity)?

11. In what ways is this situation different from other situations where you tested for a difference between two groups?

## *Modeling the Chance Variation Under the Assumption of No Difference*

You will conduct a randomization test using TinkerPlots™ to find out how likely it would be, assuming there is no difference between the percent of shifts in which a death occurred when Gilbert was working and those in which she was not working.

- Open the *Murderous-Nurse.tp* data set.
- Set up a model that will produce the fixed responses (*Death*) for the "subjects" under the null model.
- Add a linked stacks device to re-randomize the group labels.
- Run the model.

### *Plotting and Collecting the Results*

- Use TinkerPlots™ to plot the results for the trial (see instructions in margin).
- Collect two results: (1) the percentage of deaths when Gilbert worked on a shift and (2) the percentage of deaths when Gilbert was not working a shift.
- Calculate the difference between the two percentages in the *History of Results* case table.
- Carry out 500 randomized trials of the simulation in TinkerPlots™ (see instructions in margin for speeding up the simulation).

**Plotting a Trial**

- Drag a new plot into the workspace and place the *Death* attribute on the *x*-axis and the *Shift* attribute on the *y*-axis.
- Vertically stack the points.

**Speeding Up the Simulation**

- Minimize all of the objects (sampler, results table, plot of the results) except for the collection window.
- Select the collection window and from the `Objects` menu select `Inspect Collection`.
- Uncheck the `Animation On` option.
- Close the inspector window.
- In the cards object, change the number of samples to collect to *499*.
- Click the `Collect` button.

### *Evaluating the Observed Result*

- Plot the differences in percentages for the 500 simulated trials.

12. Sketch the plot below.

13. What are the cases in the plot? (Hint: Ask yourself what each individual dot represents.)

14. Where is the plot of the results centered (at which value)? Explain why this makes sense.

15. Report the approximate $p$-value (i.e., strength of evidence) based on the observed result.

16. Based on the $p$-value, how strong would you consider the evidence against the null model?

17. Based on the $p$-value, provide an answer to the research question.

18. Can the results be generalized to the population of all 8-hour shifts at the hospital? Why or why not?

19. Can the difference in the percentage of deaths be attributed to the fact that Kristen Gilbert was working? Explain. If not, provide an alternative explanation for the difference in percentages.

20. Write a brief memo in which you report the pertinent results from the analysis. When reporting the results of a simulation study, pertinent information from the analysis that needs to be included is:

    - The type of test used in the analysis (including the number of trials);
    - The null model assumed in the test;
    - The observed result based on the data;
    - The $p$-value for the test; and
    - All appropriate inferences based on the $p$-value and study design.

# *Observational Studies*

One of the biggest differences between the analysis of the data from the *Murderous Nurse* activity and those from previous activities is that the data was not randomly sampled nor were the observational units randomly assigned to conditions—Kristen Gilbert was not randomly assigned to the 8-hour work shifts. When the cases/data are not assigned to conditions they are referred to as *observational data*, and studies in which subjects are not assigned by the researcher to conditions are called *observational studies*.

The purpose of an observational study is to describe some group or situation. The only difference lies in the inferences you can make.

You typically cannot draw cause-and-effect conclusions from observational studies, because the possibility of alternative explanations always exists. For example, since Kristen Gilbert was not randomly assigned to work the shifts that she did work, there is any number of other explanations for the increase in the number of deaths, aside from attributing it to the fact that Kristen Gilbert was working, that are possible and even plausible.

Observational studies may or may not incorporate random sampling. For example, the study described in the *Latino Achievement* activity was an observational study (the conditions of Mexico and Other were not assigned by the researchers) that incorporated random sampling. Because of the random sampling, the researchers could generalize the difference in educational achievement to the larger population of Mexi-

can and non-Mexican Latin American immigrants. However, because the conditions were not randomly assigned (in fact, they were not assigned at all) it is not appropriate to make cause-and-effect statements attributing that difference solely to the intrinsic characteristic of country of origin. It is much more likely that there is another reason that explains these differences (e.g., socioeconomic differences, prior education, etc.).

In 1988, results released to the public from the *National Household Survey on Drug Abuse* created the false perception that crack cocaine smoking was related to ethnicity. The analysis, which was based on observational data (researchers cannot assign race) showed that the rates of crack use among blacks and Hispanics were twice as high as among whites. The data were re-analyzed in 1992 by researchers from Johns Hopkins University to take into account social factors such as where the users lived and how easily the drug could be obtained. They found that after adjusting for these factors, there were no differences among blacks, Hispanics and whites in the use of crack cocaine.

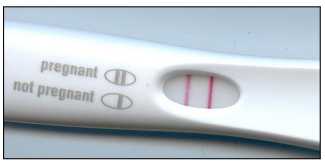
Lillie-Blanton, M., Anthony, J. C., & Schuster, C. R. (1993). Probing the meaning of racial/ethnic group comparisons in crack cocaine smoking. *Journal of the American Medical Association, 269*(8), 993-997.

To further help you understand the limitations of observational data, we would like you to read an excerpt from a research methods website.

- Read the web excerpt, *Observational Study*. This excerpt is available at `http://www.experiment-resources.com/observational-study.html`.

# *Course Activity: Pregnancy Tests*

Pregnancy tests have evolved greatly over the years. Many of the home pregnancy tests make strong claims. For example, the First Response Gold® Digital Pregnancy Test claims it can give "results as early as five days before the day of your missed period".




An important question would be

How accurate are the test results from the First Response Gold® Digital Pregnancy Test?

To answer this question, researchers conducted a clinical trial with 215 women who were trying to become pregnant. The women took the First Response Gold® Digital Pregnancy Test daily starting five days before the day of their expected period. We will use their data to learn of how to characterize the accuracy of the test.

Cole, L. A. (2011). The utility of six over-the-counter (home) pregnancy tests. *Clinical Chemistry & Laboratory Medicine*, *49*(8), 1317–1322.

The results for the tests that were taken five days before the day of their expected period appear in the table below.

| | Positive Pregnancy Result | Negative Pregnancy Result | Total |
|---|---|---|---|
| **Pregnant** | 58 | 79 | 137 |
| **Not Pregnant** | 4 | 74 | 78 |
| **Total** | 62 | 153 | 215 |

***Discuss the following questions.***

1. What is the percentage of women in the study that were pregnant?

2. What is the percentage of women in the study that were not pregnant?

## *Sensitivity and Specificity*

The two main measures of accuracy of a diagnostic test are known as *sensitivity* and *specificity*. The sensitivity of a test is the probability the test result is positive if the patient indeed has the disease/condition. The other measure, specificity, is the probability that the test result is negative when the patient does not have the disease/condition.

The sensitivity for the First Response Gold® Digital Pregnancy Test is the probability of testing positive with the First Response Gold® Digital Pregnancy Test for pregnancy when a patient actually is pregnant. Based on these data, the sensitivity is equal to

$$\frac{58}{137} = 0.423 \quad \text{or} \quad 42.3\%$$

3. Based on the above calculation, describe in words how to compute the sensitivity of the test.

The specificity for the First Response Gold® Digital Pregnancy Test is the probability of testing negative with the First Response Gold® Digital Pregnancy Test for pregnancy when a patient actually is not pregnant.

4. Describe in words how to compute the specificity of the test.

5. Find the specificity of the First Response Gold® Digital Pregnancy Test.

---


***Compare your answers to the above questions with another group.***

---

## *False Positives and False Negatives*

The two main measures of errors of a diagnostic test are known as *false positives* and *false negatives*. These errors refer to each of the possible situations that arise when the test gives a result that is wrong based on whether the patient actually has (or does not have) the disease or condition.

A false positive occurs when the test provides a positive result and the patient does not have the disease/condition. For example, a false positive would occur when the First Response Gold® Digital Pregnancy Test is positive, but the patient actually is not pregnant. The *false positive rate* of a test is the probability the test result is positive if the patient does not have the disease/condition.

6. Describe in words how to compute the false positive rate for the test.

7. Find the false positive rate for the First Response Gold® Digital Pregnancy Test.

The other important error measure is a false negative. A false negative occurs when the test provides a negative result and the patient has the disease/condition. For example, a false negative occurs when the First Response Gold® Digital Pregnancy Test is negative, but the patient actually is pregnant. The *false negative rate* is the probability the test result is negative if the patient does have the disease/condition.

8. Describe in words how to compute the false negative rate for the test.

9. Find the false negative rate for the First Response Gold® Digital Pregnancy Test.

10. Describe in words (or a diagram) how the false negative rate and the false positive rate relate to specificity and sensitivity.

11. Consider a woman that was not trying to get pregnant. Would the sensitivity or specificity be of more interest? Explain.

# *Learning Goals: Unit 2*

The activities, homework, and reading that you have completed for the second part of this course have introduced you to several more fundamental ideas in the discipline of statistics. The ideas and concepts you were introduced to in this unit form the basis for all formal statistical methods for testing hypotheses. In addition, you have learned how to use TinkerPlots™ to carry out a randomization test.

Below are the key concepts and skills from Unit 2 that you should have learned. At this point in the course, you should

## *Literacy/Understanding (Terms and Concepts)*

- Understand characteristics of distributions (e.g., shape, center variability) and be able to use them to describe and compare distributions of data, as well as, distributions of statistics.
- Understand basic terminology related to comparing groups (e.g., categorical data, quantitative data, factor, treatment variable, response variable, experiment, observational study, random assignment, random sampling, etc.)
- Understand the importance of random assignment in drawing inferences (cause-and-effect conclusions)
- Understand the importance of random sampling in drawing inferences (making generalizations)

- Understand that the scope of inferences that can be reached are based on the study design
- Understand the idea of a null model as constituting the "what if" distribution under the assumption of no effect of treatment or no group differences
- Understand the idea of a *p*-value as a quantification of the strength of evidence against a particular model
- Be able to compute the *p*-value by counting the number of results as extreme or more extreme than the observed result
- Understand ideas of how well a model works in making decisions (types of errors, specificity, and sensitivity)
- Be able to calculate sensitivity, specificity, a false positive rate, and a false negative rate given a contingency table of observed data

### *Selecting/Using Models*

- Be able to generate simulated data to produce the distribution of a numerical summary under the assumption of no effect of treatment or no group differences for randomly assigned or randomly sampled groups (distributions are centered at 0, no difference)
- Be able to generate simulated data from a model that includes two sampling devices

### *Evaluation*

- Be able to quantify the strength of evidence (p-value) of an observed result under the null model of no effect of treatment or no group differences
- Be able to assess the strength/degree of evidence against the null model (evaluate a *p*-value)

- Be able to draw appropriate inferences regarding cause and effect/generalization based on the study design
- Be able to evaluate importance of types of errors given a particular research context

## *TinkerPlots™ Skills*

You should also be able to do the following using TinkerPlots™.

- Properly create a sampler to model the random variation inherent in the assignment or sampling of two groups by
  - Using a counter to systematically reproduce the fixed outcome or responses
  - Using a random device (e.g., stacks, mixer) to randomly produce the two groups without replacement
- Create a single plot that separates the trial outcomes for the two groups
- Compute a numerical measure to summarize the difference between the two groups
- Collect the summary measure from many trials
- Plot the distribution of the summary measure
- Use the divider tool to compute the *p*-value

In the next activity, *Unit 2 Wrap-Up & Review*, you will have a chance to assess yourself on whether or not you have mastered these ideas through a variety of practice and extension problems. As a pre-cursor to this activity, you may want to review the readings and activities in Unit 2.

# *Course Activity: Unit 2 Wrap-up & Review*

## *Terminology for Unit 2*

1. At this point, you should be familiar with the following terms. Write down what each term represents as well as any notes that may help you remember.

   (a) Experiment

   (b) Observational Study

   (c) Factor/Treatment Variable

   (d) Response Variable

   (e) $p$-value

   (f) Null Model

   (g) Random Assignment

(h) Random Sampling

(i) Confounding Variables

(j) Randomization Test

(k) Sensitivity

(l) Specificity

(m) False Positive Rate

(n) False Negative Rate

2. For all studies described in Unit 2 consider whether the study was an experiment or observational study. For each study, identify the factor/treatment variable and response variable (and their levels).

| Study | Experiment/ Observational Study | Factor/Treatment Variable | Response Variable |
|---|---|---|---|
| **Memorization Study** | | | |
| **Sleep Deprivation Study** | | | |
| **Strength Shoe® Study** | | | |
| **Dolphin Therapy Study** | | | |
| **Latino Achievement Study** | | | |
| **Murderous Nurse Sudy** | | | |

3. Think about how your analysis and conclusions might have changed if you had subtracted the group means in the other direction (sleep deprived mean − unrestricted sleep mean).

   (a) What parts of your analysis would have been the same, and what parts (if any) would have turned out differently? How would they have been different (if at all)?

   (b) How would your conclusion about the study have changed (if at all)?

   (c) Investigate your predictions by making this change and re-conducting your analysis.

4. Investigate the effect that a single observation can have on this analysis.

   (a) Remove the improvement score of 45.6 from the *unrestricted sleep* group, and re-conduct the analysis. Comment on how much impact this one observation has on your analysis and conclusion.

(b) Restore the value of 45.6, but remove the improvement score of $-7.0$ from the *unrestricted sleep* group. Conduct the analysis. Comment on how much impact this one observation has on your analysis and conclusion.

(c) Investigate your predictions by making this change and re-conducting your analysis.

5. Notice that this research study involved slightly different numbers of subjects in the two groups. Suppose that you describe this study to a friend, and he argues that the study is invalid because of the unequal group sizes. Describe how you would respond to your friend, and be sure to include a description of how your analysis took these unequal group sizes into account.

## *Dolphin Therapy*

6. Suppose the results of the experiment had been that 11 subjects had improved in the dolphin therapy group (instead of ten) and only two subjects had improved in the control group (instead of three). Explain how your approximate *p*-value would have been different in this case. Also describe how the strength of evidence for the benefit of dolphin therapy would have changed.

7. Suppose the results of the experiment had been that eight subjects had improved in the dolphin therapy group (instead of ten) and five subjects had improved in the control group (instead of three). Explain how your approximate *p*-value would have been different in this case. Also describe how the strength of evidence for the benefit of dolphin therapy would have changed.

8. Suppose the study had involved twice as many subjects, 60 instead of 30, and suppose that the same proportion of subjects had improved in each group as in the original study. Describe what would have changed in how you set up the simulation analysis. Then make a prediction, and explain your reasoning, for how the approximate $p$-value, and the strength of evidence for the benefit of dolphin therapy, would have changed. Finally, conduct the simulation analysis for this new situation, and comment on whether your prediction was confirmed or refuted.

## *Teen Hearing*

Headlines in August of 2010 trumpeted the alarming news that nearly one in five U.S. teens suffers from some degree of hearing loss, a much larger percentage than in 1988. These headlines were based on a study that was described in the *Journal of the American Medical Association*. The findings were based on large-scale surveys of randomly selected American teenagers from across the United States: 2,928 teens in 1988–1994 and 1,771 teens in 2005–2006. The researchers found that 14.9% of the teens in the first sample (1988–1994) had some hearing loss, compared to 19.5% of teens in the second (2005–2006) sample.

Shargorodsky, J., Curhan, S. G., Curhan, G. C., & Eavey, R. (2010). Change in prevalence of hearing loss in US adolescents. *Journal of the American Medical Association, 304*(7), 772–778.

9. Describe (in words) the research question. Identify the factor/treatment and the response variables (and their levels) used in this study.

10. Just as with the *Dolphin Therapy* and *Sleep Deprivation* studies, this study made use of randomness in the its design. But the use of randomness was quite different in this study. Discuss what type of conclusions can be drawn from each type of study and why you can draw those conclusions for one study but not the other.

11. Are the percentages reported above (14.9% and 19.5%) population values or sample values? Explain.

12. Write out the null model for this analysis.

## *Mammography*

A mammogram is an X-ray of the breast. Diagnostic mammograms are used to check for breast cancer after a lump or other sign or symptom of the disease has been found. Routine screening is recommended for women between the ages of 50 and 74, but there are controversial findings about beginning the screening at earlier ages. In November 2009, the *Annals of Internal Medicine* reported these controversial findings, citing the large number of false positives. Gigerenzer reported the following data consistent with those seen in mammography screenings.

Mandelblatt, J. S., et al. (2009). Effects of mammography screening under different screening schedules: Model estimates of potential benefits and harms. *Annals of Internal Medicine, 151*(10), 738–747.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

| | **Positive Mammography Result** | **Negative Mammography Result** | **Total** |
|---|---|---|---|
| **Has breast cancer** | 70 | 90 | 160 |
| **Does not have breast cancer** | 700 | 9,140 | 9,840 |
| **Total** | 770 | 9,230 | 10,000 |

13. What percentage of women in this study have breast cancer?

14. Describe in words what the *sensitivity* is using the context of this problem.

15. Find the sensitivity of the mammography test.

16. Describe in words what the *specificity* is using the context of this problem.

17. Find the specificity of the mammography test.

18. If a patient tests positive for breast cancer, the patient may experience extreme anxiety and may have a biopsy of breast tissue for additional testing. If patients exhibit the symptoms of the disease but tests negative for breast cancer, this may result in the patient being treated for a different condition. Untreated cancer can lead to the tumor continuing to grow or spread.

    (a) Given the consequence of a false test result, is the sensitivity or specificity more important in this case? Explain

    (b) Find the *false positive rate* of the mammography test for breast cancer.

    (c) Find the *false negative rate* of the mammography test for breast cancer.

### *Native Californians*

Out of people living in California, is the percentage born in California different in the years 1950 and 2000? To investigate this question, a random sample of 500 California residents was drawn using data from the 1950 Census and another random sample, independent from the 1950 random sample, of 500 California residents was drawn using data from the 2000 Census. The results are shown in the table below

| | 1950 | 2000 | Total |
|---|---|---|---|
| **Born in California** | 219 | 258 | 477 |
| **Not born in California** | 281 | 242 | 523 |
| **Total** | 500 | 500 | 1,000 |

19. Compute the difference between the percentage of native California residents in the years 1950 and 2000

20. Describe how to carry out a simulation to investigate the difference in percentage of native California residents in the years 1950 and 2000.

21. Carry out a simulation using TinkerPlots™ to answer the research question: Out of people living in California, is the percentage born in California different in the years 1950 and 2000?

### *Blood Pressure*

In a 2001 study, volunteers with high blood pressure were randomly assigned to one of two groups. In the first group—the talking group—subjects were asked questions about their medical history in the minutes before their blood pressure was measured. In the second group—the counting group—subjects were asked to count aloud from one to 100 four times before their blood pressure was measured. The data presented here are the diastolic blood pressure (in mm Hg) for the two groups. The sample average diastolic blood pressure for the talking group was 107.25 mm Hg and for the counting group was 104.625 mm Hg.

| ***Talking*** ($n = 8$) | ***Counting*** ($n = 8$) |
|---|---|
| 103 | 98 |
| 109 | 108 |
| 107 | 108 |
| 110 | 101 |
| 111 | 109 |
| 106 | 106 |
| 112 | 102 |
| 100 | 105 |

Data for the Blood Pressure study.

22. Do the data in this study come from a randomized experiment or an observational study? Explain.

23. Calculate the *difference* in the means of the two groups.

24. Write out the null model for this study.

25. Use TinkerPlots™ to carry out the appropriate analysis to determine if a difference this large could reasonably occur just by chance. Comment on whether the difference in the means is statistically significant; in other words, providing strong evidence against the null model.

### *Social Fibbing*

A student investigated "social fibbing" (the tendency of subjects to give responses that they think the interviewer wants to hear) by asking students "Would you favor a policy to eliminate smoking from all buildings on campus?" She randomly assigned half the subjects to be questioned by an interviewer smoking a cigarette and the other half were interviewed by the same student but not while she was smoking. The results are displayed in the following table.

| | **Favor ban** | **Do not favor ban** | **Total** |
|---|---|---|---|
| **Smoking** | 43 | 57 | 100 |
| **Not smoking** | 79 | 21 | 100 |
| **Total** | 122 | 78 | 200 |

26. Does the behavior of an interviewer affect the responses of the people being surveyed? Use TinkerPlots™ to carry out the appropriate analysis to determine if a difference this large could reasonably occur just by chance. Comment on whether the difference in the percents is statistically significant (i.e., provides strong evidence against the null model).

27. Calculate the *difference* in the means of the two groups.

28. Write out the null model for this study.

29. Use TinkerPlots™ to do the appropriate test to determine if a difference this large could reasonably occur just by chance. Comment on whether the difference in the means is statistically significant (i.e., it provides strong evidence against the null model).

# *Unit III: Sampling & Estimation*

*An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*
—John Tukey (2002)

One of the largest aspects of statistical inference is the estimation of unknown parameters using sample data. Polling companies such as Gallup and Harris have made billions of dollars by using statistical estimation to carry out public opinion research. These companies are hired to provide snapshots of public attitudes and opinions on varied topics from politics and the economy, to social awareness and health and well-being. The results of their polls are seen on a daily basis in almost every newspaper, news blog and website the world over.

Statistical estimation is used by more than pollsters. Biologists, social scientists, and medical researchers use statistical sampling and estimation to quantify populations. For example, the Department of Natural Resources estimates the size of animal, bird, and fish populations yearly. These estimates are used to help set hunting and fishing regulations, as well as to allocate resources.

One of the largest uses of statistical sampling is in the estimation of at-risk populations. Public health advocates and researchers use statistical sampling to estimate the size of the population at risk for particular disease or behavior. For example, Boynton Health Services has sampled undergradu-

ate and graduate students at colleges and universities across Minnesota to estimate the proportion of students engaging in at-risk behaviors such as drug or alcohol use. The World Health Organization and the Joint United Nations Programme on HIV/AIDS have used sampling to estimate the size of the population at risk for HIV around the world.

World Health Organization & UNAIDS. (2011). *Guidelines for estimating the populations most at risk for HIV.* Le Mont-sur-Lausanne, Switzerland: Author.

Statistical sampling has also been used to obtain estimates of groups and populations that are conventionally very difficult to enumerate. For example, researchers at the University of California, Los Angeles have recently used sampling methods to estimate the size of the lesbian, gay, bisexual, and transgender community in the United States. Human rights activists have also used statistical sampling in estimating the size of difficult-to-count populations. For example, statistical estimates have been used throughout history to estimate the number of victims of genocides and political conflict.

Gates, G. J. (2012). LGBT identity: A demographer's perspective. *Loyola of Los Angeles Law Review, 45*(3), 693–714. http://digitalcommons.lmu.edu/llr/vol45/iss3/2

Asher, J., Banks, D., & Scheuren, F. J. (2007). *Statistical methods for human rights.* New York: Springer.

Sampling and estimation, however, are not without controversy. The United States Census Bureau has a constitutionally mandated task of quantifying the population of the United States every ten years. For the 2000 Census, the Census Bureau proposed the use of statistical sampling to address a chronic and growing problem of "undercounting" some identifiable groups, including certain minorities, children, and renters. Because the results of the Census are used to re-draw legislative boundaries and apportion the number of representatives to State and Federal Congress, the use of sampling was challenged in a lawsuit that made its way to the Supreme Court.

For more information on the constitutional requirements of the Census, see http://www.usconstitution.net/consttop_cens.html.

To read more about the challenge to statistical sampling in the use of the United States Census and the Supreme Court's decision, see http://www.civilrights.org/monitor/winter_spring1999/art2p1.html.

## *Outline of the Unit*

In this unit, you will begin exploring ideas related to statistical sampling and estimation.

In the first activity, you will be introduced to two characteristics of sample estimators—bias and precision—that allow for reasonable inferences and estimates. You will also learn how different sampling methods affect these characteristics.

Throughout the remainder of the unit, you will explore these two characteristics more deeply.

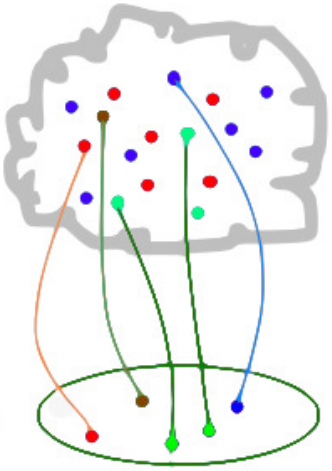
In the second activity, you will learn how to quantify variation in a distribution by computing the standard deviation. The quantification of variability through the standard deviation will provide you with another numerical summary to better describe distributions of data. It will also play a very important role in the third activity, in which you will make use of the standard deviation to quantify the precision of an estimator. This is important because it provides a numerical summary of the uncertainty in an estimate that is due to sampling, called sampling error.

After learning how to quantify the precision in an estimate, you will learn how statisticians use this measurement to provide a margin of error and interval estimates. You will also learn a method called the bootstrap, that will allow you to obtain estimates of the sampling error in order to estimate an unknown population parameter from a single observed sample of data.

The idea of the bootstrap in estimating the amount of sampling uncertainty will be developed further in an activity where you will also learn how to quantify the size of an effect, or effect size, between two groups. This is a natural extension of the ideas and group comparisons you experienced in the second unit. Here, you will learn how to answer the follow-up question to "are the groups different?", which is, "how different are the groups?"

Lastly, you will experience an activity in which you will gain a deeper understanding of how and why statisticians and researchers use the interval estimate produced by $\pm 2SE$. You will also learn about the connections between theoretical sampling from a population and empirical bootstrap sampling. Throughout the unit, you will continue to add to the knowledge you accumulated from Unit 1 and Unit 2, including both content knowledge and TinkerPlots™ knowledge.

# *Course Activity: Sampling*



In statistics, estimation refers to the process by which one makes inferences about a population or model, based on information obtained from a sample. In practice, it is often impossible to examine every unit of the population, so data from a subset, or sample, of the population is examined instead. The sample data provides statisticians with the best estimate of the exact "truth" about the population. The "truth" one is searching for in the population is typically a summary measure such as the population mean or population percentage. Summary measures of a population are called *parameters*. The estimates of these values from sample data are referred to as *statistics*.

Consider taking a sample of ten students from this class for the purpose of estimating the average number of credits an EPSY 3264 student takes per term.

***Share and discuss your responses to each of the following questions with your group.***

1. Describe at least two different ways you can choose a sample of students.

2. Share your two methods of sampling from question 1 with a group near you. Compare and contrast the different methods. Are some better or worse than others? Why or why not?

3. Choose a sampling method that you think is good based on your answer to Question 2. Using this sampling method, would you expect two different samples of students to yield the same estimate for the average number of credits? Why or why not?

4. Using that same sampling method, would your estimate of the average number of credits be a good estimate of the true average number of credits taken by EPSY 3264 students in this class? Why or why not?

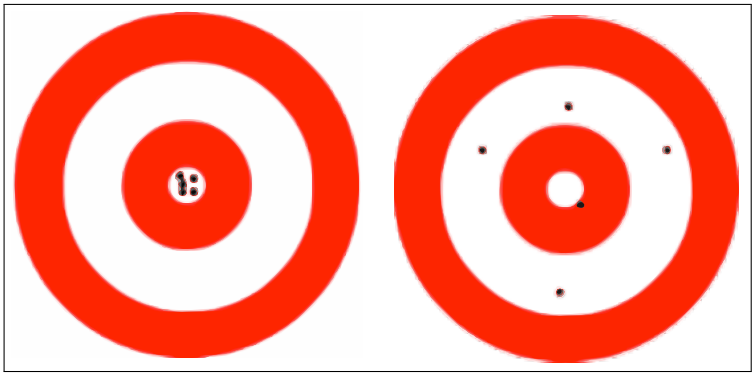
5. Using that same sampling method, would your estimate of the average number of credits be a good estimate of the true average number of credits taken by EPSY 3264 students in all sections this semester? Why or why not?

6. Using that same sampling method, would your estimate of the average number of credits be a good estimate of the true average number of credits taken by EPSY 3264 students in the last five years? Why or why not?

When estimating a parameter for an unknown model, there are several qualities that are ideal to have. Two of those qualities are *unbiasedness* and *precision*. Both of these qualities describe the estimation or sampling method used. You will examine unbiasedness in this activity and precision in upcoming activities.

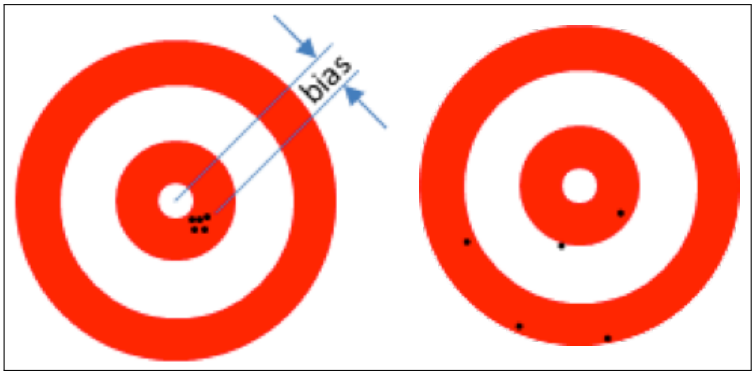
## *Unbiasedness*

Unbiasedness is a quality that indicates that the estimation method used produces a distribution of the estimated parameter that is neither systematically too large nor too small. To illustrate this, consider the following two targets which show the locations of five darts thrown overhanded (target on the left) and underhanded (target on the right).



Both throwing methods, under- and overhanded, would be unbiased. If you examine the set of throws as a whole on the target on the left, they "average out" to be on center. Now examine the throws on the target on the right. Again, even

though none of the darts thrown hit the center exactly, as a whole, the five darts "average out" to have "hit" the center. Now compare this with the targets below in which the darts were thrown under- and overhanded while the thrower had closed her eyes. (Not a good idea when throwing darts!)




In both of these targets, the throwing method would be considered biased. On "average", the throws did not hit the center of the target. It is important to note that in examining the dart throwing methods, you used the distribution of throws to judge whether or not the throwing method was unbiased. Similarly, in judging whether an estimation or sampling method is unbiased, you will have to examine a distribution of the estimates produced using that method.

Does the sampling method used impact whether the estimation is unbiased?

To help answer this research question, you are going to compare two different sampling methods using the population of 268 words in the passage on the following page. The passage is, of course, Lincoln's Gettysburg Address, given November 19, 1863 on the battlefield near Gettysburg, PA.

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.

It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

The goal in many studies is to provide information about some characteristic of a population. For example, you may want to say something about the percentage of Americans who would support a particular piece of legislation. Or, you may want to provide information about the average amount of time University of Minnesota students take to graduate. One potential solution to obtain such information would be to collect the necessary data from every member of the target population.

In many studies, however, it may not be feasible given time and money constraints to collect data from each member of

the population. In these cases it is only possible to consider data collected for a smaller subset, or *sample* from that population. In these cases, the characteristic of the population would be estimated from the sample data and inferences would be drawn about the population. The key is then to carefully select the sample so that the results estimated from the sample are representative of the characteristic in the larger population.

> The population is the entire collection of who or what (e.g., the observational units) that you would like to draw inferences about. A sample is a subset of observational units from the population. A model is a simplified representation of the population.

7. Circle ten words in the text of the Gettysburg Address such that the ten words you select constitute a representative sample (i.e., have the same characteristics) of the entire passage.
8. Describe how the ten words in your sample are representative of the 268 words in the population.

9. Record the length (number of letters not including punctuation) for each of the ten words in your sample:

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

10. Determine the average (mean) word length for your ten words. This sample average is an estimate of the average word length in the population.

---


***Add your sample estimate to the case table on the instructor's computer.***

---

11. Sketch the plot of all of the sample estimates. Make sure to label the axis appropriately.

12. The actual population average word length based on all 268 words is 4.3 letters. Where does this value fall in the above plot? Were most of the sample estimates around the population mean? Explain.

13. For how many groups in your class did the sample estimate exceed the population average? What proportion of the class is this?

14. Based on your answer to question 13, is the sample estimate just as likely to be above the population average as it is to be below the population average?

When the sampling method produces characteristics of the sample that systematically differ from those characteristics of the population, you say that the sampling method is biased. To try to eliminate potential biases, it is better to take a random sample. This should create a representative sample, no matter what variable is focused on. Humans are not very good "random samplers," so it is important to use other techniques to do the sampling for us.

### *Simple Random Sampling*

A *simple* random sample (SRS) is a specific type of random sample. It gives every observational unit in the population the same chance of being selected. In fact, it gives every sample of size $n$ the same chance of being selected. In this example you want every possible subset of ten words that could be sampled to have the same probability of being selected.

The first step in drawing a simple random sample is to obtain a *sampling frame* or a list of each member of the population. Then, you can use software to randomly select a sample from the sampling frame.

*Use TinkerPlots™ to Draw a SRS*

- Open the file *Gettysburg.tp*.
- Draw a simple random sample of ten words from the sampler.

15. Record the words and their lengths:

| | Word | Length |
|---|---|---|
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |

- Use TinkerPlots™ to find the length of each word in your sample (Hint: use the `stringLength()` function), and determine the mean word length for your ten words.

16. Record the mean word length for your ten randomly sampled words. Remember, your sample mean is an estimate of the average word length in the population.

---


***Add your sample estimate to the case table on the instructor's computer.***

---

17. Sketch the plot of all of the sample estimates from this sampling method. Make sure to label the axis appropriately.

18. This time how many students in your class obtained a sample average that was longer than the population average? What proportion of the class is this?

19. If the sampling method is unbiased the estimates of the population average should be centered "around" the population average word length of 4.295. Does this appear to be the case?

*Examining the Sampling Variation*

To really examine the long-term patterns of this sampling method on the estimate, you will use TinkerPlots™ to take many, many samples.

- Collect the mean word length for each random sample of ten words.
- Use TinkerPlots™ to draw 500 random samples of 10 words.
- Place a vertical reference line at the population average word length on the plot of the 500 sample estimates (see instructions in the margin).

**Putting a Vertical Reference Line on a Plot**

- Highlight the plot of the 500 sample estimates.
- Click on the vertical `Reference Line` button in the upper plot toolbar.


- Click and drag the vertical reference line to the population average word length value.

20. Sketch the plot of the sample estimates based on the 500 samples drawn. Make sure to label the axis appropriately.

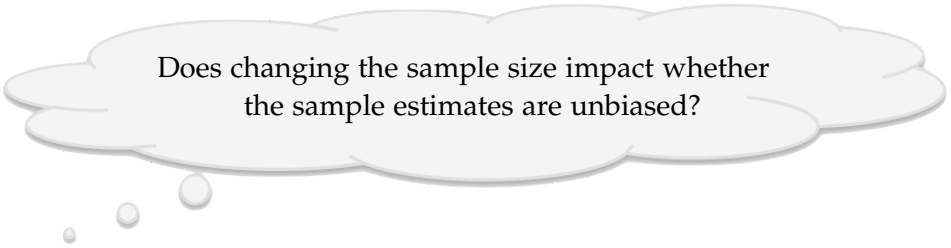
21. Record the typical (average) value for the estimate of the average word length.

22. Based on the TinkerPlots™ results from drawing many samples using simple random sampling, is this sampling method unbiased? Explain.

### *Sample Size*

Even when an unbiased sampling method, such as simple random sampling, is used to select a sample, you do not expect the estimate from each individual sample drawn to match the population average exactly. You should see, however, that the estimates are just as likely to over- or underestimate the population parameter. Because of this predictability to the variation in the possible sample estimates, inferences drawn about the population are said to be valid.

On the other hand, if the sampling method is biased, any inferences made about the population based on a sample estimate may not be valid. In such cases the estimate of the parameter is more likely to be too large or too small compared to the parameter. It is therefore very important to determine how a sample was selected before believing inferences drawn from sample results.




- Change the sample size from 10 to 25.
- Use TinkerPlots™ to draw 500 random samples of 25 words, and collect the average word length for each sample.

23. Sketch the plot of the sample estimates based on the 500 samples drawn. Make sure to label the axis appropriately.

24. Record the average value for the estimate of the average word length.

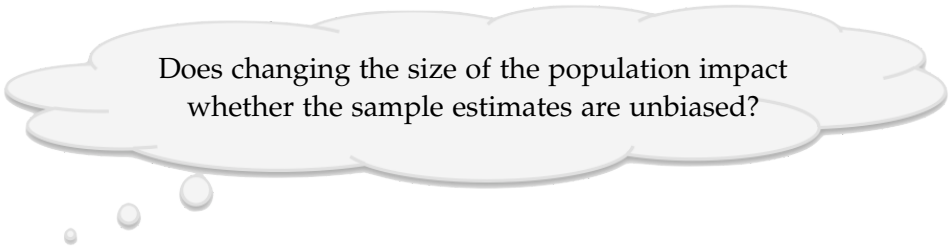
25. Does the sampling method still appear to be unbiased? Explain.

26. Compare and contrast the distribution of sample estimates for $n = 10$ and the distribution of sample estimates for $n = 25$. How are they the same? How are they different?

27. Using the evidence from your simulations, answer the research question: does changing the sample size impact whether the sample estimates are unbiased?

## *Population Size*

It is clear that changing the size of the sample does not affect whether or not an unbiased estimate is produced. Now we examine another question:




One common question is how the size of the population affects the estimates. To examine this you will increase the size of the population in our TinkerPlots™ sampler. Please see instructions in the margin for how to quadruple the size while keeping the population characteristics the same.

- Change the population size from 268 to 1072.
- Use TinkerPlots™ to draw 500 random samples of 25 words, and collect the average word length for each sample.

28. Sketch the plot of the sample estimates based on the 500 samples drawn. Make sure to label the axis appropriately.

**Adapting the Population Size in the Model**

- Highlight all of the elements (i.e., words) in the sampler.
- Under the `Edit` menu, select `Copy Cases`.
- Then select `Paste Cases` (also in the `Edit` menu).
- Paste two more times, for a total of three pastes.
- Now the population in the sampler consists of four copies of the Gettysburg Address ($4 \times 268 = 1072$ words) so that it is four times larger than it used to be (but, the population characteristics are the same).

29. Record the average value for the estimate of the average word length.

30. Does the sampling method still appear to be unbiased? Explain.

31. Compare and contrast the distribution of sample estimates for $n = 25$ now that you are sampling from a larger population to the distribution of sample estimates for $n = 25$ from before. How are they the same? How are they different?

32. Use the evidence collected from the simulation to answer the research question: does changing the size of the population impact whether the sample estimates are unbiased?

A rather counterintuitive, but very crucial, fact is that when determining whether or not a sample estimate produced is unbiased the size of the population does not matter! Even more counterintuitive might be that the precision of the sample estimate is unaffected by the size of the population! This is why organizations like Gallup can state poll results about the entire country based on samples of just 1,000–2,000 respondents as long as those respondents are randomly selected.

In summary, it is important to note three caveats about random sampling:

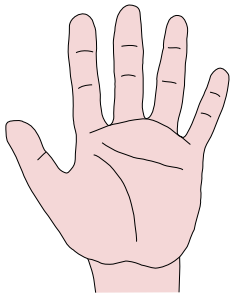
- One still gets the occasional "unlucky" sample whose results are not close to the population even with large sample sizes.
- Second, the sample size means little if the sampling method is biased. As an example, in 1936 the Literary Digest magazine had a huge sample of 2.4 million people, yet their predictions for the Presidential election did not come close to the truth about the population

See http://en.wikipedia.org/wiki/The_Literary_Digest#Presidential_poll for more information about the sampling and the 1936 election.

- While the role of sample size is crucial in assessing how close the sample results will be to the population results, the size of the population does not affect this. As long as the population is large relative to the sample size (at least ten times as large), the precision of a sample statistic depends on the sample size but not on the population size!

# *Course Activity: Comparing Hand Spans*

In this activity, you will learn about the standard deviation, a common measure of variability.



How can you quantify variability and summarize it into a single measure?

- Measure and record the hand span for each person in your group.
- Enter the data into a TinkerPlots™ case table.
- Create a plot of the hand spans for your group.
- Find the mean hand span for your group on the plot.

1. Record this mean.

2. Sketch the plot below. Label the axis and specify the units of measurement. Write names or initials above the dots to identify each case and circle the mean value.

3. What are two possible factors that would explain the variation in the measurements? That is, give two reasons why the measurements are not all the same.

4. How far is your hand span from that of the mean of your group?

5. How far from the mean are the hand spans of the others in your group?

## *The Standard Deviation*

Recall that the mean is a single number that can be used to summarize the data. In this context, it is a description of the typical hand span measurement for your group. Of course, not every student in the sample is at the typical value (in fact all of them might be different from the typical value). Thus, it is also useful to have a single number description of how different the data tends to be from this typical value.

One single number description of the variability in a sample of data is called the *standard deviation* or *SD*. If the word "typical" is substituted for the word "standard" in its name, the name standard deviation (typical deviation) makes more sense. This measure quantifies variability by determining how far data cases typically deviate from the mean value.

- Use TinkerPlots™ to create a new attribute in the table, called *Deviations*, that contains the difference between the observed data (hand spans) and the mean of your group member's hand spans. Use a formula to compute this difference (you can compute these by subtracting the mean from each observation).
- Create a plot of the *Deviations* attribute.

6. Sketch the plot below. Make sure to label the axis. Circle the mean and record its value.

7. How does the distribution of deviations compare with the distribution of hand span lengths you created in Question 2?

8. How would you interpret the values of the *Deviations* attribute?

One thing you may not have known about the mean is that it is the value that "balances" the data. In other words, the mean is the value that gets the deviations to sum to zero. This is useful when describing a typical value of the data (it is the "closest" point to all of the cases, on average). If you try to average these deviations, however, you will always get zero. This is not very useful in summarizing variation in a data set, nor in comparing the variation between two data sets. One way to alleviate this problem is to square each of the deviations before you add them together.

- Create another attribute, *SquaredDeviations*, which contains the squared values of the deviations.
- Create a plot of *SquaredDeviations*.

9. Sketch the plot below. Make sure to label the axis. Circle the mean and record its value.

10. How does the plot of the squared deviations compare to the first two plots (Question 2 and Question 6)?

Because the deviations have been squared, the mean represents the typical squared deviation.

11. Find the square root of this value.

12. What does this new value represent (interpret its value)?

13. Now use TinkerPlots™ to find the standard deviation of the original data directly. You can use the `stdDev()` function to compute this value. The value will be similar, albeit higher, than the value you obtained in Question 11. (For those of you that want a challenge, try to determine why these values are different.)

### *Using Both the Mean and Standard Deviation: A Complete Summary*

The file *Study-Hours.tp* contains responses from 100 EPSY 3264 students who responded to the survey question, *how many hours per week do you typically study?* These students' responses are a random sample from all responses obtained from all in-class sections taught from 2004–2010.

14. Find the mean of these data.

15. Using the `stdDev()` function, find the standard deviation of these data.

16. Use the mean and standard deviation to provide an interval estimate to answer the question, *how many hours a week do students study?*

17. Describe the population of students to which these sample estimates (mean and standard deviation) apply.

# *Understanding the Standard Deviation*

To help provide you with a deeper understanding of ideas related to the standard deviation, we would like you to complete a short online tutorial from *Usable Stats*.

- Complete the online, *Standard Deviation Tutorial*. The tutorial is available at `http://www.usablestats.com/tutorials/StandardDeviation`.

# *Course Activity: Kissing the 'Right' Way*

A German bio-psychologist, Onur Güntürkün, was curious whether the human tendency for right-sightedness (e.g., right-handed, right-footed, right-eyed), manifested itself in other situations as well. In trying to understand why human brains function asymmetrically, with each side controlling different abilities, he investigated whether kissing couples were more likely to lean their heads to the right than to the left . He and his researchers observed 124 couples (estimated ages 13 to 70 years, not holding any other objects like luggage that might influence their behavior) in public places such as airports, train stations, beaches, and parks in the United States, Germany, and Turkey, of which 80 leaned their heads to the right when kissing.

What percentage of couples lean their heads to the right when kissing?

Güntürkün, O. (2003). Human behaviour: Adult persistence of head-turning asymmetry, *Nature, 421*, 711.

***Discuss the following questions.***

1. Based on the data collected, provide a single number estimate to offer an answer to the research question.

2. Consider another study carried out using the same methodology, but using a different sample of 124 couples. Would you necessarily obtain the same answer to the research question (i.e., would the percentage of couples who lean their heads to the right when kissing be the same)? Explain why or why not.

3. Now imagine the same study were carried out 10 more times, each time using a different set of 124 couples from the same population. Make a conjecture about the percentage of couples from each of these studies who lean their heads to the right when kissing. Write these values down.

Remember, summary measures that describe a sample, like the mean and standard deviation, are called statistics. The standard deviation of a distribution of statistics, when each case is a statistic, is referred to as a *standard error*. The standard error is a measure of the variation due to random sampling or random assignment.

- Enter these values into TinkerPlots™ and compute the standard deviation.

4. Is this value a standard deviation or a standard error? Explain.

5. What is the source of the variation in these values (i.e., what is the reason they are different)?

## *Precision of the Estimate*

Each of the imagined studies would be interested in answering the same research question, namely what percentage of couples lean their heads to the right when kissing. Because they used different samples of couples, those studies might provide different estimates of the population percentage. For example, in the observed data, 80 of the 124 couples in the sample, or 65%, leaned to the right when kissing. In another study the researchers might have found that 77 of the 124 couples, or 62% of the couples, leaned to the right when kissing.

When estimates like these are made from sample data, they often include an additional measure of the precision of the estimate. The *precision of the estimate* is an acknowledgement of the fact that sample estimates will vary from sample to sample. To determine the precision of an estimate due to random sampling, the question that needs to be answered is, how variable are the different possible estimates when different random samples are used to make this estimate? This variability, as you learned in the previous activity, is captured by the standard error.

## *Modeling the Variation Due to Random Sampling*

Before you can compute a standard error for the estimate of the true proportion of couples leaning their heads to the right when kissing, you need to be able to draw many different samples of size 124. A major obstacle is that you do not have access to the population. You also do not have a model from which you can generate simulated data. (Note that if you had either of those two things, you would not need to estimate the proportion of couples leaning to the right when kissing, you could just determine what it was.)

What you have is the observed data. Without any other evidence as to what the true model is, the most informed choice is to *use the observed data as a stand-in*, or proxy, for the unknown model. This stand-in model can then be used to generate simulated data that represent many samples of 124 couples.

As you have witnessed throughout the course, a model generates many different samples of data. One major problem in substituting the observed data as the model from which you will generate simulated data is that the observed data values are fixed. Sampling from these values will generate the exact same values! This will not allow you to estimate the variation in the estimate across samples because the sample estimates will always be the exact same value!

In order to use the observed data as a model from which to generate simulated data without getting the exact same values, you need to sample from the observed data *with replacement*. This means that the same value can be sampled multiple times. The process of using the observed data as a stand-in for the unknown model and generating data by sampling with replacement is called *nonparametric bootstrapping*.

## *Nonparametric Bootstrapping Using TinkerPlots™*

- Set up a model in TinkerPlots™ based on the results in the observed data (80 out of 124 couples lean to the right) to generate simulated data for 124 couples.
- Plot the results for the trial.
- Collect the percentage of couples leaning to the right.
- Carry out 500 trials of the simulation.

### *Evaluating the Bootstrap Distribution*

6. Plot the results from the 500 trials and sketch the plot below. Make sure to label the axis. Circle the mean and record its value.

7. Where is this distribution centered? Explain why it makes sense that the distribution is centered at this value.

8. Compute the standard error based on this simulation.

9. What does this value represent (interpret its value)?

## *Margin of Error*

Another acknowledgment that is often made when using sample data to estimate a model is that the true parameter is likely to be different from the sample estimate simply because of random sampling. In other words, there is uncertainty in the estimate due to random sampling. It turns out that the amount of uncertainty there is in an estimate because of random sampling, is exactly the same as the variability that exists between estimates from different random samples, the standard error.

Consider the following poll reported in the New York Times on June 30, 2011:

> As the housing market slumped over the last few years with a speed and magnitude not seen since the Great Depression, aspects of homeownership have been debated as never before. There are tough questions about the role the government should take. . . includ[ing] how much of a down payment lenders should demand. Whether buyers need to come up with a 20 percent down payment—the standard for decades, but beyond the reach of many families now—is hotly debated. Fifty-eight percent of respondents say lenders should require this, while 36 percent say they should not. The nationwide telephone poll was conducted June 24–28 with 979 adults and has a margin of sampling error of plus or minus three percentage points for all adults.

In polls reported in the newspaper and online, the margin of error is almost always provided. For example, in the newspaper article above, the margin of error for the sample estimate (the percentage of all adults in the United States who believe that lenders should require a 20% down payment on a house) is given as $\pm 3\%$. In studies reported in journals, however, the margin of error may not be reported. Luckily, the margin of error can easily be computed as,

$$2 \times SE$$

10. Using the standard error from the kissing study, compute the margin of error.

## *Interval Estimates*

When statisticians report sample estimates, they provide the value of the estimate along with the quantification of the variation in the estimate expected from random sampling. As indicated previously, in popular media this is often reported as the sample statistic and a margin of error. For example, in the newspaper article above, the percentage of all adults in the United States who believe that lenders should require a 20% down payment on a house was reported as $58\% \pm 3\%$.

The same information can also be reported by adding and subtracting the value of the margin of error to the sample estimate and providing the actual interval for the estimate. For example, in the newspaper article above, the percentage of all adults in the United States who believe that lenders should require a 20% down payment on a house was reported as (55%, 61%).

Statisticians refer to this as an *interval estimate* because it gives an interval of plausible values for the percentage of all adults in the United States who believe that lenders should require a 20% down payment on a house. Based on the observed data, the best estimate for the "truth" is that 58% of all adults in the United States who believe that lenders should require a 20% down payment on a house. However, because of the uncertainty associated with sampling, it may be that "truth" may be anywhere between 55% and 61%. All are just as believable.

11. Obtain the interval estimate for the true percentage of couples that lean to the right when kissing. Use the interval estimate to provide an answer to the research question.

# *Margin of Error*

Section on Survey Research, American Statistical Association. (1998). *What is a margin of error?* Alexandria, VA: Author.

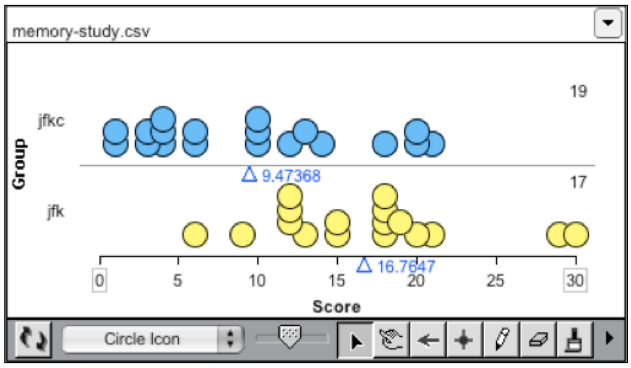
To further help you understand ideas related to sampling error and the margin of error, we would like you to read a short pamphlet put together by the American Statistical Association's Section on Survey Research.

- Read the pamphlet, *What is a Margin of Error?*. The pamphlet is available at `http://www.computing.dcu.ie/~jhorgan/margin.pdf`.

# *Course Activity: Memorization—Part II*

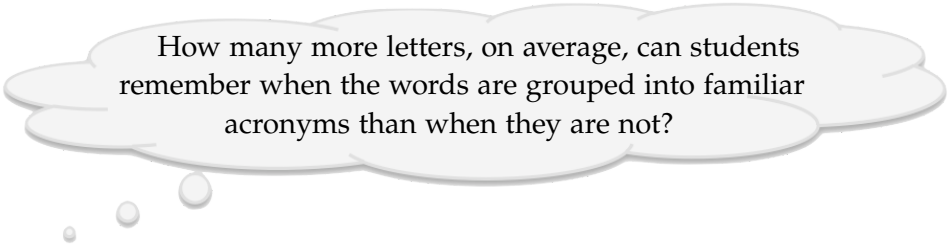
Recall the memory study that you analyzed to compare how many letters students in your class could recall when given the sequence grouped into familiar three-letter acronyms vs. unfamiliar groupings (with varying numbers of letters). The data from another class are presented below.

Data from the memorization study.




For this class's results, there was strong evidence against the null model of no difference between the two groups when the data was analyzed using a randomization test ($p$-value $< 0.01$). This suggested that the difference in the average number of letters memorized between the two treatments is not due to chance.

A natural follow-up question to rejecting the null model of no difference between the groups is:




Now you need to estimate how much better one treatment is than the other. This is called the *size of the effect*.

***Discuss the following questions.***

1. Use the observed data to find an estimate for the size of the effect. In other words, using the data, how many more letters, on average, can students remember when the words are grouped into familiar acronyms than when they are not?

## *Size of the Effect*

Size of the effect is a term used to describe the extent to which sample results diverge from the expectations specified in the null hypothesis.

For example, consider a study that compares two weight loss programs. The null hypothesis might be that there is no difference in the average weight loss between the two programs. However, the researchers might find that in their samples of participants, the average weight loss is 30 pounds more for program A. In this case, 30 pounds is the size of the effect.

As another example, consider a study in which a tutoring program is examined. Using a sample of students, the researchers find that the students involved with the tutoring program have, on average, raised their school performance by one letter grade over a control group of students. This grade increase is the size of the effect of the tutoring program.

Quantifying the size of the effect helps researchers focus on whether hypothesis test results, even those showing strong statistical evidence against the null model, are meaningful or not. Reporting the size of the effect is considered good practice when presenting empirical research findings in many fields since it facilitates the interpretation of the substantive, as opposed to the statistical, significance of a research result.

In the two examples above, the values offered for the size of the effect are estimates of the true size of the effect. Using one set of subjects, for example, researchers might find the average difference in weight loss to be 30 pounds. However, with a different set of subjects, the size of the effect might be estimated to be 32 pounds. Both are estimates of the true size of the effect, but may be different simply because of random assignment. Thus, as in the previous activity, it is important to also indicate a measurement of the *precision*.

### *Precision of the Size of the Effect*

In the previous unit, under the null model of no difference, the observations were pooled (put into one big group) and then the samples were drawn from the pooled sample. When this model is rejected (there is substantial statistical evidence against the null model), the conclusion is that the groups are different.

Under this conclusion, the strategy for modeling needs to change so that the simulation is conducted under the assumption of an alternative model. It is now assumed that the samples represent two different groups with two different mean values. Because it is believed that the samples represent two

different groups with two different mean values, the replicate data sets are drawn from the two samples separately—the first replicate sample is bootstrapped only using the observations from the first sample, and the second replicate sample is bootstrapped only using the second observed sample.

Once you get the two replicate samples, you can compute the size of the effect—in the case of the weight loss example this would be the difference in means. Repeating this process a large number of times will give us a distribution of plausible differences in means for the population under the alternative model of a difference between the two groups. Using this distribution, you can calculate an estimate of the sampling error. The following steps illustrate this process:

- Consider our sample data as representative of the two group populations.
- Sample, with replacement, from each of the observed samples separately, matching the sample sizes used in the study.
- Compute the estimate of the size of the effect (e.g., the difference in means).
- Find the estimate of the precision (i.e., the standard error or the standard deviation of the estimates).

### *Using TinkerPlots™ to Measure the Precision*

Set up models in TinkerPlots™ that can be used to simulate the 17 memory scores for the three-sequence group and 19 memory scores for the four-sequence group.

- Open the *Memorization.tp* file.
- Use TinkerPlots™ to carry out a single trial of the random samples for each group separately (see instructions in margin).
- Create plots for each of the resulting re-sampled groups.

**Setting Up the Model: Fixed Responses**

- Drag **TWO** Samplers from the object toolbar into the workspace.
- Create an empty mixer device in each of the sampler windows and set `Draw` to *1* for each sampler.
- Copy and paste the 17 memory scores for the JFK group into one of the mixers. Name this mixer *JFK_score*.
- Copy and paste the 19 memory scores for the JFKC group into the other mixer. Name this mixer *JFKC_score*.
- Change the `Repeat` value for each sampler to the respective sample size for the group, and set the mixer devices to sample `with replacement`.
- Click the `Run` button to simulate a single trial of the simulation.

2. Sketch the two plots below.

3. Compute (do not collect yet) the size of the effect (difference in the two mean score values) between the re-sampled groups.

4. Run another trial for each simulation. Explain why you do not obtain the same difference in sample means every trial, even though you are again sampling 17 subjects from one group and 19 from the other group.

- Collect the sample means for each group sampler, separately.
- Run 499 bootstrapped trials for each group sampler, separately.
- Create a table to calculate the difference in means for the 500 trials (see instructions in the margin).
- Plot the 500 differences.

**Create a Table for Difference in Means**

- Drag a table from the object toolbar into the workspace.
- Create a new attribute named *mean_difference*.
- Open the `History of Results` for the first sampler. Highlight the `mean_JFK_score` attribute and select `Copy Attribute` from the `Edit` menu.
- In the newly created table, highlight the *mean_difference* column and select `Paste Attributes` from the `Edit` menu. This will paste the `mean_JFK_score` attribute in the column to the left of the *mean_difference* column.
- Open the `History of Results` for the second sampler. Highlight the `mean_JFKC_score` attribute and select `Copy Attribute` from the `Edit` menu.
- Again, highlight the *mean_difference* column in the table and select `Paste Attributes` from the `Edit` menu.
- Right-click the *mean_difference* attribute and use the formula editor to compute the difference between the `mean_JFK_score` and the `mean_JFKC_score` attributes.

5. Sketch a plot below.

6. Where are they centered? Does this make sense? Explain.

7. Find the standard error.

## *Interval Estimate for the Size of the Effect*

Remember the estimate of the standard error gives us an indication of how variable the estimates of the size of the effect will be from sample to sample. Often, you are also interested in using the sample estimate to indicate how large the size of the effect will be for a particular population. In other words, you are interested in using the sample estimate to indicate something about the unknown parameter.

In the last activity, you used the sample estimate along with the estimate of the standard error to obtain an interval estimate. You can use the exact same method for obtaining an interval estimate for the size of the effect.

8. Compute the interval estimate of the size of the effect at 95% confidence.

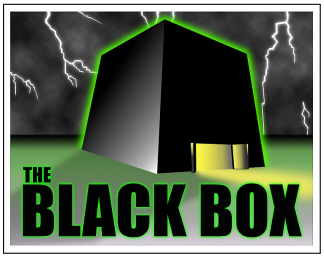
9. Interpret the interval and answer the research question.

If you want to estimate the true size of the effect, the key is to get a sense for how much variability there is between the potential sample estimates and the parameter due to random assignment of cases to the treatments or random sampling. To do this,

- Set up a simulation using the actual data in two mixers;
- Generate a large number of estimates of the size of the effect (e.g., differences in means) by sampling from each mixer with replacement;
- Determine the estimated standard error (SE) from the bootstrap distribution;
- Compute the estimated margin of error.
- Use the original estimate of the size of the effect (from the observed data) and the estimated margin of error to compute the interval estimate.

# *Course Activity: Why Two Standard Errors?*

A common statistical question is how to estimate the mean of an unknown population when you cannot access the entire population. If you are able to obtain a random sample of data from such an unknown population, you can use the bootstrap method to estimate the mean of the population and the variability of sample means of random samples drawn from the population (i.e., the standard error). Using the original random sample as a model, you use the bootstrap resampling method to obtain many random samples of it.




How can you obtain a precise estimate for a parameter using only a single sample?

## *Exploring Random Samples from an Unknown Population*

To address the question, you are going to draw random samples from an "unknown" population and examine the distribution of sample means computed from those samples.

- Open the *Population.tp* file.

A random sample of 15 values has been drawn for you from the "unknown" population. (To further convey the idea that the population is not known, it has been covered by a gray rectangle and marked with a question mark.) In addition, the sample mean has been collected into the `History of Results` table and plotted. The plot, called a *rug plot*, indicates the value of the sample mean by a short line segment rather than a circle.

- Collect 999 more random samples.
- Use the divider tool to indicate where most of the data are. To do this, think about where you might "cut off" the extreme values. It might help to look where the rug plot starts to "thin out" (i.e., noticeable space between the marks) at each end of the plot.

Remember that each mark represents a mean value from a randomly drawn sample. Each of these sample means are, in turn, an estimate for the true population parameter, which is unknown to us. Since all of the means come from random samples drawn from the exact same population, the variation seen in the rug plot is completely due to random sampling. The divider gives us a sense of where most of the sample means are, and in turn, this helps to provide an idea of how much variability can be expected in the sample estimates just because of random sampling.

1. What are the values indicated by the divider tool endpoints?

2. What is the mean value for the 1000 sample estimates?

3. How far above and below the mean value (approximately) are the divider tool endpoints?

4. Roughly, what percentage of the sample means are contained in the middle part of your divider tool?

5. What would be your best guess of a good interval estimate for the mean of the unknown population?

6. Compute the standard deviation of the sample means (i.e., the standard error).

7. How many standard errors above (or below) the mean value are the divider tool endpoints?

8. How do you think this relates to why the margin of error is estimated as $2 \times SE$?

9. Compute the endpoints of an interval based on the mean of the 1000 estimates $\pm 2 \times SE$. Roughly, what percentage of the collected sample means are inside the interval?

### *Exploring a Single Random Sample*

In practice, researchers do not draw several samples and thus, do not get to examine the characteristics of a distribution of sample estimates in order to obtain a measure for the standard error. Rather, the convention is to use a single sample to estimate the standard error. In this section, you will examine the bootstrap estimate for the standard error and interval estimate from a single sample to those estimates which were computed from re-sampling out of the "unknown" population.

- Open the provided sample of data corresponding to your group number.

10. Compute the mean for your sample.

11. Use the nonparametric bootstrapping method to obtain 100 bootstrapped means. Sketch a plot of the bootstrapped means below.

12. Find the estimate of the standard error based on the bootstrapped means (i.e., compute the standard deviation) and use it to calculate the margin of error ($2 \times SE$).

13. Using the margin of error and the sample mean, compute an interval estimate for the true population mean.

14. What percentage of the collected bootstrapped means are inside the interval? (Hint: You may want to use the divider tool.)

---


***Sketch your plot on the board. Be sure to label your axis and the lower and upper limits of your interval estimate. Also, record both the sample mean and the estimate for the standard error (based on your bootstrap distribution).***

---

### *Exploring the Bootstrap Distribution from Different Random Samples*

Examine the plots of your classmates. Each plot represents the distribution of 100 bootstrapped means. The bootstrapping for each plot is based on a different random sample of 15 values drawn from the original "unknown" population.

15. Which characteristics of the different bootstrap distributions are the same? Which characteristics are different?

16. Explain why you would expect these similarities and differences.

### *Exploring the Bootstrap Intervals Computed from Different Random Samples*

Now, consider the full computation for the interval estimate,

$$\text{Sample Estimate} \pm 2 \times SE.$$

To compute an interval estimate only requires two quantities, a sample estimate and an estimation of the standard error. Since the bootstrap allows researchers to estimate the standard error from a single sample, it is possible to compute an estimate of

the uncertainty due to random sampling and interval estimate using only one sample!

Your instructor will enter the sample mean and estimate of the standard error from the different random sample into a table in TinkerPlots™ and display the interval estimates from all of the groups in a plot. Each interval estimate is represented by a horizontal line segment. The point in the middle of each line represents the sample mean. The plot also displays the mean of the "unknown" population using a vertical reference line.

17. Examine the estimates for the standard error (SE) in the table. Are they all the same? How close are they to each other? How similar are they to the estimate of the standard error you computed based on drawing 1000 random samples from the "unknown" population (see Question 6)?

18. Examine the plot of the interval estimates. Are the endpoints for each interval estimate exactly the same? Are they close to each other? Are the widths of the intervals similar?

19. Considering how the interval estimate is computed, explain why you would expect these similarities and differences.

20. Examine the interval estimate from your specific random sample. Is the population mean value inside your interval estimate?

21. Again, examine the plot of all the interval estimates. How many of the interval estimates include the population mean value inside the interval? How many of the interval estimates do not include the population mean value inside the interval?

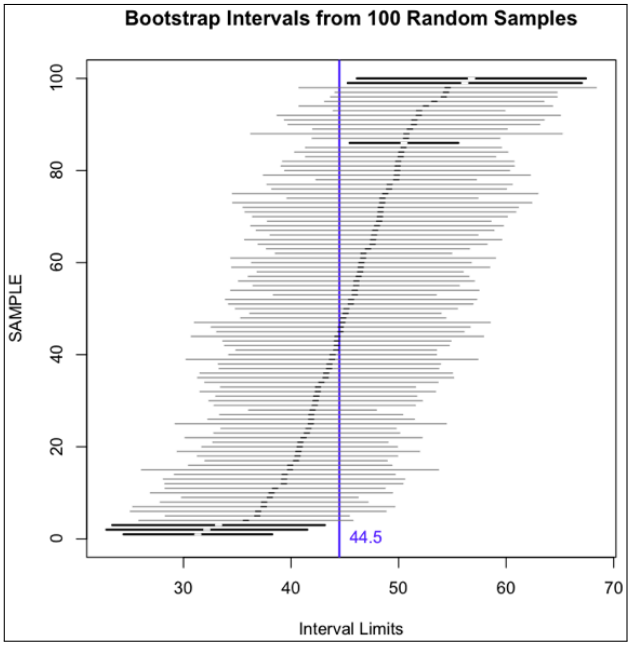
### *What does it Mean to be 95% Confident?*

Interval estimates are sometimes referred to as *confidence intervals*. In interpreting these intervals, researcher will often say something like, "I am 95% confident that the population mean is between $A$ and $B$. What do they really mean when they make a statement like this?

Not every interval estimate included the population mean. This can happen just because of chance (unlucky sampling). But what are your chances of producing a bootstrap interval estimate that includes the true population parameter (the population mean in this case)?

The figure on the next page shows the bootstrap interval for each of 100 random samples drawn from the "unknown" population represented as a horizontal line segment. The sample mean is represented by a white mark in the middle of the interval. The vertical blue line represents the population mean.

To help you answer the questions below, the intervals have also been sorted in order of their sample means (the random sample having the lowest sample mean is on the bottom and the random sample having the highest sample mean is on top) and color-coded (if the interval includes the population mean value, it is drawn in grey, otherwise it is drawn in black).




The bootstrap interval for each of 100 random samples drawn from the "unknown" population. If the interval includes the population mean value, it is drawn in grey, otherwise it is drawn in black.

22. How many of the intervals include the population mean value (the grey intervals)? What percentage is this?

23. How many of the intervals do not include the population mean value (the black intervals)? What percentage is this?

24. Consider only the intervals that include the population mean value. What are the lowest and highest sample means for this set of intervals?

25. Compare the values from the last question to the endpoints of the values indicated by the divider tool endpoints you found in Question 1?

When you draw a random sample from an unknown population, you have a 95% chance of obtaining a sample that will produce an interval estimate based on the bootstrap that includes the population parameter. Because of this, you say that you are 95% confident that the true population parameter is in the interval estimate.

# *Learning Goals: Unit 3*

The activities, homework, and reading that you have completed for the third part of this course have introduced you to several more fundamental ideas in the discipline of statistics. The ideas and concepts you were introduced to in this unit form the basis for all formal statistical methods for statistical estimation. In addition, you have learned how to use TinkerPlots™ to bootstrap the standard error for a statistic.

Below are the key concepts and skills from Unit 3 that you should have learned. At this point in the course, you should

## *Literacy/Understanding (Terms and Concepts)*

- Understand the impact of sampling method on the bias and precision of estimates
- Understand desirable properties of estimators such as unbiasedness and precision
- Understand that sample size impacts the precision of an estimator, but not the bias
- Understand that the size of the population does not impact the precision of an estimator, nor the bias
- Understand how to quantify the variation in a data set using the standard deviation
- Understand standard error as a measure of precision of an estimate (i.e., measure of variability between potential random samples or random assignments)

- Understand the use of standard error in computing the margin of error
- Understand the idea of the nonparametric bootstrap method and the resulting bootstrap distribution
- Understand the idea of using bootstrapping to estimate the standard error for an estimator (mean, percent/proportion, difference of means)
- Understand an interval estimate as containing a point estimate and margin of error
- Understand why the margin of error is equal to two standard errors
- Understand size of effect as a measure of how two groups differ
- Understand that when estimating the difference in means or proportions between two groups, the simulation is carried out assuming that the alternative hypothesis of group differences is true (i.e., the data are not pooled together, instead the data from each group is used independently to generate bootstrap samples)

### *Selecting/Using Models*

- Be able to use a set of sample data as a model to simulate data (bootstrap method) using sampling with replacement
- Be able to generate a bootstrap distribution
- Be able to use the data as a model (nonparametric bootstrap) to find the standard error of a statistic
- Be able to use two sets of sample data in a model that assumes group differences to obtain a bootstrap distribution and approximate a standard error

### *Evaluation*

- Evaluate the validity of an estimate based on the sampling method used to obtain the data
- Evaluate the precision of an estimate by interpreting the standard error
- Evaluate the uncertainty in an estimate by interpreting the standard error and margin of error

### *TinkerPlots™ Skills*

You should also be able to do the following using TinkerPlots™.

- Sample with replacement using a sampling device
- Use two sampling devices (each sampling without replacement) to bootstrap under the assumption of group differences
- Compute the standard deviation or standard error using the `stdDev` function

In the next activity, *Unit 3 Wrap-Up & Review*, you will have a chance to assess yourself on whether or not you have mastered these ideas through a variety of practice and extension problems. As a pre-cursor to this activity, you may want to review the readings and activities in Unit 3.

# *Course Activity: Unit 3 Wrap-up & Review*

## *Terminology for Unit 3*

1. At this point, you should be familiar with the following terms. Write down what each term represents as well as any notes that may help you remember.

   (a) Bias

   (b) Precision

   (c) Representative

   (d) Population

   (e) Sample

   (f) Model

   (g) Parameter

(h) Statistic/Sample Estimate

(i) Standard Deviation

(j) Standard Error

(k) Sampling Variation/Sampling Error

(l) Margin of Error

(m) Interval Estimate

(n) Size of the Effect

### *Sampling*

2. Consider estimating the average word length for all words written on the internet using a sample of 100 pages from Wikipedia. Explain why that particular sampling method would be considered biased.

3. Identify the direction of the bias. Explain your reasoning. (In other words, does the sampling method tend to produce an estimate which would overestimate or underestimate the average word length for all words written on the internet?)

### *Comparing Airlines—Revisited*

Recall again the activity where you compared two regional airlines, Mesa and American Eagle, to determine if they are equally reliable with respect to arrival delays. You were given data of arrival time delays for both airlines to five different cities. Two of those cities were Madison, WI and Fort Wayne, TX.

The difference in mean arrival delay times for the two airlines was calculated by taking the mean arrival delay time for Mesa Airlines and subtracting it from the mean arrival delay time for American Eagle Airlines. The difference in means was $-15.9$ for Madison, WI (American Eagle arrived, on average, 15.9 minutes earlier than Mesa), whereas the difference in means was $-20$ for Fort Wayne, TX (American Eagle arrived, on average 20 minutes earlier than Mesa).

Use the *Airlines-Variability.tp* file to help you answer the following questions.

4. Use TinkerPlots™ to calculate the standard deviation of the arrival delay times for Madison, WI. Do the same for Fort Wayne, TX. Report the values of each standard deviation.

5. For Madison, run 500 bootstrap trials to estimate the standard error of the bootstrapped difference in means. Calculate and report this value.

6. For Fort Wayne, run 500 bootstrap trials to estimate the standard error of the bootstrapped difference in means. Calculate and report this value.

7. Which of the cities has the larger standard error?

When a a randomization test was carried out to examine the the null model of no difference between the two mean arrival times, the $p$-value for Madison, WI was 0.132 whereas the $p$-value for Fort Wayne, TX was 0.226.

8. Considering the two standard errors, explain why the $p$-value for Madison, WI is smaller than the $p$-value for Fort Wayne, TX, even though Fort Wayne, TX has the larger observed difference in means?

## *Rating Chain Restaurants*

The August 2012 issue of *Consumer Reports* included ratings of 102 chain restaurants. The ratings were based on surveys that readers sent in after eating at one of the restaurants. The article reported that the survey results were based on 110,517 visits to full-service restaurant chains between April 2011 and April 2012, and reflected the experiences of their readers, not necessarily those of the general population.

http://www.consumerreports.org/cro/magazine/2012/08/america-s-best-restaurant-chains/index.htm

9. Do you think that the sample here was chosen randomly from the population of Consumer Report readers? Explain.

10. Why do the authors of the article make this disclaimer about not necessarily representing the general population?

11. To what population would you feel comfortable generalizing the results of this study? Explain.

## *Emotional Support*

In the mid-1980s, Shere Hite, a prominent sex researcher, undertook a study of women's attitudes toward relationships, love, and sex by distributing 100,000 questionnaires in women's groups. Of the 4,500 women who returned the questionnaires, 96% said that they gave more emotional support than they received from their husbands or boyfriends.

Hite, S. (1987). *Women and love: A cultural revolution in progress.* New York: Alfred A. Knopf.

12. Comment on whether Hite's sampling method is likely to be biased in a particular direction. Specifically, do you think that the 96% figure overestimates or underestimates the proportion of all American women who give more emotional support than they receive from their husbands or boyfriends?

An *ABC News/Washington Post* poll in the same year surveyed a random sample of 767 women, finding that 44% claimed to give more emotional support than they received.

http://www.highbeam.com/doc/1P2-1350996.html

13. Which poll result do you think is more representative of the population of all American women? Explain.

### *Balsa Wood*

Student researchers investigated whether balsa wood is less elastic after it has been immersed in water. They took 44 pieces of balsa wood and randomly assigned half to be immersed in water and the other half not to be immersed in water. They measured the elasticity by seeing how far (in inches) the piece of wood would project a dime into the air.

Rossman, A. J., Chance, B. L., & Lock, R. H., (2009). *Workshop statistics: Discovery with data and Fathom* (3rd ed.). Emeryville, CA: Key College Publishing.

Use the *Balsa-Wood.tp* file to help you answer the following questions.

14. The observed difference in mean elasticities between the two groups is 4.16 inches. Explain why it is more appropriate to produce a bootstrap interval, as opposed to simply reporting this value, for estimating the actual treatment effect of immersing balsa wood in water.

15. Produce an interval estimate for the size of the effect of immersing balsa wood in water. Describe the process by which you produce this interval, and also interpret what the interval means in the context of this study.

## *Microsort®*

The *Genetics and IVF Institute* is currently studying methods to change the odds of having a girl or boy. MicroSort® is a method used to sort sperm with X- and Y-chromosomes. The method is currently going through clinical trials. Women who plan to get pregnant and prefer to have a girl can go through a process called X-Sort®. As of 2008, 945 women have participated and 879 of those women have given birth to girls.

http://www.microsort.com/

16. Compute an interval estimate of the percentage of female births for women that undergo X-Sort®.

17. Interpret your interval.

18. Why do you set the `Repeat` value in your simulation equal to the sample size?

19. Suppose more data has been collected since 2008. If the number of women had increased to 3,000 but the observed percentage of femal births remained the same, what would you expect to happen to your interval?

20. Test out your conjecture by creating a new interval using a sample size of 3,000. Report your new interval estimate. Was your expectation in question 15 correct?

21. How many trials did you run in your simulations?

22. What is the difference between sample size and number of trials?

## *Marijuana*

Pope and Yurgelun-Todd studied whether frequent marijuana use is associated with residual neuropsychological effects. College undergraduate students were recruited to participate in the study. A total of 65 heavy marijuana users and 64 light users participated. The students took a neuropsychological test which involves sorting cards. The average number of cards correctly sorted for the heavy marijuana users was 51.3 and for the light marijuana users was 53.3 .

Pope, H. G., & Yurgelun-Todd, D. (1996). The residual cognitive effects of heavy marijuana use in college students. *Journal of the American Medical Association, 274*(7). 521–527.

Use the *Marijuana.tp* file to help you answer the following quiestions.

23. Carry out a randomization test to determine whether heavy marijuana users sort fewer cards correctly, on average, compared to light marijuana users? Report the pertinent results below.

24. Provide an interval estimate for the size of the effect. Report the pertinent results and interpret the interval below.

# *Bibliography*