

Stat 216 Course Pack Fall 2015

Activities and Notes

Sections: 2, 3, 6, 11, 14, 18, 20, and 21 meeting
Tuesday and Thursday

with two Common Hour Exams, and Common Final



Department of Mathematical Sciences
Montana State University

STAT 216 Introduction to Statistics

Fall 2015 Calendar of Topics

for Sections 2, 3, 6, 11, 14, 18, 20, and 21 meeting Tuesdays and Thursdays

TUESDAY	THURSDAY
August 25 Detecting Fraud (1) Class Begins	27 Descriptive Stats (2) Aug 28: Last Day to Add On-Line
September 1 Sampling (3)	3 Helper–Hinderer (4) Sept 4: Last Day to Drop On-Line
8 Hyp Test 1 proportion(ESP) (5)	10 Estimate 1 proportion (6)
15 What “confidence” means (7) Sept 14: Last Day to Avoid a W	17 Test & Estimate 1 Proportion (MIT) (8)
22 Unit 1 Review (9) Common Hour Exam I 6:00 - 7:50 pm	24 No Class
29 Exp vs Obs Study(10)	October 1 Textbook Cost – CI for μ (11)
6 Peanut Allergies (12)	8 Weight Awareness $p_1 - p_2$ (13)
13 Energy Drinks, $\mu_1 - \mu_2$ (14)	15 Birth Weights, $\mu_1 - \mu_2$ (15)
20 Hyp Test 1 mean (16)	22 Correlation/slope (17)
27 Test: “Is slope zero?” (18)	29 Types of Errors (19)
November 3 Unit 2 Review (20) Common Hour Exam II 6:00 - 7:50 pm	5 Normal Distribution (21)
10 Z inference for p (22)	12 Z inference for $p_1 - p_2$ (23)
17 t distributions (24)	19 t inference for μ (25) Nov 19: Last Day to Withdraw
24 No Class	26 Thanksgiving Holiday
December 1 Sample Size Effects (26)	3 Review (27) Last Day of Class
Final Exam Week: December 7 – 11 Common Hour Stat 216 Exam: Wednesday, ??, 10:00 – 11:50 am Rooms: TBA	

Stat 216 Syllabus Fall 2015

People

- Your Instructor: (Write contact info here)

- Student Success Coordinator: Jade Schmidt
email: roskam@math.montana.edu Office: Wilson 2-260 406-994-5357
- Course Supervisor: Dr. Robison-Cox
email: jimrc@math.montana.edu Office: Wilson 2-241 406-994-5340

Course Materials

You need to buy the Stat 216 Course Pack with **Tuesday – Thursday** on it from the MSU Bookstore. There is another pack for MWF which is different.

We will also use several online web applications – so you need access to a computer. You will work as a group of three and one of your group needs to bring a computer for each class meeting. The free online textbook *Intro Stat with Randomization and Simulation* at <https://www.openintro.org/stat/textbook.php> will be used for some of its explanations.

Other materials, such as readings and “Quizorks” (our word for very important homework sets) will be downloaded from D2L, so be sure you can log in to the MSU D2L (Brightspace) system: <https://ecat.montana.edu/>. If you have problems, view the help on that page.

Learning Outcomes for STAT 216

- Understand how to describe the characteristics of a distribution.
- Understand how data can be collected, and how data collection dictates the choice of statistical method and appropriate statistical inference.
- Interpret and communicate the outcomes of estimation and hypothesis tests in the context of a problem.
- To understand when we might make causal inference from a sample to a population.
- To understand how selection of a sample influences the group to which we might make inference.

CORE 2.0: This course fulfills the Quantitative Reasoning (Q) CORE 2.0 requirement because learning statistics allows us to disentangle what's really happening in nature from “noise” inherent in data collection. It allows us to evaluate claims from advertisements and results of polls and builds critical thinking skills which form the basis of statistical inference.

Comments and concerns: We are always looking for ways to improve this class and we want students to be successful. The first step is to discuss your comments or concerns with your instructor. If they are not resolved, contact the Student Success Coordinator, Jade Schmidt.

Course Description

This section of Stat 216 is designed to engage students using a simulation approach to inference using web apps. Use of small group discussion activities and daily assignments have been shown by the research to be effective. Upon completion of this course, you should have an understanding of the foundational concepts of data collection and of inference and you will appreciate the fundamental role that statistics plays in all disciplines. In addition, statistical summaries and arguments are a part of everyday life, and a basic understanding of statistical thinking is critical when it comes to helping you become an informed consumer of the numerical information they encounter on a daily basis. You will be exposed to numerous examples of real-world applications of statistics that are designed to help you develop a conceptual understanding of statistics.

Note: this course will be a lot of work, and attendance every day is really important for your success.

Please think seriously about this as you decide if this course is the right fit for you.

Prerequisites

You should have completed a 100-level math course (or equivalent) with a grade of C- or better (Alternatives: a good score on Math portion of SAT or ACT, or a 3.5 on the MPLEX exam). You should have familiarity with computers and technology (e.g., Internet browsing, word processing, opening/saving files, converting files to PDF format, sending and receiving e-mail, etc.). See the Technology section of the syllabus for more details.

Technology

- **Web Applets** We will be utilizing web applets at <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/> These run in a web browser, but may have trouble with older versions of the Microsoft IE browser.
- **Technology Policy:** This course utilizes technology extensively. Bring a laptop or tablet to class (your phone might get you by for web apps, but you'll also need a word processing program). You will need at least one laptop within your group each day.

Math Learning Center in 1-112 Wilson Hall is a very important resource providing help on Stat 216 topics. Fill in the hours here.

Assessment

Your grade in this course will be based on the following:

- **Quizorks: 15%** These assignments will help you learn the course material and software through reflection and practice and are essential preparation for the exam.
Format: Your instructors will tell you if you submit these as electronic files uploaded to D2L or as hard copies. If electronic, it needs to be in a format we can read. Adobe pdf is our standard. Submissions we can't read will not count.
- **Online (D2L) Exercises: 10%**
- **Common Hour Exam I: 20%** Taken individually, not in groups. You may bring a one page sheet of notes.
- **Common Hour Exam II: 20%** Taken individually, not in groups. You may bring a one page sheet of notes.
- **Final Exam: 25%.** This exam will be cumulative in content. Again, you will be allowed to bring in one page of handwritten notes for the final exam.

- **Attendance/Participation/Preparation: 10%** . Class participation is an important part of learning, especially in courses like this one that involve group cooperation.

Participation/Attendance: Students can miss class/arrive late/leave early once (1 day) before they will be penalized for non-participation due to an absence. For each day missed thereafter, the students overall grade will be reduced 1% (up to 5%).

Preparation: The in-class activities and out-of-class assigned readings are the primary source of information for this course. Take them seriously, work through them with care, and they will be very valuable on exams. As a way to provide further emphasis to the activities and readings, most classes will begin with a Readiness Assessment Test (RAT) with questions covering the previous class's activity and readings required for the class.

Late or Missed Work: If you cannot be in class, it is your responsibility to notify the instructor and your group members with as much advance warning as possible. In general, make-up exams or late homework assignments will not be allowed. Case-by-case exceptions may be granted in only extreme cases at the discretion of the instructor (daily work) or Student Success coordinator (exams). You must provide documentation explaining your absence for the instructor to determine whether an exception should be granted. If you fail to provide documentation as requested then you will not be able to make-up missed work at all.

Letter grades will be assigned using a 10 point scale. As an approximation (which will be fine tuned at the end of the semester) 94 - 100 = A, 90 to 93 = A-, 87 to 89 = B+, etc.

Some Department Policies:

- Do not attempt to turn in any assignment in the math office. They will not be accepted.
- Do not call or email the math office for information on grades.

University Policies and Procedures

Behavioral Expectations

Montana State University expects all students to conduct themselves as honest, responsible and law-abiding members of the academic community and to respect the rights of other students, members of the faculty and staff and the public to use, enjoy and participate in the University programs and facilities. For additional information reference see MSU's Student Conduct Code at: http://www2.montana.edu/policy/student_conduct/cg600.html . Behavioral expectations and student rights are further discussed at: <http://www.montana.edu/wwwds/studentrights.html> .

Collaboration

University policy states that, unless otherwise specified, students may not collaborate on graded material. Any exceptions to this policy will be stated explicitly for individual assignments. If you have any questions about the limits of collaboration, you are expected to ask for clarification.

Plagiarism

Paraphrasing or quoting another's work without citing the source is a form of academic misconduct. Even inadvertent or unintentional misuse or appropriation of another's work (such as relying heavily on source material that is not expressly acknowledged) is considered plagiarism. If you have any questions about using and citing sources, you are expected to ask for clarification.

Academic Misconduct

Section 420 of the Student Conduct Code describes academic misconduct as including but not limited to plagiarism, cheating, multiple submissions, or facilitating others' misconduct. Possible sanctions for academic misconduct range from an oral reprimand to expulsion from the university.

Section 430 of the Student Code allows the instructor to impose the following sanctions for academic misconduct: oral reprimand; written reprimand; an assignment to repeat the work or an alternate assignment; a lower or failing grade on the particular assignment or test; or a lower grade or failing grade in the course.

Academic Expectations

Section 310.00 in the MSU Conduct Guidelines states that students must:

- A. be prompt and regular in attending classes;
- B. be well prepared for classes;
- C. submit required assignments in a timely manner;
- D. take exams when scheduled;
- E. act in a respectful manner toward other students and the instructor and in a way that does not detract from the learning experience; and
- F. make and keep appointments when necessary to meet with the instructor. In addition to the above items, students are expected to meet any additional course and behavioral standards as defined by the instructor.

Withdrawal Deadlines

September 4, 2015 is the last day to withdraw without a "W" grade. University policy is explicit that the adviser and instructor must approve requests to withdraw from a course with a grade of "W".

Group Expectations

We have all been in groups which did not function well. Hopefully, we've also all had good experiences with working in groups. Our use of groups in this course is based on educational research which provides strong evidence that working in groups is effective and helps us learn. By expressing your opinions and catching each other's mistakes, you will learn to communicate statistical concepts. The statistical concepts you will be learning are partly "common sense" ideas (for instance, gathering more data provides a better foundation for decision making), but they are often phrased in odd ways. We find it really helps to talk about them with others.

Detecting Fraud

Randomness is a key concept in statistics, but how good are we at detecting randomness, or better yet, at detecting fraud? We will investigate this idea today.

With your group:

1. Write down a sequence of 45 coin flips (so 45 heads and tails) that you (as a group) think is random.
2. Now flip your coin 45 times and write down the result of each flip.

3. Flip one more coin. If you get heads, label the second sequence of actual flips as sequence “A”, the “made up” flips as sequence “B”. If you get tails, label them oppositely. Don’t let other groups see which is which.

Write both sequences on the board in shorthand with the number of consecutive heads or tails, so the sequence HHHTTHHT will be recorded as:

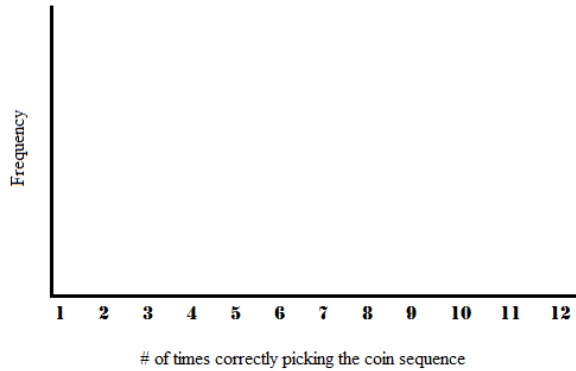
3H
2T
2H
T

4. As the other groups write their sequences on the board, guess whether A or B was really the coin flip sequence. Mark your guesses for the sequence from the coin in the table below. Each person should guess individually and not as a group! When you come to your own group, flip a coin and mark “A” for Heads, “B” for Tails.

Group	1	2	3	4	5	6	7	8	9	10	11	12	13
Guess (A or B)													

5. Each group will now reveal their true “coin flip” sequence. How many correct guesses did you have? Others in your group? Your instructor?

6. Plot your number correct on the board and copy the plot here. Is the instructor’s point unusual?



Two possible explanations for an unusual result:

- Your instructor was merely guessing which sequence was generated by the coin. If this is the case, how many times would you expect the instructor to have been correct?
- Your instructor has super secret knowledge of randomness and could detect the fraudulent sequences. If this is the case, how many times would you expect the instructor to have been correct? *Hint: you do not need to specify a number here, just a direction from a particular number.*

7. Which of these two explanations seems more plausible given the data? Explain your group's choice.
8. If the class was just guessing when writing down which sequence was the coin, your results give an example of how many sequences a person *could* guess correctly out of _____. Which values are unlikely, according to the plot? Is the instructor's result unlikely?
9. What does this tell you about your instructor? Do you think he/she was guessing or can detect fraud? Explain your answer using the dot plot of the class results.

Further Exploration

10. What could your instructor do to make you more sure they can in fact detect fraud?

11. If an instructor had correctly identified the coin sequence for 20 of 24 groups and a different instructor had correctly identified 10 of 12 groups, which instructor would you think is better at detecting fraud? Or would they be equally effective? Explain.

Take Home Messages

- In this course we will learn how to evaluate a claim by comparing observed results (Instructor's guess) to a distribution.
- Blind guessing between two outcomes will be correct only about half the time. We can create data (via computer simulation) to fit the assumption of blind guessing.
- Unusual results will make us doubt the assumptions used to create the distribution. A large number correct is evidence that a person was not just blindly guessing.

Assignment

- Trade contact info with your group members. Decide who will bring a computer to the next class.
- Log in to this course on D2L.
- Look through the course resources.
- Complete **Exercise 1** on working in groups.
- Read pages 9–12 for the next class. You will be quizzed over them.

Descriptive Statistics

Data is everywhere. We take for granted the fact that our smart phones, smart TV's and other hi-tech gadgets store huge amounts of data about us. We have quickly become used to being able to call up all sorts of information from the web. To handle data we first have to distinguish several types of data which are handled and plotted differently.

As an example, suppose that we want to filter out spam messages before they hit your email inbox. We can keep track of several attributes of an email, and each email will have its data on a single line in the data file (one line is called a “**case**” or a “record”). It may look like this:

spam	num_char	line_breaks	format	number
0	21.70	551	html	small
0	7.01	183	html	big
1	0.63	28	text	none
0	2.45	61	text	small
0	41.62	1088	html	small
⋮	⋮	⋮	⋮	⋮
0	15.83	242	html	small

Where the **variable** in each column tells us:

spam is 1 if the message is known to be spam, 0 otherwise.

num_char counts the length of the message in thousands.

line_breaks counts the number of lines of text.

format is either “html” or “text”.

number is small if text contains a number < 1 million, big if a number over 1 million is included, and none otherwise.

We will divide variables into two main types:

Categorical variables tell us about some attribute of the case which is not numeric, for example: hair color or favorite sport. The categories can be numeric (like zip codes) if it makes no sense to “average” them together.

Quantitative variables are numbers which can be averaged together. They can be integers (like counts) or precise measurements like milliliters of beer in a stein.

Data summaries vary with data type

Categorical variables are summarized with tables like this:

html	13
text	37

which says that 13 of the messages were in html format, and 37 were plain text. We could also say that 26% ($= 13/50 \times 100\%$) of the emails were in html format.

Quantitative variables are summarized with measures of center (mean or median) and spread, and sometimes with quartiles.

mean or “average” is found by summing all values and dividing by the size of the sample (we label sample size as n). With a “sample” of values, we call the first one x_1 , the second x_2 , and so forth, and we call the mean “x bar” which is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

For the number of characters in the emails, we get

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.598.$$

median is a number which has half the values below it and half above it. It is not affected by extreme values in the way that the mean is. The number of characters in an email has some large values which inflate the mean, but the median is smaller at 6.89 thousand characters.

first quartile labeled Q_1 , has one fourth of the values below it and three-fourths above. It is also called the 25th percentile.

third quartile labeled Q_3 , has three fourths of the values below it and one-fourth above. It is also called the 75th percentile.

Inter-Quartile Range or IQR, is the distance between the first and third quartiles. It is a measure of **spread** of the values. For the 'numbers of characters' data, Q_1 is 2.536 and Q_3 is 15.411, so $IQR = 15.411 - 2.536 = 12.875$.

Standard Deviation labeled s is roughly the average distance from each point to the mean of the sample. We do not expect you to compute it, but the formula is

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

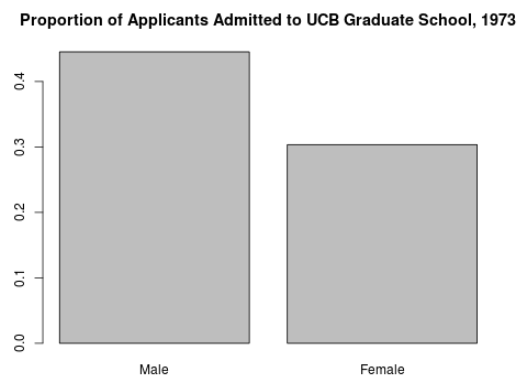
which, for the data we are considering, is 13.125.

It is an important measure of **spread**.

Plotting Data

As with numeric summaries, the type of data determines the appropriate plot.

Categorical variables are plotted using a bar chart. (Note, one could use a pie chart, but then it is much harder to compare two areas of the pie than with the bar chart.) For a more interesting example, we'll consider the admissions rate of applicants to UC-Berkeley grad school in 1973 separated by gender. (Gender is categorical and so is “admitted or rejected”, so the plot allows us to compare one categorical variable split by another. This seems more interesting than just looking at one variable – like admission rates for all applicants.)

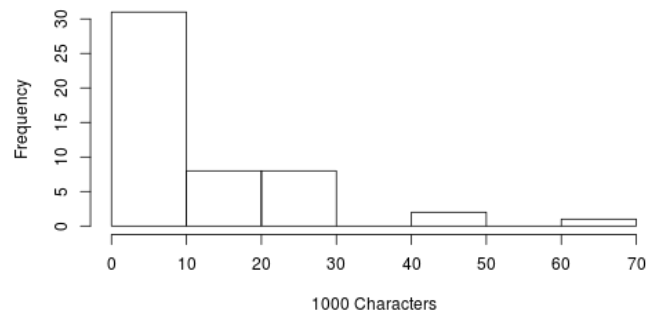


Quantitative variables are plotted with dot plots, histograms, density plots, and boxplots.

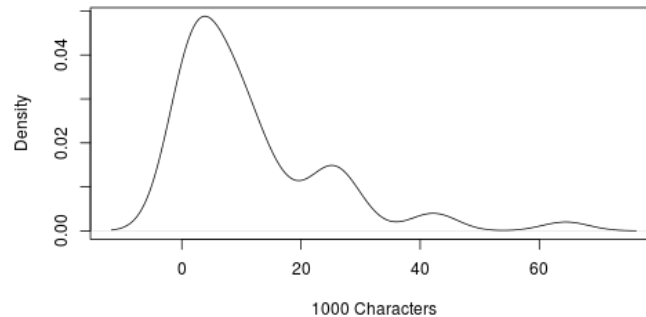
dot plots represent each point with a dot above the number line. This works well with small sample sizes. If the data are too close together to distinguish, we might stack them up to remove any overlap.



histograms divide the axis into “bins” and count the numbers of points falling into each bin. The height of each bin might show the count (frequency) of values in the bin or the proportion (relative frequency) for the bin. These plots work with moderate to large sized data sets. Choosing the best number of bins can be hard.



density plots are basically like smoothed off relative frequency histograms.



box-and-whisker plots show the quartiles of the distribution, making a box from Q_1 to Q_3 (median is also Q_2), and then showing whiskers which extend to the minimum and maximum value. If those extremes are too far out, the whisker usually stops at the last point within $1.5 \times \text{IQR}$'s of either Q_1 or Q_3 and flags points beyond $1.5 \times \text{IQR}$ as “outliers”, or unusual points. Half of the data will be included in the box, and half will be outside the box.



One more idea is important in describing a sample of quantitative values is the **skew** of a distribution of values.

A distribution is skewed if the histogram tapers off to one side. For example, the num_char variable above shows strong right skew because the histogram and density plots taper down to the right, and the boxplot has a long “right tail” (longer whisker to right and outliers to right). If those same plots look roughly the same on each side, we say the data are “symmetrically distributed” rather than saying “unskewed”.

Got Data?

Data are everywhere. We take for granted the fact that our smart phones, smart TV's and other hi-tech gadgets store huge amounts of data about us. We have quickly become used to being able to call up all sorts of information from the web.

Statistics is all about making sense of data, so we first need to pay some attention to the main types of data we will be using.

1. Which variable is of a different type?
 - A. The cell phone carrier you use.
 - B. The monthly fee charged by your cell phone provider.
 - C. Whether your cell phone has buttons or touch screen.
 - D. The manufacturer of your cell phone.

Circle the odd ball and explain why its different.

2. Got it? – Let's just check again for the different data type.
 - E. Amount you spend on textbooks this term.
 - F. Number of credits you're signed up for.
 - G. How much student loan you'll take out this term.
 - H. The area code of your phone number.

Again circle one and explain.

3. One thing we need to be comfortable with is summarizing data. For each of the above variables, A through H, how would you summarize data collected from each person in class today?

You've read about two main types of data:

Quantitative takes numeric values which we can average.

Categorical falls into one of two or more categories. The categories can have numeric labels (like zip codes), but it makes no sense to average them. (some call this "Qualitative", but we don't like to use two words starting with Q)

4. For which variables on the previous page, A through H, would the **mean** be informative?

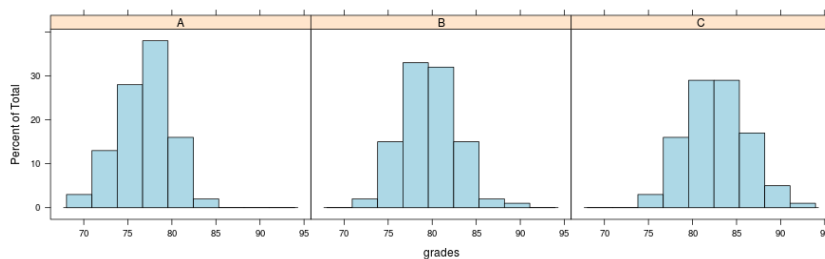
We also need to summarize categorical data, so we use proportions: the number in a certain category divided by the total number.

5. For which variables on the previous page, A through H would the **proportions** be informative?

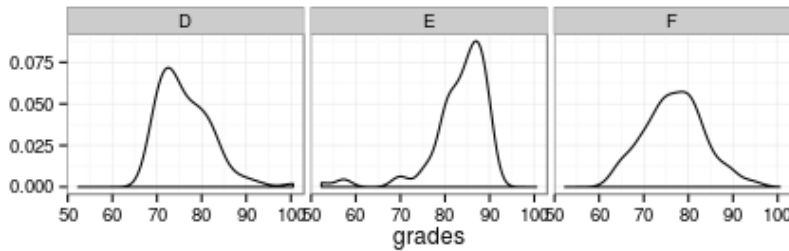
Comparing Distributions

Now we'll focus on quantitative data.

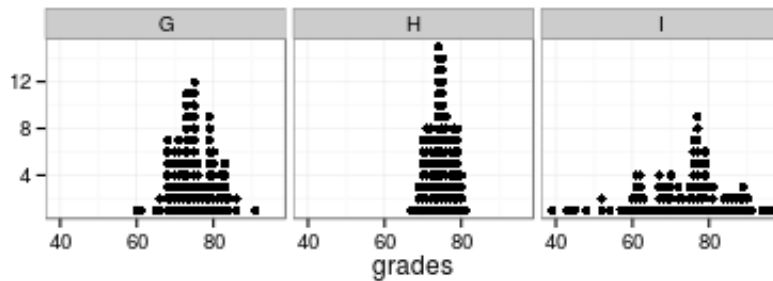
6. Suppose you are choosing which professors' class to enroll in. You have three choices, and have data on the grade distribution for each shown as histograms. Which class seems to have the best grade distribution? Explain.



7. Here's another set of three distributions of exams scores. The density plots shown are essentially smoothed off histograms. Which do you prefer? Explain why.



8. And here's a third set as a dot plot. Each point is one student's exam score – stacked up when several people have the same score. Which class do you prefer? Explain the differences.



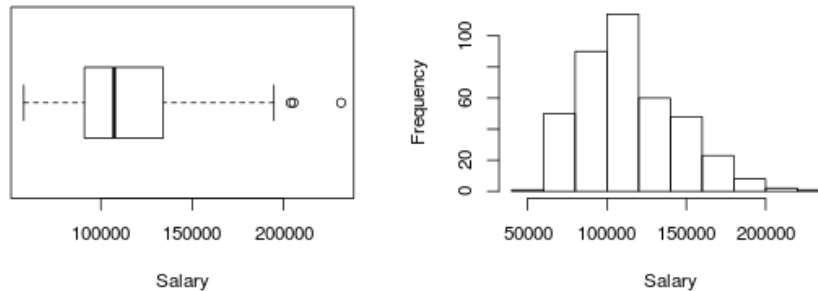
9. When comparing distributions there are several things to consider:
- (a) Comparing location or center (measured by mean or median) tells us which class did best “on average”.
 - (b) Comparing spread (interquartile range or standard deviation) tells us which class is generally closest to its mean.
 - (c) Comparing skew (could be left or right) to symmetric tells us which tail stretches out more. (Let's hope that there are more high grades than low ones.)

In the three problems above, which comparison were you making? For each set of comparisons, fill in center, spread, or skew.

(6) _____ (7) _____ (8) _____

10. Of the three comparisons above, which was easiest and which was hardest? Explain.

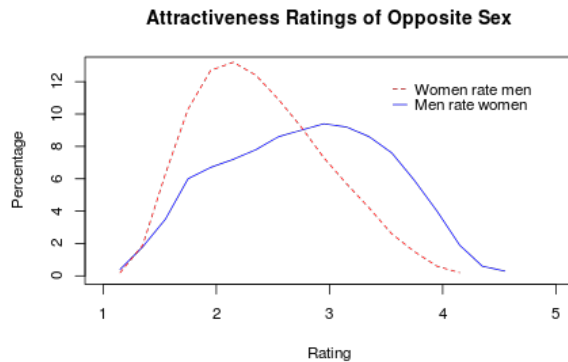
11. You should have read about mean, median, standard deviation, IQR, boxplot and histograms. Apply what you learned to these data: 2009 professor's salaries at a college in the US.



- (a) Is salary skewed (if so which way?) or does it have a symmetric distribution?
- (b) Are any points flagged as outliers? If so, describe them.
- (c) Give approximate values for the median and the first and third quartiles. Also compute the IQR.
- (d) For these data, which is a better summary: mean and standard deviation? or median and IQR? Why?
12. In Christian Rudder's book *Dataclysm* (2014) he shows plots of how men rate the attractiveness of women (data from the online dating site OKcupid) on a scale of 1 to 5 – the solid line in this plot. Y axis is the percentage of women who get this ranking. The line connects what would be the centers at the top of each bar of a histogram, like “hollow Histograms” on page 41 of the OpenIntro stat book.) The dashed line was added by forcing in a perfectly symmetric distribution. Describe the skew of the solid line using the dashed line as a reference.



13. So men have some “biases” about female attractiveness. What if we go the other way and have women rate men? Are the men using OKcupid really ugly? Describe what’s going on here.



Take Home Message:

- To learn about the world, we collect data. Two main types:
 - Categorical – summarize with proportions
 - Quantitative – describe center (mean or meadian) spread (SD or IQR) and shape of distribution (symmetric, left-skewed, right-skewed).
- Plots:
 - Categorical – use bar charts. Pie charts waste ink and are harder to read.
 - Quantitative – Dot plots, histograms, boxplots.
We describe center (mean or median), spread, and skewness based on these plots.

Assignment

- Quizork 1 - Descriptives
A template is posted on D2L. Your completed Quizorks must be exported as a pdf file and uploaded to the D2L dropbox for Quizork2.
- Read pages 18 – for the next class.

Population and Sample

The science of statistics involves using a **sample** to learn about a **population**.

Population: all the units (people, stores, animals, ...) of interest.

Sample: a subset of the population which gets measured or observed in our study.

Statistical Inference: making a statement about a **population parameter** based on a **sample statistic**.

Parameter: a number which describes a characteristic of the population. These values are never completely known except for small populations which can be enumerated. We will use:

μ to represent the population mean.

σ to represent the population's standard deviation (spread).

p to represent a population proportion.

ρ (the Greek letter "rho") for correlation between two quantitative variables in a population.

β_1 slope of a true linear relationship between two quantitative variables in a population.

Statistic: a number which describes a characteristic of the sample and can be computed from the sample. We will use:

\bar{x} to represent the sample mean (or average value).

s to represent the population's standard deviation (spread).

\hat{p} to represent a sample proportion. (We often use a hat to represent a statistic.)

r for correlation between two quantitative variables in a sample.

$\hat{\beta}_1$ slope of the "best fitting" linear between two quantitative variables in a sample.

In this first unit, we will focus on parameter p using sample statistic \hat{p} to estimate it.

Representative Samples

Because we want the sample to provide information about the population, it's very important that the sample be **representative** of the population.

In other words: we want the statistic we get from our sample to be **unbiased**.

Sampling problems:

Convenience Sample is made up of units which are easy to measure, for example, to assess people's opinions on federal college loan programs, we interview students on a university campus. Or to assess the presence of noxious weeds in the state, we select only plots of ground which are within 100m of a secondary highway.

Non-response bias: If people refuse to answer questions for a phone survey, or do not return a mailed survey, we have a “non-response.” When will non-responses cause bias in the results?

Ideal Samples

Ideally we will have a list of all units in the population and can select units **at random** to be in our sample.

Random selection assures us that the sample will generally be representative of the population. A **simple random sample** is selected so that every sample of size n has the same chance of being selected. You can think of this as pulling names out of a hat (although it’s better to use the computer to select samples since names in the hat might not be well mixed).

Simple random sampling is not the only way to get a random sample, and more complex schemes are possible. If you run into a situation in which the population is divided into strata (for example university students live either on campus, in Greek houses, or non-Greek off campus housing, and you want to sample from each) you can use a stratified sample. We will only use simple random sampling (SRS) in this course, and suggest that you consult a statistician or take more statistics classes if you need more complexity.

Non-response bias can be addressed with more work. We would have to do a second (or further) attempt to contact the non-responders, then check to see if they differ (in some important way) from those who responded the first time. Again, this is a situation in which you would need further statistical expertise.

Bias can also result from the wording of a poll, so writing questions is a science in its own right. People tend to try to please an interviewer, so they might, for example, soften their attitudes toward breathing second-hand smoke if they know the interviewer smokes.

Sampling

If we can measure every unit in a **population**, we then have a **census** of the population, and we can compute a population **parameter**, for instance a proportion, mean, median , or measure of spread. However, often it costs too much

time or **money**

so we cannot take a census. Instead we sample from the population and compute a **statistic** based on our **sample**. The science of statistics is all about using data from the sample to make inferences about the population.

This lesson focuses on how to get a good sample. We need a way to select samples which are representative of the population.

The box below contains 241 words which we will treat as our population.

1. Circle ten words in the passage below which are a representative sample of the entire text. (Each person does this, not one per group).

Four college friends were so confident that the weekend before finals, they decided to go to a city several hours away to party with some friends. They had a great time. However, after all the partying, they slept all day Sunday and didn't make it back to school until early Monday morning.

Rather than taking the final then, they decided to find their professor after the final and explain to him why they missed it.

They explained that they had gone to the city for the weekend with the plan to come back and study but, unfortunately, they had a flat tire on the way back, didn't have a spare, and couldn't get help for a long time. As a result, they missed the final.

The Professor thought it over and then agreed they could make up the final the following day. The four were elated and relieved.

They studied that night and went in the next day at the time the professor had told them. He placed them in separate rooms and handed each of them a test booklet, and told them to begin.

They looked at the first problem, worth 5 points. It was something simple about exploratory data analysis. "Cool," they thought at the same time, each one in his separate room. "This is going to be easy."

Each finished the problem and then turned the page. On the second page was written:

For 95 points: Which tire?

2. Explain your method of selection. How did you choose your ten words?
3. Suppose we want to estimate the mean (average) length of all words in our population. Is that a parameter or a statistic?
4. What is the average word length for your sample?

STOP!

Give your sample means to your instructor.

5. To evaluate a method of estimation, we need to know the true parameter and we need to run our method lots of times. That's why we chose a small population which we know has mean word length of 4.26 letters. You are giving your estimate to your instructor so that we can see how well your class does as a whole. In particular we want to know if people tend to choose samples which are biased in some way. To see if a method is biased, we compare the distribution of the estimates to the true value. We want our estimate to be

on target = unbiased.

Then the mean of the distribution matches our true parameter.

While we're waiting to collect all groups sample means we will look at another method:

Simple Random Sampling

- (a) Point your browser to

<http://www.rossmanchance.com/applets/OneSample.html?population=gettysburg>

- (b) Click and erase all the data in the box.
- (c) Copy the word length data from D2L. Select the entire file with control-A, copy it to the clipboard with control-C (or use the right mouse button to copy) and paste it into the applet's data box (use control-V or the mouse option). Click . You should see a plot of all data values with summary information. This is our population of 241 words.
- (d) Click "Show Sampling Options". Change Number of Samples to and Sample Size to . Click . Write out the 10 word lengths in that sample.

6. Record the average (mean) word length for the ten randomly sampled words. Remember, your sample average is an estimate of the average word length in the population. This value should appear on the bottom of the data plot and in the right hand plot of the applet page.
7. Click Generate 1 sample again and record the next mean.
8. Click Generate 1000 samples three times and record the mean and standard deviation of all the sample means. (See upper left of rightmost plot.)
9. If the sampling method is unbiased the estimates of the population average should be centered around the population average word length of 4.257. Does this appear to be the case? Copy the plot here and describe what you see.
10. **Class Samples** Now your instructor will display the estimates from each person in the class. Sketch the plot of all of the sample estimates. Label the axes appropriately.
11. The actual population mean word length based on all 241 words is 4.257 letters. Where does this value fall in the above plot? Were most of the sample estimates around the population mean? Explain.
12. For how many of us did the sample estimate exceed the population mean? What proportion of the class is this?

13. Based on your answer to question 12, are “by eye” sample estimates just as likely to be above the population average as to be below the population average? Explain.
14. Compare the applet plot from question 8 with the plot from 10. Which method is closer to being **unbiased**? Explain.

Examining the Sampling Bias and Variation

To really examine the long-term patterns of this sampling method on the estimate, we use software to take many, many samples. **Note:** in analyzing real data, we only get **one** sample. This exercise is **NOT** demonstrating how to analyze data. It is examining how well our methods work in the long run (with many repetitions), and is a special case when we know the right answer.

We have a strong preference for unbiased methods, but even when we use an unbiased estimator, the particular sample we get could give a low or a high estimate. The advantage of an unbiased method is **not** that we get a great estimator every time we use it, but rather, a “long run” property when we consider using the method over and over.

Above we saw that Simple Random Sampling gives unbiased estimates. People picking a representative sample are often fooled into picking more long than short words. Visual choice gives a biased estimator of the mean.

Even when an unbiased sampling method, such as simple random sampling, is used to select a sample, you don’t expect the estimate from each individual sample drawn to match the population mean exactly. We do expect to see half the estimates above and half below the true population parameter.

If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid. Random sampling avoids this problem. Next we’ll examine the role of sample size. Think of larger samples as providing more information about our population.

Does changing the sample size impact whether the sample estimates are unbiased?

15. Back in the web app, change sample size from 10 to 25. Draw 3000 random samples of 25 words, and write down the mean and standard deviation of the sample means (rightmost plot).

16. Sketch the plot of the sample estimates based on the 3000 samples drawn. Make sure to label the axis appropriately.

17. Does the sampling method still appear to be unbiased? Explain.

18. 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?

19. Compare the spreads of the plots in 8 and 16. You should see that in one plot all sample means are closer to the population mean than in the other. Which is it? Explain.

20. Using the evidence from your simulations, answer the following research questions. Does changing the sample size impact whether the sample estimates are unbiased? Does changing the sample size impact the variability of sample estimates? If you answer yes for either question, explain the impact.

Population Size

Now we examine another question:

Does changing the size of the population impact whether the sample estimates are unbiased?

21. Increase the size of the population. Click “Population” under the data box. How large a population do you now have? Do mean and SD change?
22. With sample size set to , draw a few single samples to see if they look similar, then draw 3000 random samples and record the average (mean) of all the average word lengths.
23. Sketch the plot of the sample estimates based on the 1000 samples drawn. Label the axis appropriately.
24. Does the sampling method still appear to be unbiased? Explain.
25. 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?
26. 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?

27. When we actually collect data, we only get a single sample. In this exercise, we started with a known population and generated many samples. How did we use many samples to learn about properties of random sampling?

A rather counter-intuitive, but crucial fact is that when determining whether or not an estimator produced is unbiased, the size of the population does not matter. Also, the precision of the estimator is unaffected by the size of the population. For this reason, pollsters can sample just 1,000-2,000 randomly selected respondents and draw conclusions about a huge population like all US voters.

Take Home Messages

- Even with large samples, we could be unlucky and get a statistic that is far from our parameter.
- A biased method is not improved by increasing the sample size. The Literary Digest poll: http://en.wikipedia.org/wiki/The_Literary_Digest#Presidential_poll of 2.4 million readers was way off in projecting the presidential winner because their sample was biased. If we take a random sample, then we can make inference back to the population. Otherwise, only back to the sample.
- Increasing sample size reduces variation. Population size doesn't matter very much as long as the population is large relative to the sample size (at least 10 times as large).
- Add your summary of the lesson. What questions do you have?

Assignment

- D2L Exercise 2
- Read

Helper – Hinderer

Do young children know the difference between helpful and unhelpful behavior? A study in *Nature*¹ reported results from a simple study of infants which was intended to check young kids' feelings about helpful and non-helpful behavior.

We'll watch the video of the puppet show the kids watched and see the choice they had to make. The research question is:

Are infants able to notice and react to helpful or hindering behavior observed in others?

Data: Of the 16 kids, 14 chose the “helper” toy and 2 chose the “hinderer”.

Discuss with your group and fill in:

1. What proportion of the infants chose the helper toy? Include the correct notation. (p for a population proportion, or \hat{p} for the sample proportion.)
2. Suppose the infants really had no preference for one toy or the other, and the puppet show had no effect. What sort of results (numbers) would you expect to see?
3. Think back to our “Fraud Detection” activity on the first day of class. What sort of evidence made us think the instructor really could distinguish coin flip sequences from made up sequences? Note: it depended not just on one answer from the instructor, but also on the “background” distribution from the whole class.
4. How could you use coin flips to model a child's choice of toy? For 16 kids?
5. In using the coin, what assumption are you making about the kids' preferences?

¹ Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557-559.

6. In statistical language the idea of “no preference” is called the **null hypothesis** and it is written in terms of the population proportion, p = the true proportion of infants which chose the helper toy, as

$$H_0 : p = 0.5.$$

We also have an **alternative hypothesis**, labeled H_a , which tells us the direction the researchers would like to conclude is true. For this situation, they think there might be a preference toward the helper, so they would like to conclude that H_0 is wrong, and that really

$$H_a : p > 0.5 \text{ is true.}$$

Under H_0 , is it possible that 14 out of 16 infants could have chosen the helper toy just by chance?

7. If infants had no real preference, would the observed result (14 of 16 choosing the helper) be very surprising or somewhat surprising, or not so surprising? How strong do you believe the evidence is against the null hypothesis?

8. Carry Out the Simulation

If you'd like to see that happen, use the <http://shiny.math.montana.edu/jimrc/spin/> Spinner web app. Set number of categories to , labels to , Percentages to , Stop after , Stop after spins, and click to see a simulation of 16 kids choosing helper or hinderer when the two are equally likely.

9. Set trials, Run, and sketch your plot of 1000 trial results.
10. To see how unusual it is to get 14 or more “helpers” add the counts (for 14, 15, 16) in the table below the plot. How many of yours are this extreme? Circle these dots on your plot above. Check with the other groups nearby. Do we all agree?

11. Do you think that babies are just blindly grabbing one of the toys? Explain.

Strength of Evidence

The observed result gets compared to the distribution from the simulation to gauge the evidence against H_0 . That's how the scientific method works. We formulate a hypothesis which can be falsified, then see if the data collected argue against the hypothesis. Sometimes our result provides a lot of evidence against the null model – when the observed result is very unlikely – while other times it has very little evidence against the null model – when the observed result is likely under the null model. To explain to others how likely or unlikely the observed result is under the null model, we report the “strength of evidence” – also called the **p-value**.

The strength of evidence is quantified by answering the question: “If H_0 is true, what proportion of the simulated results are as unusual as (or even more unusual than) the observed result?” For example, consider the results from “Fraud Detection” in Figure 1. This instructor got 9 correct out of 12. The simulation assumed $H_0 : p = 0.5$, and counted the number of heads in 12 flips of a fair coin. (One head represents correctly identifying which of two sequences was generated by coin flips.) The whole process was simulated 1000 times and the number of outcomes at 9 or above on the plot are those as extreme or more extreme as the instructor's score. The chance is $74/1000 = 0.074$ of getting a result this extreme when H_0 is true. We can think of 0.074 as the strength of evidence against H_0 for 9 correct matches. It is the probability of obtaining results as extreme or more extreme when H_0 true.

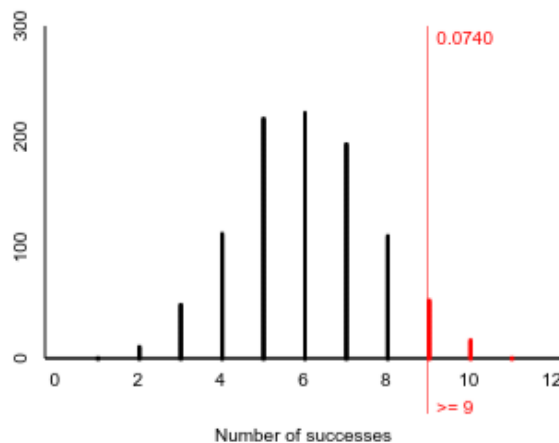


Figure 1: Simulation results obtained from the null model. The outcomes 9 and higher (74 out of 1000 trials) were as or more extreme as one instructor's number correct and indicate the strength of evidence of 0.074.

To help interpret strength of evidence, we offer this picture:



Figure 2: Numeric p-value and strength of evidence

The important point is that **smaller** p-values (in red) provide **stronger** evidence against H_0 .

For this instructor, we would say that there is some evidence that they can detect fraud, but it is not super strong because a p-value of 0.07 is not really small.

12. Use your plot from above to quantify the strength of evidence for the observed result of 14 out of 16 infants choosing the helper toy. Give the numeric p-value and a verbal description of the evidence it provides.
13. What does this suggest about infants making their selections based only on random chance?
14. Put the following steps into their proper order:
 - (a) gather data
 - (b) formulate a hypothesis
 - (c) report strength of evidence
 - (d) simulate a distribution
 - (e) compare observed results to the distribution

Take Home Messages

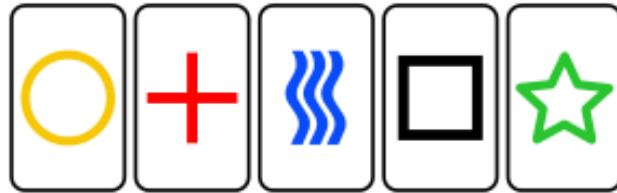
- Setting up null and alternative hypotheses is very important.
They should be set in the planning stages of the study, not after looking at the data. The equals sign always goes into H_0 , but the value we set $= p$ is not always .5. The direction of the inequality in H_a must match the researcher's ideas – what they would like to show. It can be $<$, $>$, or \neq . The latter means they are looking for a difference in either direction.
- It's important to know the definition of the p-value. We assume H_0 is true to compute it. We use a simulation based on the value for p in H_0 to calculate the p-value.
- The idea of p-value is very important in statistics. It will follow us all the way through the course. Stronger evidence means **smaller** p-value. Large p-values mean the data are not unusual under H_0 .
- In any hypothesis test, we report p-values to the reader.
- In the remaining space add your own summary of the lesson. Any questions?

Assignment

- Quizork 3 is posted on D2L. Turn in as a pdf file to the D2L drop box.

Can Humans Sense Each Others' Thoughts?

Mind readers often appear in spooky SciFi movies. (Insert spooky music here.) Is it possible that some people can read minds? In the 1930's Dr. J.B. Rhine at Duke University designed experiments to see if some undergraduate students could tell which card (see the five "Zener" cards below) a "sender" was looking at. The deck of cards (5 of each type) was shuffled and a card drawn at random. After each attempt, the card was returned to the deck and the deck was reshuffled (we call this sampling with replacement). Therefore each of the five card types has an equal chance of occurring at any draw.



Rhine found one exceptional subject in his extrasensory perception (ESP) research, Adam Linzmayer, an economics undergraduate at Duke. Linzmayer went through the experiments in 1931, and correctly identified 36% of 25 cards as the "receiver" in the study. We will start by investigating this student's result as Rhine did, hoping to prove that Linzmayer does have extrasensory perception.

Step 1. State the research question.

1. Based on the description of the study, state the research question.

Step 2. Design a study and collect data.

Linzmayer correctly identified the card 9 out of 25 times in one trial.

2. What were the possible outcomes for 1 attempt (guessing one card) that Rhine would have recorded?
3. Your answer above gives the outcomes of the variable of interest in the study. Is this variable quantitative or categorical?

Step 3. Explore the data.

With categorical data, we report the number of successes or the proportion of successes as the "statistic" gathered from the sample.

4. What is the sample size in this study? $n =$
5. Determine the observed statistic and use correct notation to denote it.
6. Could Linzmayer have gotten 9 out of 25 correct even if he really didn't have ESP and so was randomly guessing between the five card types?
7. Do you think it is likely Linzmayer would have gotten 9 out of 25 correct if he was just guessing randomly each time?

Step 4. Draw inferences beyond the data.

Two things could have happened:

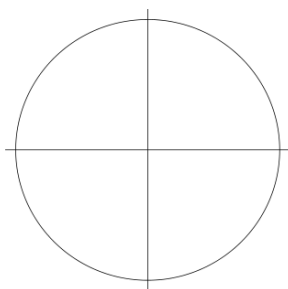
- He got over one third correct just by random chance – no special knowledge.
 - He is doing something other than merely guessing and perhaps has ESP.
8. Of the two possibilities listed above, which was Rhine trying to demonstrate (the alternative) and which corresponds to “nothing happening” (the null)?
 9. What is the value of the true parameter if Linzmayer is picking a card at random? Give a specific value and use correct notation to denote it.
 10. If Linzmayer is not just guessing and did have ESP, what values might the true proportion take on? Again, use correct notation to denote this range of values.

Is the observed statistic (9/25) in this interval?

11. When writing the null and alternative hypotheses, we may use words or we may use symbols. Rewrite the null and alternative hypotheses in both words and notation by combining your answers from 8 – 10.
 H_0 :

H_a :

12. Think of a “spinner” on a game board. How would you subdivide and color it so that each spin would be equivalent to Linzmayer randomly guessing one card and getting it right/wrong with the null hypothesis probability. (Hint: you do not need 5 segments.) Sketch your spinner on the circle below and shade the area where he gets it right just by chance. Put a paper clip on the paper with pen to hold one end on the center. Spin 25 times and count the number of successes.



13. Now we'll use a web app to speed up the process. Go to <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/> and click “Enter / Describe Data” under the “One Categ.” menu. Enter the counts to show how many Linzmayer got right and got wrong. (These should add up to 25, but neither is 25.) Click “Use These Data” and record his proportion correct.
14. Now from the “One Categ.” menu choose “Test”. and enter the value from 9 as the True Proportion. Run 5000 samples and sketch the plot below.
15. Check the summary statistics inside the plotting window. Does the mean agree with the null or alternative hypothesis? Explain why it should.
16. What proportion did Linzmayer get correct?
Type that value in to the box just after “than” below the plot. Select the direction (less, more extreme, or more) based on the alternative hypothesis in 11. Click and record the proportion of times this occurred.
Would you consider this an unlikely result?

17. Go back to figure 2 to report the strength of evidence against H_0 . Give the numeric and the verbal strength of evidence.

Step 5: Formulate conclusions.

Based on this analysis, do you believe that Linzmayer was just guessing? Why or why not?

Are there ways other than ESP that a person could do well as a “receiver”? Explain.

Another part of the scientific method is a reliance on replication. Other scientists tried to replicate this study and could not find another person like Linzmayer.

Take Home Messages

- This activity was much like the previous one (Helper–Hinderer), except that the null hypothesis value was not one-half. (Here “at random” was 1 of 5, not 1 of 2)
- Again note how H_0 is used to compute the p-value. The alternative comes into play only when we need to see which direction to count as “more extreme”.
- Both examples we’ve done so far have used a $>$ alternative, but that is not always the case.
- And finally: other reporting on Linzmayer suggests that he was cheating, rather than reading minds.
- Use the remaining space for any questions or your own summary of the lesson.

Interval Estimate for a Proportion

If we call someone a “rat”, we don’t mean that they are nice to be around, but rats might not deserve their bad reputation. Researchers examining rat’s capacity for empathy designed a study in which a pair of rats were placed in the same cage. One was trapped in a cramped inner cage, while the other could move around much more, and could also free the trapped rat if it chose to do so. Of thirty pairs of rats in the experiment, 23 of the free rats released the trapped rat even though they then had to share the available food.



The lab rats used in the study are genetically identical to other rats of the same strain, and can be assumed to be a “representative sample” from the population of all rats of this strain. Researchers need a good estimate of the true proportion of these rats who would free another rat trapped in the inner cage.

Step 1. State the research question.

1. Based on the description of the study, state the researcher’s need as a question.
2. There is a light gray vertical line in the center of the plot. What is the value (on the x axis) for this plot and why is it marked?

Step 2. Design a study and collect data.

2. What actions of the free rat will be recorded?
3. Your answer above gives the outcomes of the variable of interest in the study. Is this variable quantitative or categorical?

4. What is the parameter the researchers were interested in? Describe it in words and use proper notation to denote it.

Step 3. Explore and summarize the data.

5. What is the sample size in this study? $n =$
6. Determine the observed statistic and use correct notation to denote it.
7. If the experiment were repeated with another 30 pairs of rats, do you think you would get exactly 23 who opened the cage again? Explain.

Step 4. Draw inferences beyond the data.

The last point is simple, but really important. When we repeat the same experiment, we do not get exactly the same results. Why is that? (Yes, you need to write an answer right here! The future of the world – no, I mean your success in this course – depends on it.)

We know exactly what proportion of rats in the sample showed empathy, and that number makes a good estimate of the same proportion of empathetic rats in the population. However, the fact that not all rats, and not all samples are the same tells us we need to expect some variation in our sample proportion when we repeat the experiment.

A single number like the one you computed in 6 does not tell the whole story. We want to let our reader know “how good” this estimate is. One way to report the quality of an estimate is to give a range of values – an interval estimate – instead of a single “point estimate”.

Because we now have easy access to computers, we can run a **simulation** to see how variable the statistic might be. We only get one sample of real data, but we can create lots of simulated datasets which represent other values which might have been observed.

8. Your group will get 30 cards on which you will write the observed outcomes from (2) – one for each of the 30 pairs. We don't care about order, just that we get the right numbers of cards for each outcome. Next we simulate another experiment on another sample of 30 rat pairs. We can't actually get more rats and study them, so we “recycle” the numbers we have.
- (a) Shuffle your cards and draw one at random. Record the outcome for this pair.
 - (b) Replace the card into the deck. This is a simple but powerful idea. By sampling **with replacement** we have the same conditions for every draw, and the probability of each outcome stays the same. Shuffle, draw a card, and record the outcome.
 - (c) Repeat until you have 30 outcomes chosen at random. What proportion of your rats were freed?

The process you just used is called **bootstrapping** (which means to make something out of little or nothing), and the 30 outcomes are called a bootstrap **resample**. It's not a sample – we only get one of those – and we can repeat the resampling process many times.

9. Reshuffling is slow, so we want to speed up the process by using the computer. Our goal is to see what other outcomes we might have gotten for different samples of 30 rat pairs. We will again use the One Categ. web app at <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/>. Select “Enter / Describe Data” to enter the rat data to look like:

Freed	23
Not	??

Then choose “Estimate” from the “One Categ” menu. What proportion of the rats were freed in your one resample?

10. Now resample 100 times and copy the picture you get here.

Where is the distribution centered?

How spread out are the sample outcomes? (SE stands for standard error, which is the standard deviation of the resampled values.)

11. The center should seem reasonable. Why is the distribution centered at this value?

12. You should have 101 resampled values in your plot. If not, go back to “Data Entry” and change the spelling on the labels (caps to lower case, or vice versa). Then come back to “Estimate” and click just once.
Below the plot we have options for confidence limits for our interval estimate.
 - (a) Click and count: How many red points in the left tail?
How many reds in the right tail?
How many blue points in the middle?
 - (b) Click and count: How many red points in the left tail?
How many reds in the right tail?
How many blue points in the middle?
 - (c) Click and count: How many red points in the left tail?
How many reds in the right tail?
How many blue points in the middle?
 - (d) Explain how the confidence limit is related to the number of blue points.
 - (e) Play with the “Confidence Limit” buttons more to explain: How are the endpoints of the interval estimate related to the colors of the points in the plot?
 - (f) Predict: what will happen to the numbers in each tail for, say, a 90 % interval, if we double the number of resamples?
 - (g) Click again and explain: were you right?

13. We need to spend more time on the meaning of “Confidence”, but first let’s review: Explain how one dot in the plot was created. (I suggest going back to how you did it manually in 8.)

Take Home Message

Several very BIG ideas:

- We only get one sample, but we can create many “resamples” using sampling with replacement (also called bootstrapping).
- Interval estimates are better than point estimates.
 - They don’t pretend to be exact. Any exact value is almost certainly wrong.
 - By looking at the width of an interval we can evaluate the quality of the data. Wide intervals are not very useful. Skinny intervals are more informative.
 - We can pretend that we know the true value of a parameter in order to test our methods.
 - Our methods are not “fail safe”, but are actually designed to have a certain error rate, for example, 5% of the time our 95% confidence intervals will fail to cover the true parameter.
- Our confidence in a particular interval is actually in the process used to create the interval. We know that using this process over and over again (go out and collect a new random sample for each time) gives intervals which will usually cover the true value. We cannot know if a particular interval covered or not, so you have to be tolerant of some uncertainty.
- Use the remaining space for any questions or your own summary of the lesson.

What Does “Confidence” Mean?

We’ve seen that we can obtain several different interval estimates for a parameter. Each is centered at the statistic (here \hat{p}), but they have different widths.

To understand the meaning of the term “confidence”, you have to step back from the data at hand and look at the process we use to create the interval.

1. Select a random sample from a population, measure each unit, and compute a statistic like \hat{p} from it.
2. Resample based on the statistic to create the interval.

To check to see how well the techniques work, we have to take a special case where we actually know the true parameter value. Obviously, if we know the value, we don’t need to estimate it, but we have another purpose in mind: we will use the true value to generate many samples, then use each sample to estimate the parameter, and finally, we can check to see how well the confidence interval procedure worked by looking at the proportion of intervals which succeed in capturing the parameter value we started with.

Again go to <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/> and select “Confidence Interval Demo” from the “One Categ” menu.

The first slider on this page allows us to set the sample size – like the number of rat pairs in the experiment. Let’s start with .

The second slider sets the true proportion of successes for each trial or spin (each rat pair). Let’s set that at or 75% which is close to the observed \hat{p} of the rat study.

You can then choose the number of times to repeat the process (gather new data and build a confidence interval [CI]: 10, 100, 1000 or 10K times) and the level of confidence you want (80, 90, 95, or 99%).

We’ll start with simulations of a % CI.

The upper plot shows 100 \hat{p} ’s – one from each of the 100 simulations.

The second plot shows the interval estimate we get from each \hat{p} . These are stacked up to put smallest estimates on the bottom, largest on top. The vertical axis has no real meaning.

14. Click on a point in the first plot to see its corresponding CI in the second plot. Especially try the largest and smallest points. Which intervals do they create (in terms of left or right position)?

15. There is a light gray vertical line in the center of the plot. What is the value (on the x axis) for this plot and why is it marked?
16. What color are the intervals which do not cross the vertical line? How many are there?
17. What color are the intervals which cross over the vertical line? How many are there?
18. Change the confidence level to 95%. Does the upper plot change? Does the lower plot? Describe any changes.
19. If you want an interval which is stronger for confidence (has a higher level), what will happen to its width?
20. Try each confidence level in turn and record the coverage rate (under plot 2) for each.

80	90	95	99

Take Home Message

- Interval estimates are better than point estimates.
- Our confidence in a particular interval is actually in the process used to create the interval. We know that using this process over and over again (go out and collect a new random sample for each time) gives intervals which will usually cover the true value. We cannot know if a particular interval covered or not, so you have to be tolerant of some uncertainty.
- Use the remaining space for any questions or your own summary of the lesson.

MIT – the Male Idiot Theory

The usually serious *British Medical Journal* enjoys a bit of fun in each Christmas issue. In December 2014 they published a study of the MIT – “Males are Idiots Theory” based on data collected from the Darwin Awards.

“Winners of the Darwin Award must die in such an idiotic manner that ‘their action ensures the long-term survival of the species, by selectively allowing one less idiot to survive.’²⁰ The Darwin Awards Committee attempts to make a clear distinction between idiotic deaths and accidental deaths. For instance, Darwin Awards are unlikely to be awarded to individuals who shoot themselves in the head while demonstrating that a gun is unloaded. This occurs too often and is classed as an accident. In contrast, candidates shooting themselves in the head to demonstrate that a gun is loaded may be eligible for a Darwin Award—such as the man who shot himself in the head with a ‘spy pen’ weapon to show his friend that it was real.¹⁸ To qualify, nominees must improve the gene pool by eliminating themselves from the human race using astonishingly stupid methods. Northcutt cites a number of worthy candidates.^{12–21} These include the thief attempting to purloin a steel hawser from a lift shaft, who unbolted the hawser while standing in the lift, which then plummeted to the ground, killing its occupant; the man stealing a ride home by hitching a shopping trolley to the back of a train, only to be dragged two miles to his death before the train was able to stop; and the terrorist who posted a letter bomb with insufficient postage stamps and who, on its return, unthinkingly opened his own letter.”²

The authors examined 20 years of data on the awards, removing awards given to couples “usually in compromising positions” so that each remaining winner was either male or female. Of the 318 remaining awards, 282 were given to males and 36 were awarded to females.

They ask the question: “If we look only at people who do really stupid things, what is the gender breakdown?” or “Are idiots more than half male?”

1. What population is represented by these winners of the Darwin Awards?
2. What is the parameter of interest?
3. What statistic do we obtain from the sample? Give proper notation, the statistic’s value, and explain it in words.

²Lendrem, B. A. D., Lendrem, D. W., Gray, A., & Isaacs, J. D. (2014). The Darwin Awards: sex differences in idiotic behaviour. *BMJ*, 349, g7094.

4. Looking at the research question, “Is the group of idiots in the world more than half male?”, we set up the null hypothesis to assume “just half” and the alternative to be “more than half” male.

- (a) State null and alternative hypotheses in symbols and words.

H_0 :

H_a :

- (b) How would you mark cards and randomly draw from them to obtain one simulated proportion drawn from the distribution when H_0 is true?

- (c) Input the data under “One Categ” in <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/> and use the “Test” page to obtain the distribution of 1000 or more sample proportions under H_0 . Sketch the picture you get here.

- (d) How unusual is the sample statistic from 3 relative to the distribution you created? Explain in words where it falls relative to the plotted points.

- (e) How strong is the evidence against the null hypothesis? What do you think about the idea that idiots are half male?

5. Instead of considering a test of the true population proportion, we will switch gears and now estimate it.
 - (a) What is our “point” estimate of the true proportion of idiots who are male (the sample statistic)?
 - (b) In order to generate simulated data,
 - i. How many individual “idiots” do we generate for one resample?
 - ii. Explain how you would mark 318 cards and use them to simulate the gender of one individual, and then another.
 - iii. What probability of being male is used?
 - iv. After resampling 318 individuals, what number do you compute?
 - (c) Use the web applet to create 1000 or more resamples from the original data. Plug in the numbers from the data.
 - i. Where is this distribution centered?
 - ii. What is the spread of the distribution of resampled proportions?
 - (d) Find a 95% confidence interval for the true proportion of idiots who are male.
 - (e) Explain what the word “confidence” means for this confidence interval.
6. Compare results from the hypothesis test and the interval estimate. If the null hypothesis is true, what value should be included in the 95% CI? Explain. Do the two methods agree to some extent?

Take Home Message:

- You just did two inferences on the same parameter. First, we tested the null hypothesis that half the world's idiots are male.
You should have reported very strong evidence against that null hypothesis (less than 1/1000). We can feel quite confident that the true proportion of males in this exclusive group is more than one half.
- Secondly, we computed a 95% confidence interval for the true proportion of idiots who are male and you interpreted the interval. In 5e you should have explained the long-run coverage property of the method.
- There is a correspondence between testing and estimating. The values inside the interval you found are consistent with the data, or **plausible**. Because 0.50 is not in the interval, it is not a plausible value for this parameter.
- Use the remaining space for any questions or your own summary of the lesson.

Unit 1 Review

Vocabulary Define each term:

- sample
- population
- statistic
- parameter
- types of variables
- measures of center
- measures of spread
- estimation bias
- Null hypothesis
- Alternative hypothesis
- Strength of evidence
- Confidence interval

Simulation

1. If we repeat the “Helper – Hinderer” study and 10 of the 16 infants chose the helper (6 chose hinderer):
 - (a) How would you assess the strength of evidence using the same simulation we already performed?
 - (b) What strength of evidence against the null hypothesis does this new data provide?

- (c) If 13 kids chose the helper toy, what is the strength of evidence against the null hypothesis?

- (d) If we redid the study with 8 infants, and 7 chose the helper, is this stronger, weaker, or the same amount of evidence against the null hypothesis?
- (e) Explain how would you rerun the simulation for only 8 infants.

- (f) Perform the simulation for 8 infants and compare the strength of evidence provided by 7 choosing the helper. Was your hunch correct? Explain any differences.

- 2. A German bio-psychologist, Onur Güntürkün, was curious whether the human tendency for right-sidedness (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.
 - (a) What parameter is of interest?

 - (b) What statistic do we obtain from the sample? Give proper notation, the statistic's value, and explain it in words.

- (c) We can set the null hypothesis as we have before, but don't know before collecting data whether the alternative should be greater or less than one half. We therefore use a "two-sided" alternative with a \neq sign.

- i. State null and alternative hypotheses in symbols and words.

H_0 :

H_a :

- ii. How would you mark cards and randomly draw from them to obtain one simulated proportion drawn from the distribution when H_0 is true?

- iii. Use the 'One Categ – Test' applet to obtain the distribution of 1000 or more sample proportions under H_0 . Sketch the picture you get here.

- iv. How unusual is the sample statistic from 2b relative to the distribution you created? Explain in words where it falls relative to the plotted points.

- v. How strong is the evidence against the null hypothesis? What do you think about the idea that only half of couples lean right when kissing?

- (d) Now estimate the true population proportion.
- i. What is our “point” estimate of the true proportion of couples who lean right?
 - ii. In order to generate simulated data,
 - A. How many couples do we generate for one resample?
 - B. Explain how you would mark 124 cards and use them to simulate the lean of one couple, and then another.
 - C. Each couple leans right with what probability?
 - D. After resampling 124 individuals, what number would you compute?
 - iii. Use the web applet to create 1000 or more resamples from the original data.
 - A. Where is this distribution centered?
 - B. What number describes the spread of the distribution?
 - iv. Compute a 99% confidence interval.
 - v. Explain what the word “confidence” means for this situation.
- (e) Compare results from the hypothesis test and the interval estimate. If the null hypothesis is true, what value should be included in the 99% CI? Explain. Do the two methods agree to some extent?

Unit 2

Does Music Help Us Study?

Suppose you have a big test tomorrow and need to spend several hours tonight preparing. You'll be reviewing class notes, rereading parts of the textbook, going over old homework – you know the drill.

1. Which works better for you: turn on music, or study in silence? Circle one:

- A. With music
- B. In silence

If you like to study with music (at least some times) describe:

- (a) what volume?
- (b) with lyrics? or instrumental?
- (c) what general category do you prefer?

2. A researcher wants to know if some types of music improve or hurt the effectiveness of studying. Suppose we want to address this question by getting college students to fill out a survey.

- (a) The survey will ask for details on the music type each student prefers for studying, but we will also need a way to measure how effective their studying is. How could we measure a **response** to use for comparison – to see how much people are learning while studying?
- (b) In discussing the response, you probably found difficulties which make it hard to compare people. What differences in students make it hard to get a clear comparison between different music types? List at least three variables that we should consider. For each: is it categorical or quantitative? You should have at least one categorical and one quantitative variable.

3. Another option for studying the effect of music type on studying is to **assign treatments**, as in this study from 2014.

Perham, N. and Currie, H (2014). Does listening to preferred music improve reading comprehension performance? *Applied Cognitive Psychology* **28**:279–284.

They used four levels of the variable **sound**: “disliked lyrical music (DLYR), liked lyrical music (LLYR), non-lyrical music (NLYR) and quiet (Q)” and each subject chose music they liked with lyrics (LLYR), while the instrumental music (NLYR) was picked by the researchers, and subjects were screened to be sure they did not enjoy “thrash” music, which was used for DLYR. Subjects were told to ignore the music, and had to read 70 lines of text, then answer four multiple choice questions about the reading (taken from SAT exams). They repeated the task for three more readings (with 4 questions each), and the proportion correct was recorded.

- (a) Is use of the SAT questions an improvement over your choice of response in 2a? Explain why or why not.

- (b) Is use of the four **sound** treatments an improvement over asking students how they study? Explain why or why not.

4. In an **experiment** levels of the explanatory (treatment) variable are **assigned** to subjects (or units if people are not involved). In order to allow statistical inference, we should assign the treatments at random, making it a **randomized experiment**. In an **observational study** we simply record levels or values of the explanatory variable instead of assigning them. Looking back at the studies above, which was an experiment?

Which was an observational study? Explain how you know this.

Advantages of Randomized Experiments

To make sure we're all thinking of the same response for our study on the effect of music while studying, we'll focus on using the SAT reading comprehension scores as our response. Music (or quiet) will be played while our subjects read and answer the questions.

6. In 2b, above you mentioned several attributes of people which would indicate who does better on a test. One such variable would be IQ. Smarter people tend to get higher scores on the SAT.

We refer to a variable like IQ as a **lurking** variable when we do not measure it and take it into account. What other lurking variables did you identify in 2b (or add some here to get at least three) which would cause some people to do better on SAT than others?

7. If we don't measure IQ and don't adjust for it, we won't be able to tell whether one group did better because it had higher mean IQ, or because they were assigned the more effective treatment. Let's see what happens to mean IQ (and another variable - SAT prep) if we randomly separate 12 people into treatment (music) and control (quiet) groups of 6 each.

Treatment			Control		
Name	IQ	SAT prep	Name	IQ	SAT prep
Andy	104	Y	Peter	106	Y
Ben	118	Y	Maria	90	N
Betty	79	N	Marti	97	N
Jorge	94	Y	Mikaela	98	N
Kate	106	N	Patty	89	N
Russ	88	Y	Shawn	85	Y

Mean IQ of treatment group: 98.2

Mean IQ of control group: 93.8

Difference in means: 4.4

Write Name, IQ, and whether or not they took an SAT prep class (Yes or No) for each person on an index card. (If the cards are already started, check that you have the right names and values.)

- (a) Mix the cards thoroughly, and deal them into two piles of six each, labeling one "T" and the other "C". Compute the mean IQ for each group and take the difference ($T - C$).

- (b) Plot your difference as instructed by your teacher.

8. As with many techniques in statistics, we want to see what happens when we repeat the process many times. Doing this by hand many times gets tedious, so we will use the computer to shuffle and compute means for us.

Go to: <http://shiny.math.montana.edu/jimrc/IntorStatShinyApps> and click “Enter Data” under One of each.

Select and .

group	iq
t	104
t	118
t	79
t	94
t	104
t	90
c	97
c	98
c	88
c	106
c	89
c	85

Click several times. Record the difference in means for this shuffle.

9. As we said above, we need to think about repeating the shuffling process over and over. Click and sketch the right-hand plot (above). Describe the center, spread, and shape of this distribution.

center

shape

spread

10. Do we get the same pattern in the right hand plot if we run another batch of shuffles, say 10,000 this time? Do center, shape, and/or spread change?
11. Does randomization always make mean IQs the same between the two treatment groups? Explain.

12. Does randomization tend to balance out mean IQ in the long run, after many trials? Explain.
13. **Very Important Question:** In general, how similar are group mean IQs when we randomly assign people into two groups?
14. Another lurking variable would be the fact that some people have taken a short course as an SAT prep and others haven't. If the course does what it claims, then it could be the reason for one group to score higher than the other. We will look at the proportions who have taken an SAT prep course in the treatment and control groups.
- (a) Is “took SAT prep course” a categorical or quantitative variable?
- (b) Again go to the web app and click “Enter Data” under Two Categ. Suppose we are randomly assigning 100 people to our two groups, and that 28 of them have taken SAT prep, 72 have not. Put in 14 Successes for each group, and 36 Failures for each. Click Use These Data and make sure that we have 50 in each group (1 = treatment, 2 = control), 28 “Successes” (took SAT prep) and 72 “Failures”. (This is just one way to get the right totals at the edges, or margins, of the table. Because we'll be doing lots of shuffles, the groups do not have to start out as balanced, but it makes it easy to get the right marginal totals.) Click “Test”
fixme:
and click Shuffle. (Leave the Cards option checked)
What totals do you get for blue and green cards? What do the colors represent?
- (c) For this shuffle, what is the difference in proportions? Click Shuffle four more times, and record all 5 differences.

(d) Change number of shuffles to and click . Sketch your plot here.

(e) Compare with other groups. Do the pictures look the same?

center

shape

spread

15. When we randomly assign people to two groups:

(a) Is it possible for a categorical lurking variable like SAT prep to be imbalanced across the two groups? Explain.

(b) Will the lurking SAT variable “usually” be poorly balanced across the two groups? Explain.

16. In general, how similar is the proportion of people who have taken SAT prep in the treatment group to the same proportion in the control group?

17. If you ran an experiment where you randomly assigned people to either listen to music or silence, would you have to worry about the effect of SAT prep courses on the results? Explain.

Take Home Messages

- Vocabulary: response variable, explanatory variable, experiment, randomized experiment, observational study, lurking variable.
- This lesson is critical for understanding how experiments differ from observational studies. When we assign treatments at random, we “even out” any lurking variables, so we can say that differences we observe are caused by the explanatory variable (the treatment). We call this **causal inference**.
- Our use of the web app today was to see what happens to means of a lurking variable when we randomly split people into two groups. You should have concluded that the means tend to be approximately equal (difference in means is centered at zero), and that the distribution of the difference in means is symmetric. Any positive value has a negative counterpart which just involves swapping the labels ($T \longleftrightarrow C$). We will use these apps later to evaluate treatment effects, but today we are only looking at lurking variables.
- Use the remaining space for any questions or your own summary of the lesson.

Bootstrap Confidence Interval for μ

Traditionally in intro stat classes we have used z and t -tables but we will use the power of computers to do the work for us.

Problem:

We need to know the sampling distribution to know how far away our statistic might be from our parameter. We know the sampling distribution of \bar{x} is centered at the population mean, μ , and we know some things about its spread and shape. However, the sampling distribution of \bar{x} depends on the unknown parameter μ . How can we estimate μ ?

Solution:

Use the “Resampling” or Bootstrap distribution as a substitute for the unknown sampling distribution. BUT:

We only draw **one sample from the population!**

Hang onto that idea, because we will use our one sample in an almost magical way to generate something very much like the sampling distribution.

A **bootstrap resample** is the same size as the original data, and consists of data points from the original data. The only difference is that the resampling is done “with replacement” so a bootstrap resample typically contains several replicates of some values and is missing other values completely. We can repeat this process many times and store the statistics generated from each resample. The result is a bootstrap distribution (or a resampling distribution) which can be used as a replacement for the unknown sampling distribution. In particular, we can use the spread (standard error) of the bootstrapped sample statistics as a substitute for the spread (standard error) of our statistic.

Go to this bootstrapping applet:

<http://www.math.montana.edu/~jimrc/randomization/BootDemo.html>

1. The counts shown are all the values in the population, which are amounts (in 10’s of dollars) stat students in a prior semester spent on textbooks.

(a) Click Sample and we’ll get a random sample of size 8 from this population. The population then disappears because we never can observe an entire population. Some of your numbers might be the same, but they came from different individuals in the population. Click Get New Sample at the bottom of the page, and you’ll get a new sample. How many samples do we collect in one study?

(b) Click 1 Resample and watch what happens. Click slower 1 or 2 times and watch it again. What is this button doing?

- (c) Slow it down to where you can answer these questions: For one resample, which of the original eight values got used more than once? which not at all?
- (d) Get 8 cards from your instructor and write each of the 8 values in your sample on a card. Create your own bootstrap resample to mimic what the computer does. Which of these methods works?
 - i. Select one card at random, leave it out, and select another card. Continue until you use all the cards.
 - ii. Select one card at random and write down its value. Replace it, reshuffle, and select another. Continue until you've written down eight values.
- (e) What statistic are we interested in (from the sample)? Compute it for the resample.
- (f) Click in the “Many Resamples” choices.
 - i. Explain what values are being plotted.
 - ii. One of our favorite quiz/exam questions is “What does one dot represent?”. Explain where the values came from and what statistic was computed from them.
- (g) Click in the “Many Resamples” choices. Write down the interval estimate. Count (approximately) how many circle centers are outside the red lines at the left and at the right.

Repeat twice more. Write down each confidence interval and guess how many points fall outside each.

- (h) Click 1000, 5000, and 10000 in turn. Write down three CI's for each. Compare the CI's. Are some groups off-center compared to others? More variable?
- (i) Go back to 500 resample. What happens to length of intervals when we change confidence levels? Hint: choose a different confidence level with the buttons, then click again to obtain the interval.
- from 95% to 99% -- intervals _____
- from 95% to 90% -- intervals _____
2. When we started, we saw the whole population of counts, and which has true mean $\mu = 34.5$
- (a) Look back at all the intervals you wrote down. Which ones contain the true value?
- (b) Click . Compute a 90% confidence interval for the mean using 1000 bootstrap iterations. Show the interval and write “covered” if it contains 34.5, “missed” if it does not. Do again, write the interval and “missed” or “covered”.

Repeat 8 more times to get a total of 10 samples with 2 intervals for each.

Does coverage depend more on the sample or on the particular resample?

3. With proportions we used $\hat{p} \pm 2SE$ as our confidence interval. For means, we have extra variation from not knowing the spread, σ , so the correct multiplier depends on sample

size as well as confidence level. For sample size $n = 8$, the multiplier is $t_7^* = 2.36$ for 95% confidence, 3.50 for 99% confidence, and 1.89 for 90% confidence. The web app shows standard error of the resampled means as SD, so we use this as our SE. Build 90, 95, and 99% CI's using the $\bar{x} \pm t^*SE$ method. Also write the bootstrap intervals to compare.

- (a) Compute the mean of your sample (from the 8 values, not the “Mean” printed)

$\bar{x} =$

- (b) a 90% CI for μ is (show work)

- (c) a 95% CI for μ is (show work)

- (d) a 99% CI for μ is (show work)

4. Is there a pattern when you compare the two methods? Are bootstrap percentile methods always wider? shifted? relative to the $\bar{x} \pm t^*SE$ intervals?

5. Challenge: based on what you've seen so far in this course what will happen to our interval estimates if we change sample size from 8 to 4? From 8 to 16?

Will smaller sample size shift the center?

Will smaller sample size change the width?

Will larger sample size shift the center?

Will larger sample size change the width?

Try it and record what happens to center and spread. (Yes, it is important to write it down. It will show up on the exam.)

Take Home Messages

- We only get one SAMPLE, but from it we can generate many resamples.
- We can use the resampling distribution to see how much samples vary. It is a substitute for the unknown sampling distribution.
- Whether the interval includes the parameter or not depends mainly on our luck in sampling. Most samples give statistics close to the parameter, but some can be farther away.
- We can use the bootstrap information in two ways:
 - to compute the SE of the statistic
 - to find percentiles of the resampling distribution.

Either method can give a confidence interval. With symmetric data, the two should agree well. These data are skewed to the right, and the bootstrap percentile intervals are preferred.

- Use the remaining space for any questions or your own summary of the lesson.

Peanut Allergies

Peanut allergies are so common in children that all parents have to closely watch the snacks they bring to an elementary classroom party. Any hint of peanuts in a cookie can endanger some kids. A recent article suggests that peanut allergies can be prevented by giving babies (ages 4 to 10 months) some peanut protein every week until they are 5 years old. They randomly divided 500 infants into “Eat peanuts” and “avoid peanuts” groups and followed them to age five when each child was checked for peanut allergies.

The researchers want to answer this question:

Does feeding children peanut protein prevent peanut allergies?

Discuss the Following Questions

1. Is there a treatment condition in this study? (If so, what?)
2. What is the response variable in this study?
3. Are the variables above quantitative or categorical?

Results: 5 of the 245 infants getting peanut protein, (2%) showed allergic symptoms to peanuts, whereas in the peanut avoidance group, 35 (13.7%) of 255 infants developed allergic symptoms to peanuts. (BTW, the two groups started with equal, somewhat larger numbers of infants, but there were dropouts who are assumed ignorable.)

4. Organize the results into a 2 by 2 table

	Peanuts (1)	Avoiders (2)	total
Allergic			
Not Allergic			
Total			

5. Of the 245 subjects assigned to the eat peanuts, what proportion improved? We will label this \hat{p}_1 because it is an estimate of p_1 , the true proportion who would become allergic if all infants ate peanut protein. As in the notes for day 4, we ornament p with a “hat” on top to show that this is an estimate (or a statistic) computed from the observed sample. Finally, the “1” subscript is to demark the first (treatment) group.

6. Of the 255 subjects assigned to the control condition, what proportion improved? We’ll call this \hat{p}_2 , using a “2” for the control group.

7. Find the difference between the proportion of subjects assigned to the “eat peanuts” condition that became allergic and the proportion of subjects assigned to the control condition that became allergic. $\hat{p}_1 - \hat{p}_2 =$

8. What proportion of all 500 subjects improved? This is called a marginal distribution because it just uses totals. If the treatment has no effect, then this will be a good estimate of the true overall probability that any infant will develop peanut allergy, so label it \hat{p}_T where T goes with “Total”.

9. 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 risk of becoming allergic to peanuts in these subjects; (2) the differences between the two treatment groups; and (3) whether or not the data appear to support the claim that peanut eating is effective.

In statistics, we use data from a sample to generalize back to a population. Here are some **critical questions**:

- Does the higher allergy rate in the control group provide convincing evidence that the peanut eating 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 500 subjects into groups (i.e., the luck of the draw)?
- Is it reasonable to believe the random assignment alone could have led to this large of a difference?
- Just by chance did they happen to assign more of the subjects who were going to improve into the peanut treatment group than the control group?

One way to examine these questions is to consider what you would likely see if 40 of the 500 kids were going to develop the allergy (the number of infants who did in our sample) regardless of whether they ate peanuts or not. If that is the case, you would have expected, on average, about 20 of those subjects to end up in each group (the null model suggests this).

We will answer this question by using a web applet to conduct a **permutation** (or randomization) test which lets us see the results one can get just due to variation in random assignment. We'll operate under the null model assumption that the control and peanut conditions have no effect on developing a peanut allergy. With two proportions, we use

$$H_0 : p_1 = p_2$$

Note: hypotheses are always about parameters. Never use a “hat” on the p 's in the hypothesis. As before, the direction of the alternative depends on what the research is intended to show: no difference (could go either way, so use \neq), less than, or greater than. You must specify which proportion is being subtracted from the other, because it will change the direction of the alternative.

The term “permutation” just means that we are mixing up, or permuting, the group assignments. In physical terms, it's shuffling the cards and redealing them into two groups. Because this is a randomized experiment, it's also fine to call this a “randomization” test. We are looking at what might have happened if treatments were equally effective, and we reassigned individuals to (possibly different) groups.

10. The null hypothesis is: $H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$ or $p_{treat} = p_{control}$. Is the researcher's question looking for an increase, decrease, or change in either direction? Fill in the blank with $<$, $>$, or \neq for the alternative hypothesis:

$$H_A : p_{treat} \text{ _____ } p_{control}$$

Go to the web page: <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/> and select “Enter Data” under 2 Categ. Enter the numbers and labels from the table in 4. The proportions should agree with those above, but let’s check:

- (a) The proportion of infants in Peanut group who became allergic:
 - (b) The proportion of infants in Control group who became allergic:
 - (c) The difference in proportions between the two groups:
-
11. Click “Test” under “2 categ.”, and generate 1000 shuffles and sketch the plot below.
-
-
-
-
-
-
-
-
-
-
12. Where is the plot of the results centered (at which value)? Explain why this makes sense.
-
-
-
-
-
-
-
-
-
-
13. Report the approximate p-value (i.e., strength of evidence) based on the observed result. (Reminder: we did this in the helper – hinderer study on Day 6.)

Go back to “Enter Data” and change labels slightly to clear the plot and then generate another 5000 random shuffles. How much does the strength of evidence change?

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

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

16. Another study on the effects of a different therapy had a p-value of 0.25. How would you report those results?

17. A third study computed p-value to be 0.73. How would you report those results?

18. Write up the pertinent results from the analysis on your own paper. When reporting the results of a permutation test, pertinent information from the analysis that needs to be included is:
 - The type of test used in the analysis (including the number of trials [shuffles]);
 - The null model assumed in the test;
 - The observed result based on the data;
 - The p-value and strength of evidence of the test and your conclusion; and
 - The appropriate scope of inference based on the p-value and the study design. Include:
 - How were the subjects selected? If they are a random sample from some population, then our inference goes back to the population.
 - Were treatments assigned? If treatments were assigned at random, then we can state a causal conclusion.

Take Home Messages

- We are conducting a permutation test which simply mixes up the labels. Because of random treatment assignment, this is also a randomization test.
- We tested to see if two proportions were equal. This is much like what we did in Unit 1 with a single proportion, except that the null hypothesis states that the two population proportions are equal (instead of one proportion coming from “blind guessing”).
- Question 18 asks you to write up results. Communicating and reading statistical results is a very important skill. We will keep doing this though the rest of the semester. We hope you can dive right into the task, but if you have any questions, please ask. You need to get this skills down – the sooner the better.
- Any questions?

What's Wrong With Your Weight?

A study³ in the *American Journal of Preventative Medicine*, 2003 looked at the self perception people have of their own weight. The participants in the *National Health and Nutrition Examination Survey* (NHNES) of the Center for Disease Control were asked if they thought themselves underweight, overweight, or about the right weight. Body Mass Index was also computed based on physical measurements, and the self-rating was compared to their actual classification. The NHNES is a survey of about 5000 people each year which is representative of the entire US population. The authors looked at data from 1999 when people were asked for their own perception of their weight. Interestingly, about the same proportion of men were wrong as women, but the way in which they were wrong might have a different pattern. This table shows a random subsample taken from only the **people who were wrong** in their self-perception.

Self Perception	Gender	
	Female	Male
Over Estimated	50	10
Under Estimated	20	59
Total	70	69

The parameter of interest is $p_1 - p_2$, the true difference in proportions who over-estimate their weight between women and men. We want to estimate how large the difference is, but first, as a review (as in Peanut Allergies), we'll do a test to see if the two proportions are equal.

1. State the null and alternative hypotheses in proper notation and in words.

H_0

H_a

2. Go to the web apps page: <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps> and select **Two Categ**, "Enter Data". Type our numbers into their cells so that "Success" is an overestimate, "Group 1" is Female and change the other labels accordingly.
 - (a) What proportion of women who are wrong about their weight overestimate (rather than underestimate) in this sample? What proportion of men? Take the difference between the two.

³Chang, V. W., & Christakis, N. A. (2003). Self-perception of weight appropriateness in the United States. *American Journal of Preventive Medicine*, 24(4), 332-339.

- (b) Switch to the “Test” page, and generate 1000 shuffles. Describe what the computer does for a single “shuffle”.
- (c) Why is the plot centered where it is centered?
- (d) Write down the p-value and strength of evidence.
- (e) Which is the more plausible explanation:
- These results occurred just by chance? (We just got unlucky and saw something weird) or
 - Men and Women who don’t know what their weight is really do differ in their self-perception of being over versus under weight.
- (f) A shorthand report of statistical results often does not report p-value and strength of evidence, but instead might say “the results were statistically significant at the 0.01 level”. This means that the researchers determined before collecting data that they wanted very strong evidence for any effect, and that they would decide to “reject H_0 ” if the p-value was less than 0.01, and that the p-value they computed did turn out to be less than 0.01. Here’s a summary of the steps:
- Decide what significance level (α) to use, typically 0.10, 0.05 (most common) or 0.01, before collecting data.
 - Collect and analyze data to get a p-value.
 - Reject H_0 if the p-value $< \alpha$. Note: we never “accept” H_0 . If p-value $> \alpha$, we say that we “Fail to reject H_0 ”. The testing procedure can provide evidence that H_0 is false, but it is not intended to show that H_0 is true. It’s like in our justice system when a defendant is found to be “not guilty” rather than “innocent”. The

jury is saying that the evidence is not strong enough to “demonstrate beyond a reasonable doubt” that the defendant committed the crime.

- If H_0 is rejected, some might report “the results were statistically significant at the α significance level.” (They should specify α , but if they don’t, we’d guess they used 0.05, the most common level.) This is not as informative as reporting the p-value, but you will read this in research articles, and need to know what it means. Reports for STAT 216 must contain the actual p-value.

Suppose we had set $\alpha = .01$ prior to data collection. What is your “decision” about $H_0 : p_1 = p_2$? (Either reject or fail to reject)

3. Now we will estimate the difference. You might want to look back at Activity 6 to review the reasons we like interval estimates instead of point estimates.

To build confidence intervals for a difference in proportions, we use the same app and data, just pick “Estimate” under “Two Categ.”

- (a) What is the point estimate for the difference in population proportions?
- (b) The output shows results from 1 resample of the data. They have generated 70 “Female” responses using $\hat{p}_1 = 50/70$ and 69 “Male” responses using $\hat{p}_2 = 10/69$. What is the difference in proportions (over) for this one sample?
- (c) Unlike the **test** we did above, there is no assumption that proportions are the same for men and women. Instead we use a bootstrap resample of the data on women and another bootstrap resample of the data on men to get an estimate of the sampling distribution.
Generate more resamples until you are sure that you understand how they are created. For the 70 women in one resample, what is the probability that each will “OverEstimate” her weight?

How would you do the resample with 70 index cards? Explain what to write on each card and how to randomly extract a card for each woman.

- (d) When you’re sure you know what it is doing, click 1000 several times. Note that it adds more rather than getting rid of the old ones. Record center and SE of this

distribution (upper right corner of plot).

(e) Obtain four confidence intervals, one for each confidence level and write them here.

(f) Also compute the approximate 95% CI using the $\pm 2SE$ method.

(g) Compare the two 95% intervals you created.

(h) Based on the 95% CI, is it plausible that $p_1 = p_2$? Explain your answer by referring to your CI.

(i) Interpret the meaning of this confidence interval in the context of this problem. What do we mean when we say “confidence”?

4. Summarize the hypothesis test results in a report as in question 18 of the Activity 10. Include all five bulleted points.

Note: This study did not randomly assign gender to people, it just observed whether they were male or female. The proper name for the test in 2 then is “permutation”, not “randomization”, and it was not an experiment. You may assume that the samples of men and women are representative of populations of people who are wrong about their weight status relative to ideal weight.

Take Home Messages

- With observational studies we can still conduct a permutation test, but it is not a randomization test.
- As in Activity 10, we tested to see if two proportions were equal. We had very strong evidence of a difference in proportions, but because we don't randomly assign gender, we can only say that we observed an **association** between gender and over/under estimation, not that gender causes this to happen.
- The new part of this assignment is the confidence interval for the difference in proportions.
- Just reporting "statistical significance" has many problems.
 - Statistical significance is not the same as **importance** of the results. Results could be very important, yet the p-value might not be very small, also small p-values could be attached to unimportant results.
 - Journals must filter articles so that they only publish high quality research. For years, many journals thought that small p-values were an indicator of quality, but using p-values as a filter cuts out some important results.
 - By increasing sample size, we shrink the spread of the sampling distribution which tends to make p-values smaller. It's important to have large enough sample size to get a good view of the true parameter, but a really large sample lets us "split hairs" and label unimportant results as "statistically significant".
 - The choice of a significance level must depend on the amount of evidence required before we're willing to give up H_0 . It is not the same for all situations. The values commonly used are just based on tradition and are not especially useful. A study with p-value 0.051 is not really different from a study of the same question which gives a p-value of 0.049.
- We have just used confidence intervals to estimate the difference between two proportions. Recall from Activity 6: our confidence is in the process used to create the interval. When used over and over on many random samples, 95% of the intervals created will cover the true parameter of interest (here $p_1 - p_2$) and 5% will miss the true parameter.
- When testing, we assume H_0 is true and the distribution is centered at 0. When computing a bootstrap confidence interval, we are centered at the statistic, or point estimate, $\hat{p}_1 - \hat{p}_2$.
- Use the remaining space for any questions or your own summary of the lesson.

Energy Drinks

From Red Bull to Monster to – you name it – in the last few years we’ve seen a large increase in the availability of so called “Energy Drinks”.

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

1. Why are energy drinks popular?
2. What claims are made in the advertising of energy drinks?
3. How do energy drinks interact with alcohol?
4. An experiment tried to compare the effects of energy drinks with and without alcohol on human subjects. Pharmacology is the study of how drugs affect the body, and “psychopharmacology” studies effects of drugs on the nervous system. An article in *Human Psychopharmacology* in 2009 reported on an experiment intended to tease out some of the effects and to compare an energy drink without alcohol to one with alcohol and to a non-energy drink. The research question is:

Does neuropsychological performance (as measured on the RBANS test) change after drinking an energy drink? After drinking an energy drink with alcohol?

Higher RBANS scores indicate better memory skills.

Go to the site:

<http://shiny.math.montana.edu/jimrc/IntroStatShinyApps>. Select Enter / Describe Data under One of Each. Select PreLoaded Data as the data entry method, and select REDAvsCntrl and Use These Data to load today's data.

treatment	RBAN
REDA	6.84
REDA	-9.83
REDA	-0.02
REDA	-9.12
REDA	-10.07
REDA	-19.34
REDA	3.97
REDA	-16.37
REDA	-21.02
Control	6.33
Control	1.65
Control	-3.58
Control	3.3
Control	-6.6
Control	3.29
Control	1.8
Control	1.8
Control	2.98

Examine the boxplots and dotplots. Describe any differences in the response (Change in RBANS) you see between Red+A and Control groups.

Center

Spread (SD and IQR)

Shape

The researchers used a computer randomization to assign the subjects into the groups. We'll shuffle cards instead.

5. Take 18 index cards and write the numbers 1 through 18 in a top corner, and the score of each individual in the middle starting with 6.84 for card 1, on to -21.02 for card 9, then continue with the second row. Line them up in the two rows like this data table:

RED+A	6.84	-9.83	-0.02	-9.12	-10.07	-19.34	3.97	-16.37	-21.02
Control	6.33	1.65	-3.58	3.30	-6.60	3.29	1.80	1.80	2.98

Compute the two means and take their difference: $\bar{x}_1 - \bar{x}_2$.

6. We want to test the hypothesis that the means are equal:
 $H_0 : \mu_1 = \mu_2$ (no difference in mean 'change in RBANS score' between REDA and control groups.) versus:
 $H_a : \mu_1 \neq \mu_2$

Consider this important question:

If the treatments have no effect on RBANS scores, then where do the observed differences in distributions and in means come from?

6. Discuss this within your group and write down your answer. Don't say that it has anything to do with the drink they were given because we are assuming the drinks are all having the same effect. (Give this about 2 minutes of serious discussion, then go on. If you get stuck here, we won't have time to finish the activity.)

7. Turn the index cards over and slide them around or shuffle by some other method, until you think they are thoroughly mixed up. Line up the shuffled cards (turned face up again) in 2 rows of 9 and compute the mean of each row.

Does the difference in the new means agree well with the original difference? If not, how much has it changed?

8. Suppose the first persons' change in RBANs was going to be 6.824 no matter which drink she was given, that the second would always be -9.83, and so on to the last person's score of 2.98. If we re-shuffle the people and deal them into two groups of 9 again and label them RED+A and Control, why do the means change? (You are describing a model of how the data are generated)

9. Go back to the applet and select under One of Each.
- (a) Do the means and SD's in the summary table match what we had earlier? Did they subtract in the same order as we did?
 - (b) What are the means for control and RED+A in the reshuffled version? The difference?
 - (c) Explain how our shuffling the cards is like what the computer does to the data.
 - (d) Click three times. Where is the plot centered? Why is it centered there?
 - (e) Below the plot, keep and enter the **observed difference in means from the original data** in the last box. Click . What proportion of the samples are this extreme?
10. There are other reasons that one person might show more change in RBANS than another person. Write down at least one. (Again, don't get stuck here.)

11. Lurking variables were discussed on Activity 10. When we randomly assign treatments, how should the groups compare on any lurking variable?
12. Are you willing to conclude that the differences we see between the two groups are caused by REDA? Explain your reasoning.
13. Write your interpretation of this interval.
14. Build a confidence interval using “estimate $\pm t^*SE$ ” where the estimate is the observed difference in means, $t^* = 2.12$, and using the st.dev. from the plot as our SE. Compare to the percentile interval above.
15. Write up the results of the hypothesis test as a report using the five elements from Activity 12. Be sure to refer to the response variable as “change in RBANS”, not just RBANS score.

Take Home Messages:

- If there is no treatment effect, then differences in distribution are just due to the random assignment of treatments. This corresponds to a “null hypothesis” of no difference between treatment groups.
- By randomly applying treatments, we are creating groups that should be very similar because differences between groups (age, reaction to alcohol, memory) are evened out by the random group allocation. If we see a difference between groups, then we doubt the null hypothesis that treatments don’t matter. Any difference between groups is caused by the treatment applied. Random assignment is a very powerful tool. When reading a study, it’s one of the key points to look for.
- Use the remaining space for any questions or your own summary of the lesson.

Reference

Curry K, Stasio MJ. (2009). The effects of energy drinks alone and with alcohol on neuropsychological functioning. *Human Psychopharmacology*. **24**(6):473-81. doi: 10.1002/hup.1045.

Birth Weights

Lab tests with animals have shown that exposure to tobacco smoke is harmful in many ways. To make connections to humans has been more of a challenge. One dataset which might help us connect tobacco use of pregnant women to birth weights of their babies comes from a large set of data on **all** births in North Carolina. We will examine a random sample of size 200 from the much larger dataset. The two variables provided are **habit** (either smoker or nonsmoker) and **weight** (baby's weight at birth measured in pounds).

Discuss

1. Could there be some physiological reason why birthweights for the children of the 28 smokers might differ from the birth weights of the babies born to nonsmokers? Write down what you and your group know about smoke and nicotine to hypothesize a connection to birthweight.

2. If the connection you are thinking about is real, would it tend to increase or decrease birth weights of babies born to smokers? Or could the effect go either way?
 - (a) What is the response variable in this study? Is it quantitative or categorical?

 - (b) Is there an explanatory variable in this study? If so, name it and tell which type of variable it is.

 - (c) Select the “Pre-loaded” data (birthweights) from the One of Each menu at <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps> Compute means for weight by habit and compute the difference in means.

 - (d) Is the difference between the means large enough to convince you that babies born to smoking mothers are lighter than those born to nonsmokers? Why or why not?

Studies that Use Random Sampling

The big differences between this study and the previous studies where you compared two conditions is the subjects in this study were a **random sample** from a larger population. The use of random sampling versus the use of random assignment changes the type of inferences that can be made.

A random sample is one in which the method used to choose the sample from the population of interest is based on chance.

Reminder: When studies employ random **assignment**, we are able to draw cause-and-effect conclusions about the **treatment effects**. Randomization evens out the influences of all possible lurking variables, and allows us to conclude that treatments really made a difference.

With data on birth weights, can we assign a baby to have a smoking versus nonsmoking mother?

The habit variable splits these subjects into two populations, and we have a sample from each. In studies with random **sampling**, the goal is to describe the sample data, to compare groups, and infer any differences back to the broader population(s) from which the sample(s) was/were drawn. Even though we might have reasons from animal studies to think smoking causes certain changes, we do not expect these data to provide **causal** evidence of such a connection. We might want to know how large the difference in means (the true population means) is between two groups. That's really an estimation question.

In the peanut allergy study we found strong evidence of a difference due to the therapy, but the inference only applied to the people in the study because we could not say that the sample of subjects was representative of a larger (all infants?).

An ideal study would start with a random sample from the population of interest so that we can make inference back to the population, and it would use random treatment allocation to allow causal inference. In practice, we must often settle for a convenience sample, so our inference only extends back to a subset of the population.

MODELING THE BIRTH WEIGHTS

You will conduct a **permutation** test to find out how likely it would be to see this large a difference in sample means if the two populations really have the same overall mean birth weight. The software does not distinguish between random assignment and random sampling, so it calls it a randomization test. We prefer “permutation” to emphasize that we are not assigning treatments, but we are mixing up the labels on the responses and allowing them to come from different groups. By doing the relabeling many times, we can see what results are expected when the populations really have the same distribution of responses.

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

Copy the means of each group and the difference in means from # 2 here.

7. Check results for the first resample. What is the mean birth weight to smokers from this simulated trial?

What is the mean birth weight to nonsmokers for this single simulated trial?

What is the difference in means between these two groups?

Evaluate the Results

8. Plot the differences in means from 1000 or more simulated trials. Sketch the plot below.
9. What are the cases in the plot?
10. Where is the plot of the results centered (at which value)? Explain why this makes sense.
11. We're not told exactly what the researchers were thinking ahead of time, but let's assume that the alternative hypothesis is that smoking moms tend to have lighter babies. What

is the alternative hypothesis of interest? Do you need to count Greater than or Less than or more extreme results to find the p-value?

12. Put the observed difference in the little box under the plot, chose the proper direction for comparison, and 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 births in North Carolina? to all births in the US? Why or why not?
16. Can the researchers say that the difference in the average birth weight is caused by the mother's smoking habit? Explain. If not, provide an alternative explanation for the differences.
17. Write-up the results of the simulation study. When reporting the results of a simulation study, pertinent details from the analysis that need to be included are: (as in Activity 12)
 - 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, whether it is one or two sided; and
 - The *scope of inference* based on the p-value and study design.

18. Finally, use the web app to create a 99% bootstrap percentile interval estimate of the difference in true mean babies weights from non-smoking to smoking mothers.
- (a) Obtain a 99% bootstrap percentile interval by generating 5000 samples and write it here.
- (b) Write your interpretation of this interval.
- (c) To write up a report on a confidence interval, we must include:
- The observed result based on the data;
 - The confidence level and the interval;
 - The method used to create the interval estimate (for bootstrap include the number of resamples);
 - The interpretation in the context of the data collected.

Take Home Messages:

- In an **observational study**, no treatment is applied, different groups are just observed.
- If we have a random sample from a population, we can infer results back to the population from which the subjects were drawn.
- When we do not randomly assign treatments, we cannot be sure that there are no lurking variables. Therefore, other explanations for the observed results are possible, and we cannot infer a **causal** relationship between our explanatory and response variables. It is just an association.
You may have heard the term “correlation does not imply causation” used in cases like this. It’s not quite accurate because (wait for Unit 3) correlation is a term for linear association between two quantitative variables. Here we have a categorical explanatory variable and quantitative response.
- It is possible that a causal effect exists, for example, we know that nicotine is a vaso-constrictor and poorer blood flow could reduce babies weights. This is a stats class, so you are learning how far statistics can go with inference. It is always possible to go further based on other natural laws or explanations, or laboratory results on simpler systems. It is important to use other information, and we want you to distinguish stat inference from other ways of learning about the world.
- Use the remaining space for any questions or your own summary of the lesson.

Arsenic in Toenails

Arsenic poisoning can be caused by drinking water containing a high concentration of arsenic. Symptoms from low-level poisoning include headaches, confusion, severe diarrhea and drowsiness. When the poisoning becomes acute, symptoms include vomiting, blood in the urine, hair loss, convulsions, and even death. A 2007 study by Peter Ravenscroft found that over 137 million people in more than 70 countries are probably affected by arsenic poisoning from drinking water.⁴

Scientists can assay toe nail clippings to measure a person's arsenic level in parts per million (ppm). They did this assay on 19 randomly selected individuals who drink from private wells in New Hampshire (data in the table below). First, they would like to know just what is the mean arsenic concentration for New Hampshire residents drinking from private wells.

An arsenic level greater than 0.150 ppm is considered hazardous. A second question is, “Is there evidence that people drinking the ground water in New Hampshire are suffering from arsenic poisoning?”

0.119	0.118	0.099	0.118	0.275	0.358	0.080	0.158	0.310	0.105
0.073	0.832	0.517	0.851	0.269	0.433	0.141	0.135	0.175	

Step 1. State the research question.

1. Based on the description of the study, state the two research questions to be answered.
2. Which research question should be answered using a hypothesis test and which should be answered using a confidence interval?

Step 2. Design a study and collect data.

3. What is the variable in the study? Is this variable quantitative or categorical?

⁴Ravenscroft, P. (2007). The global dimensions of arsenic pollution of groundwater. *Tropical Agriculture Association*, 3.

4. Define the parameter of interest in the context of the study. What notation should be used to denote it?

Step 3. Explore the data.

With quantitative data, we typically report and study the average value, or the mean.

5. What is the sample size in this study? $n =$
6. Calculate the observed statistic and use correct notation to denote it (check your answer with another group!).
7. Could your answer to 6 have happened if the arsenic concentrations in New Hampshire residents are not hazardous?
8. Do you think it is likely to have observed a mean like the one you got in 6 if the arsenic concentrations in New Hampshire residents are not hazardous?

Step 4. Draw inferences beyond the data.

We'll start with the first research question asked because we have done confidence intervals for a single mean back in Activity 10.

The First Research Question: How high is the mean arsenic level for New Hampshire residents with a private well?

9. Explain why this question is better answered using a confidence interval than by conducting a hypothesis test.
10. Explain how you can use a deck of 19 cards to create the bootstrap distribution. (Go back to Activity 11 if you don't remember.)
11. Use the One Quant option in the web applet to use the pre-loaded data (arsenic) and then generate a bootstrap distribution with 5000 or more bootstrap statistics. Draw the plot below and record the summary statistics.

Explain how one dot on the plot was created and what it represents in the context of the problem.

12. Create a 95% confidence interval using margin of error $ME = 2.11 \times SE$.
13. Create a 95% confidence interval using the bootstrap Percentile Method.
14. How similar are the confidence intervals in 13 and 12?
15. Would you expect a 90% confidence interval to be wider or narrower? Explain, then give a 90% (percentile) confidence interval.
16. Interpret the 90% confidence interval from 15.

The Second Research Question: Is the mean arsenic level for New Hampshire residents with a private well above the 0.15 threshold?

There are two possibilities for why the sample average was 0.272. List them here and label them as the null and alternative hypotheses also write the null and alternative in notation.

H_0 :

H_a :

Is the alternative hypothesis right-tail, left-tail, or two-tail?

We can simulate the behavior of arsenic concentrations in New Hampshire ground water if we

assume the null hypothesis which gives a specific value for the mean. The two key ideas when creating the reference distribution are:

- The resamples must be consistent with the null hypothesis.
- The resamples must be based on the sample data.

We can use cards like we did for the CI above, but we have to change the values so that they are consistent with the null, $\mu = 0.15$.

18. How you could modify the sample data so as to force the null hypothesis to be true without changing the spread? (Do not spend more than 2 minutes on this question.)

19. Next we will simulate one repetition of the 19 toenail values collected by creating a deck of 19 cards to simulate what the data would look like **if the null hypothesis** were true.

(a) What is the null value in this study?

(b) How far is the sample mean from this null value?

(c) We need to shift the original data so that is it centered on the null value. Subtract the value from (b) from each of the data numbers to get:

-0.003	-0.004	-0.023	-0.004	0.153	0.236	-0.042	0.036	0.188	-0.017
-0.049	0.710	0.395	0.729	0.147	0.311	0.019	0.013	0.053	

What is the mean of the above values? Why do we want this to be the mean?

20. To speed up the process, we use **Test** option under One Quant at <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/>.

- Above the main plot, change the value for the null hypothesis to the one in our null. (the just barely safe level) The software will shift the data to hve this mean.

Look at the box on the right labeled **Original Sample**. Does the mean match your answer to 6? If not, consult with your instructor!

21. What is the statistic from the first resample?
22. Where in the output did you find it?
23. Explain in the context of the problem what the one dot on the main plot represents.
24. Generate 5000 randomization samples. Copy the summary statistics and the plot of the randomization distribution below
25. Where is the distribution centered? Why does that make sense?

Remember why we conducted this simulation: to assess whether the observed result (mean of 0.272) would be unlikely to occur by chance alone if the ground water in New Hampshire is not hazardous.

26. Locate the observed result on the randomization distribution. Does it appear to be likely or unlikely to occur under the null hypothesis? Explain your reasoning.
27. Just how unlikely is the observed result? Calculate your p-value using the web app and the appropriate direction and cutoff value.

How many randomization samples had a mean at least as extreme as the observed result?

28. Is the approximate p-value from your simulation analysis (your answer to 27) small enough to provide much evidence against the null hypothesis? If so, how strong is this evidence? Look back to the guidelines for assessing strength of evidence using p-values given in the ESP activity.

Step 5: Formulate conclusions.

29. Based on this analysis, what is your conclusion about the residents in New Hampshire who own a private well based on this study?
30. Can you extend your results to all of New Hampshire residents? All New Hampshire residents with a private well? Explain your reasoning.

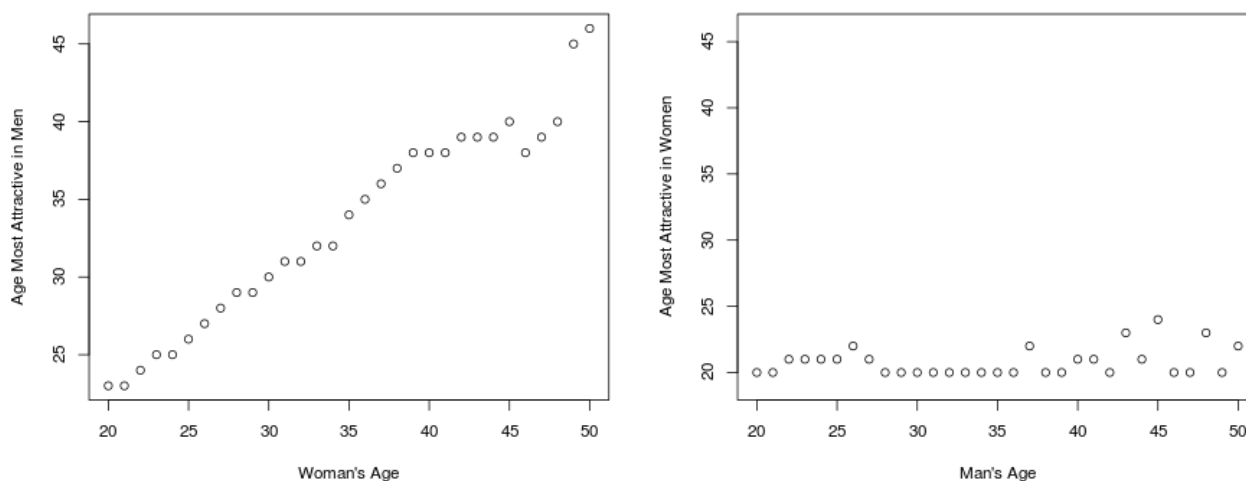
Take Home Messages

- We first reviewed building a CI for a single mean.
- You need to know when to discuss means versus proportions. If the response has two categories, then we deal with proportions. If the response is quantitative, then we estimate and test means.
- The new twist today was to do a simulation for testing $H_0 : \mu = \mu_o$ that the mean is some particular value. We had to modify the data to make H_0 true, shifting it from its center at \bar{x} to be centered at μ_o . Then we resampled it as if for a bootstrap confidence interval, and located the observed statistic (\bar{x}) to see how far out in the tails it was (the p-value).
- Use the remaining space for any questions and your own summary of the lesson.

Who Looks Good to You?

Christian Rudder works for the dating web site OKcupid, and has written a book, *Dataclysm* about some surprising data collected from the web site.

As an example, here are plots he gives for women and for men. The horizontal axis is the age of the man or woman being interviewed. The vertical axis is the age which they think looks most attractive in the opposite sex.



There are clearly big differences between men and women, so we want to describe them with statistics.

1. Suppose you're talking to a friend over the telephone, and you need to explain that the same two variables have a different relationship for women than for men. How would you describe the two plots?
2. What statistical summaries differ in the two plots?
3. As a review, in Algebra class you would have learned an equation for a linear relationship between x and y . What letters did you use for slope and intercept? What does "slope" mean? What does "intercept" mean?

In Statistics, we use the following equation for a “true” regression line:

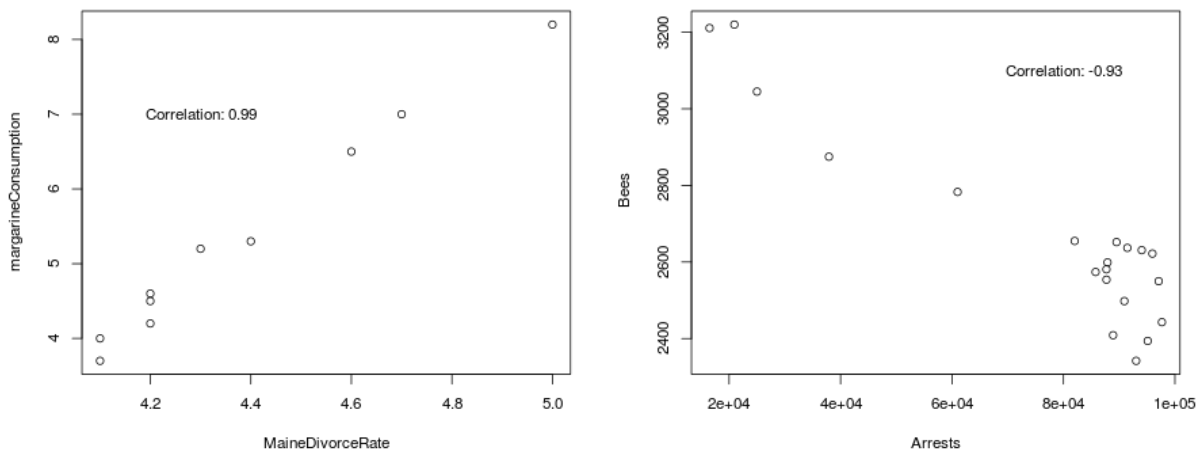
$$y = \beta_0 + \beta_1 x + \epsilon$$

and when we estimate the line we add hats to the parameters, β_0 and β_1 , and also to the left side to say we have an estimated response, \hat{y} .

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

4. Take a guess at the slopes for the two plots above.
5. **Correlation** measures the strength and direction of a **linear** relationship between two quantitative variables. It is a unit-less number between -1 and 1 with zero meaning “uncorrelated” and 1 or -1 meaning perfect correlation – points fall right on a line. The sample correlation is called r , and the true correlation is ρ , the Greek letter “rho”. The sign of the correlation tells us if the one variable increases as the other does (positive) or decreases (negative).

Go to <http://www.tylervigen.com/> and find a “spurious” correlation which has correlation coefficient, r less than -0.90 and one that has $r > 0.99$. Here are the two variables plotted without year (a lurking variable).

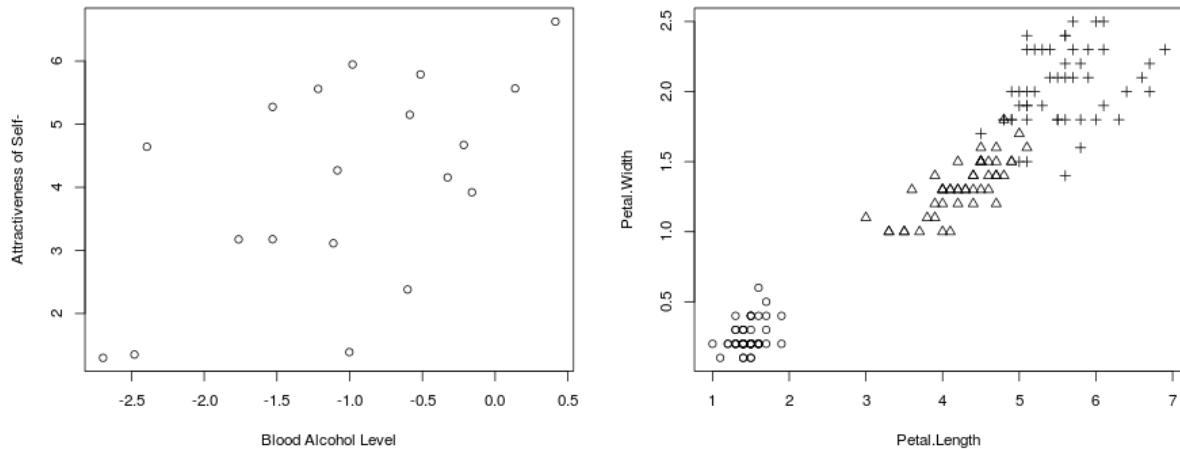


The point here is that if you search through lots of variables, you can find pairs that increase in the same way, or oppositely.

Just to show you found the site, what variables are in the first plot, and what is their correlation?

6. Why are the values on that page called “spurious”?

7. Correlations in the following plot are either 0.96 or 0.56. Which is which?



The first is data recreated from summary stats given for a study of how attractive men felt they were and their blood alcohol level (log scale, so negative numbers do make sense). The second shows measurements of iris petals. The clusters are for three different species. Within species correlations are quite different: 0.33, 0.79 and 0.32, but with all the data, correlation is higher.

8. Look back at the Age-Attraction plots from OKcupid. Guess what those correlations are for women and for men.
9. Correlation contest:
Go to <http://www.rossmanchance.com/applets/GuessCorrelation.html>. Click
Track Performance, then each member of your group guesses correlation for 5 s.
(Click between each person.) The first plot below Track Performance tells you the correlation between your guesses and the true values. What is it? What's the best one in your group?

Slope

10. In Algebra, a line is a collection of points satisfying an equation. In Statistics we start with data and have to find a good line to represent the relationship between two variables. When there is a lot of scatter in the points, many lines look reasonable. Go to <http://www.rossmanchance.com/applets/RegShuffle.htm> to see data on people's foot length versus height.

- (a) Is this a linear relationship?
- (b) Positive or Negative?
- (c) Strong, Moderate, or Weak?
- (d) Guess the correlation, then check with the button below the plot.

We'll use this app for the rest of this activity.

11. Click **Show Movable Line:** ☐ and move the center by dragging the large green square in the middle and adjust the slope by dragging either end of the line up or down. Get the line which you think best fits, write its equation here:

$$\widehat{\text{height}} = \text{---} + \text{---} \times \text{footlength}$$

12. Click **Show Regression Line:** ☐ and write the equation here:

$$\widehat{\text{height}} = \text{---} + \text{---} \times \text{footlength}$$

- (a) Was your slope too large? too small? about right?
- (b) What height does this line give for a person whose foot length is 0?

(This is an example of **extrapolation**: predicting y for an x value outside the range of observed x 's.)

13. Let's see how much we can change slope and correlation by adding just one more point. Give it a new "x" value of 60 cm. Pick a "y" value which you think will change the general pattern we see between length and height. Can you get the correlation to go close to zero? I'm not having luck with "move observations" but you can edit the last line of data to try new "y" values until you get a correlation of about zero.

- (a) What are the coordinates of the added point?
- (b) Now what is the slope of the regression line?
- (c) Is correlation resistant to outliers? Is slope? Explain.

14. Click Revert to go back to the original data. Have it show the regression line and the residuals. You can't see from the plot, but points below the line have negative residuals, points above the line have positive residuals according to this definition:

$$\text{residual} = \text{observed} - \text{predicted or } e = y - \hat{y}$$

- (a) Which residual is largest? Find the (x, y) pair in the data table associated with that point.
- (b) Compute it's predicted value using the equation given. Also compute the residual for the one furthest below the line.
- (c) Now click **Show Squared Residuals**. These are important because we are using the "Least Squares" line. It picks slope and intercept to minimize the sum of all the squared residuals. Write down SSE (sum of squared errors).

Any other line will have larger SSE.

Take Home Messages:

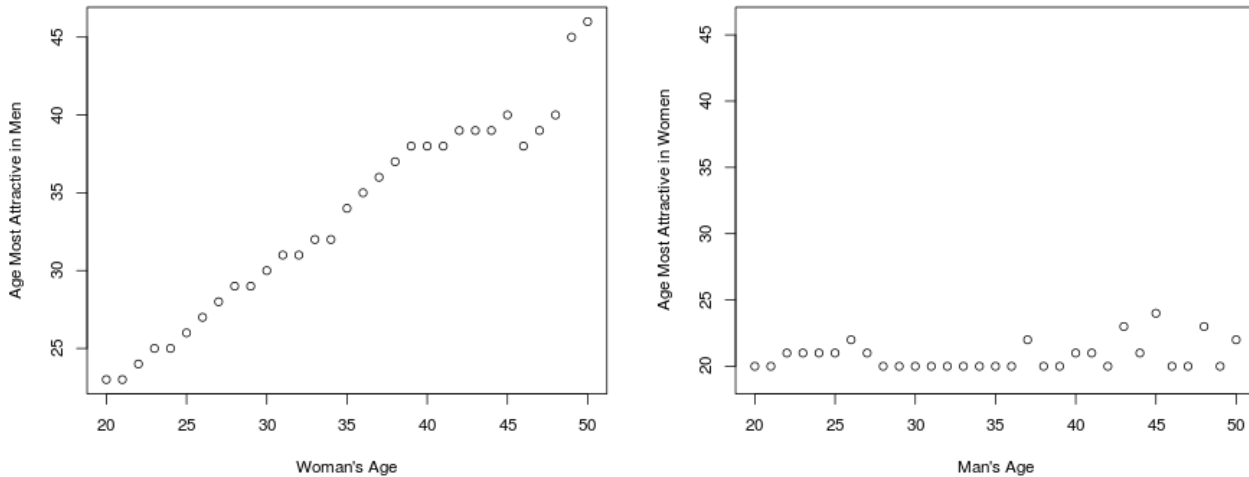
- It's not right to speak of the correlation between gender (or any categorical variable) and age, or between two categorical variables.
- It only works for linear relationships. We can have very strong nonlinear association with correlation near zero.
- Positive relationships mean large values in one variable are generally paired with large values in the other variable (and small with small). Negative relationships pair large with small.
- Correlation has no units and is restricted to the interval (-1,1). Both end of the interval indicate very strong correlation. Near zero, we say the two variables are uncorrelated.
- Neither correlation nor slope are resistant to outliers. A change in one point can completely change these values.
- Slope of the "Least Squares" line is given the label $\hat{\beta}_1$ because it estimates the true slope, β_1 . It is related to correlation.

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x}$$

where s_y is the Standard Deviation (SD) of the responses, and s_x is the SD of the explanatory variable.

Is Correlation Zero? Is Slope Zero?

Recall the plots we started with last time of “most attractive age”:



Least squares regression lines:

$$\text{Women: } \hat{y} = 9.02 + 0.70x$$

$$\text{Men: } \hat{y} = 19.57 + 0.0343x$$

1. What would you guess women aged 36.5 would say is the most attractive age of men?
2. What would you guess men aged 49.5 would say is the most attractive age of women?
3. Discuss this alternative with your group. Perhaps the age of the men really doesn't matter, and we'd be better off estimating their preference by using the mean “most attractive age for women” which is $\bar{y} = 20.78$ for all men, just ignoring the men's age. Does that seem like a reasonable way to describe the second plot: “men of all ages find 20.8 years to be the most attractive in women”? Write down your group's conclusion.

BTW: If any of you women over age 23 find this depressing, Rudder does say in his book that when men go to search for women on the dating site, they do adjust for their own age and ask to see profiles of older women if they are older themselves.

4. Go to the website <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/>, select Two Quant. and “Enter Data”. The OKcupid data is preloaded as either `WomenRateMen` or `MenRateWomen`. Use the men’s rating of women for now.

Consider this line: $\widehat{mostAttrWomen} = 20.77 + 0 \times mansAge$

- (a) Where have you seen the intercept estimate before?
 - (b) If you plug in any age for men, say 18 or 54, what result will you get from this line?
 - (c) What does that say about the relationship between x and y ?
 - (d) What will be the true slope for y based on x if there is no relationship? Use correct notation.
5. If we want to test the null hypothesis “no linear relationship between men’s age and the age of women they find most attractive”, what is the null value of the true slope? Use β_1 as the true slope and fill in the hypotheses. Use a “right-tailed” alternative.
- $H_o :$
- H_a
6. When you select “Test” for these data, a “Shuffled Data” plot appears in the middle of the page. For each x value, there is a line from the original (blue) y value to the new shuffled y value (green). Does this shuffle follow H_0 or H_a ? Explain.
7. Is the least squares line in the lower plot flatter or steeper than the one one in the upper plot? Is $\hat{\beta}_1$ closer or further from zero?
8. Compute the p-value. Explain which shuffles it counts.
9. State your decision and conclusion using $\alpha = 0.05$.

10. Switch from slope to correlation. What is the sample correlation, and what is the p-value for a test of $H_0 : \rho = 0$ versus $H_a : \rho > 0$?

11. Now test to see if slope is zero when we compare women's age (now this is x) to the age of men they find most attractive (our new y). Again use a "right-tailed" alternative.
 - (a) State the hypotheses.
 $H_o :$

 H_a

 - (b) Go back to "Enter Data" and load the women's data. What is the equation of the least squares line?

 - (c) Create 1 random shuffle of the data. Explain (yes, again – it's important) what is being shuffled.

 - (d) Compute the p-value and interpret it.

 - (e) State your decision and conclusion using $\alpha = 0.05$.

 - (f) Switch from slope to correlation. What is the sample correlation, and what is the p-value for a test of $H_0 : \rho = 0$ versus $H_a : \rho > 0$?

12. Are the men and women shown in these plots a random sample from a larger population? Are they representative of some larger population?

13. Was some treatment randomly assigned?

14. What is the scope of inference?

15. Write a report on the two hypothesis tests we just did.

Take Home Messages:

- A slope of zero is very “special” in that it says we would predict the same value for the response, \hat{y} for all values of x . That means that there is no linear relationship between the two variables.
- The OKCupid data gives us one example where slope is close to zero and another where slope is far from zero. Our conclusions should be quite different.
- The mechanics of computing p-value have not changed. We assumed H_0 was true, and created shuffled data consistent with H_0 . For each dataset, we computed a slope, and plotted a histogram for slopes under H_0 . P-value counted the number of times a slope was as or more extreme as the one we observed divided by the number of shuffles. The only difference is that we had the computer find slopes instead of proportions or means. You can easily click the correlation button to get a test of $H_0 : \rho = 0$. P-values will agree with the test for slope = 0.
- Use the remaining space for any questions or your own summary of the lesson.

On Being Wrong 5% of the Time

Our confidence in a 95% confidence interval comes from the fact that, in the long run, the technique works 95% of the time to capture the unknown parameter. This leads to an old cheap joke:

Statisticians are people who require themselves to be wrong 5% of the time.

We hope that's not really true, but decision making leads to a dilemma:
If you want to never be wrong, you have to always put off decisions and collect more data.

Statistics allows us to make decisions based on partial data while controlling our error rates.
Discuss these situations and decide which error would be worse:

1. A criminal jury will make an error if they let a guilty defendant go free, or if they convict an innocent defendant. Which is worse? Why?

2. The doctor gives patients a test designed to detect pancreatic cancer (which is usually quite serious). The test is wrong if: it says a healthy patient has cancer (a false positive), or if it says a patient with cancer is healthy (a false negative). Which is worse? Why?

3. A weather forecaster working at an airport in Indonesia on December 28, 2014 had to decide if it was too dangerous to allow Air Asia Flight 8501 to fly to Singapore. The flight was allowed, resulting in the deaths of all 162 people aboard. Errors don't get much worse than that, but what would the cost be of grounding a flight?

4. Large chain stores are always looking for locations into which they can expand – perhaps into Bozeman. When would a decision to open a store in Bozeman be wrong?
When would a decision to not open a store in Bozeman be wrong?
Which is the worse error?

Two Types of Error.

Definitions:

- To reject H_0 when it is true is called a Type I error.
- To fail to reject H_0 when it is false is called a Type II error.

To remember which is which: we start a hypothesis test by assuming H_0 is true, so Type I goes with H_0 being true.

This table also helps us stay organized:

H_0 is:	Decision:	
	Reject H_0	Do not reject H_0
true	<i>Type I Error</i>	Correct
false	Correct	<i>Type II error</i>

Which is worse?

The setup for hypothesis testing assumes that we really need to control the rate of Type I error. We can do this by setting our significance level, α . If, for example, $\alpha = 0.01$, then when we reject H_0 we are making an error less than 1% of the time. So α is the probability of making an error when H_0 is true.

There is also a symbol for the probability of a Type II error, β , but it changes depending on which alternative parameter value is correct.

Justice System and Errors

Refer to this reading about the justice system:

<http://www.intuitor.com/statistics/T1T2Errors.html>

In both the justice system and in statistics, we can make errors. In statistics the only way to avoid making errors is to not state any conclusion without measuring or polling the entire population. That's expensive and time consuming, so we instead try to control the chances of making an error.

For a scientist, committing a Type I error means we would report a big discovery when in fact, nothing is going on. (How embarrassing!) This is deemed more critical than a Type II error, which happens if the scientist does a research project and finds no "effect" when, in fact, there is one.

Type II error is harder to control because it depends on these things:

- The null hypothesis has to be wrong, but it could be wrong just by a small amount or by a large amount. For example if we did not reject the null hypothesis that treatment and control were equally effective, we could be making a type II error. If in fact, if there was a

small difference, it would be hard to detect, and if the treatment was far better, it would be easy to detect. This is called the effect size, which is [difference between null model mean and an alternative mean] divided by standard error.

- Sample size. P-values are strongly affected by sample size. With a big sample we can detect small differences. With small samples, only coarse or obvious ones.
- Significance level. The fence, usually called α (alpha), is usually set at .10, .05 or .01 with smaller values requiring stronger evidence before we reject the null hypothesis and thus lower probability of error.

Instead of limiting the probability of Type II error, researchers more often speak of keeping the power as large as possible. Power is one minus the probability of Type II error. Go to the Power Demo page: <http://shiny.math.montana.edu/jimrc/IntroStatShinyApps> and click “Power Demo” under “One Quant” data.

5. Set Sample size to 8, SD to 2, Alternative Mean to 2, and significance level to 0.01. What is the power?

Increase sample size until you get power just bigger than 0.80. How large a sample is needed?

6. Return to sample size 8. Adjust SD to get power just over 0.80. Do you make it larger or smaller? What value worked?

What is your effect size?

7. Return to SD = 2. Change Alternative Mean to get power just over 0.80. Did you make it larger or smaller? What value did you settle on?

What is your effect size?

8. How do the effect sizes in 6 and 7 compare?

How do SD and Alternative Mean work together to determine power?

9. Change significance level to 0.05. What happens to power?

Change it to 0.10. What is the power?

10. In which direction does power change when we decrease the significance level?

11. Suppose that we are planning to do a study of how energy drinks effect RBAN scores similar to the study we read about on Day 11. From previous data, we have an estimate of standard deviation of 3.8. We plan to use a significance level of $\alpha = .05$, and want to be able to detect an increase in mean RBAN score of 2 with 90% power. How large must our sample size be?

If we choose $\alpha = .01$, how large a sample is needed?

12. Now suppose that we are using the same visual discrimination task used to study sleep deprivation. Historical data provides an estimate of $SD = 13$. We want to use $\alpha = .05$ and need to detect an increase in mean score of 6 with 80% power. How large a sample is needed?

If we want to limit the chance of Type II error to 10% or less, how large a sample size is needed?

13. Suppose we do another study on energy drinks with alcohol using Control and REDA. This type we test hand-eye coordination using $H_0 : \mu_{control} = \mu_{REDA}$ versus alternative $H_a : \mu_{control} > \mu_{REDA}$.

(a) What would be a Type I error in this context?

(b) What would be a Type II error in this context?

Take Home Message

- Errors happen. Use of statistics does not prevent all errors, but it does limit them to a level we can tolerate. We have labels for two types of error.
- The talk about probability of error is based on the sampling distribution given random assignment of treatments or random sampling. It's really a "best case" scenario, because there could be other sources of error we have not considered. For example, we could have not sampled from some part of the population, or we could have errors in our measuring tools.
- If you are designing a study, you might need to consult a statistician to help determine how large a sample size is needed. You'll need to decide what α to use, what the underlying variation is (σ), and how large a difference you need to detect with a certain level of power.
- Use the remaining space for any questions or your own summary of the lesson.

Unit 2 Wrapup

Vocabulary

- Response, explanatory variable
- Random Assignment (why do we do it?)
- Random Sampling
- Lurking Variables
- Causal inference (or just association)
- Scope of Inference
- Permuting labels, permutation test, randomization test
- Bootstrap process: CI for μ
Percentile method
estimate $\pm t^* SE$
- What points must be included in a statistical report?
- Stat significance is not the same as importance or practical significance.
- Interpretation of Confidence Interval
- Correlation, Slope
- Type I Error probability is limited to α
- Type II Error power = $1 - \beta$
- What settings affect power of a study?

We have built confidence intervals and done hypothesis tests for one mean, difference in proportions, difference in two means. And we did hypothesis testing for a slope (or correlation) being 0. (Could also estimate slope with a CI, but didn't have time).

1. For all studies in Unit 2 consider whether the study was an experiment or observational study. What was the explanatory variable? the response?

Study	Experiment?	Explanatory Vble	Response Vble
Study Music			
Book Cost			
Peanut Allergies			
Nonideal Weight			
Energy Drinks			
Birth Weight			
Arsenic			
Attraction			

Extensions

2. Peanut Allergy Study

- (a) Suppose the results of the experiment had been that 4 had become allergic in the peanut group (instead of 5) and only 36 had become allergic in the control group (instead of 35). Explain how your approximate p-value would have been different in this case. Also describe how the strength of evidence for the benefit of peanut protein would have changed.

- (b) Suppose that all counts were divided by 5, so we had 1 allergy in the treatment group and 7 in the controls (out of 49 and 51 kids). Explain how your p-value would have been different in this case. Also describe how the strength of evidence for the benefit of peanut protein would have changed.

More Examples

The following exercises are adapted from the CATALST curriculum at <https://github.com/zief0002/Statistical-Thinking>.

3. Teen Hearing Study

Headlines in August of 2010 trumpeted the alarming news that nearly 1 in 5 U.S. teens suffers from some degree of hearing loss, a much larger percentage than in 1988.⁵ The findings were based on large-scale surveys done with randomly selected American teenagers from across the United States: 2928 teens in 1988-1994 and 1771 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.

- (a) Describe (in words) the research question. List the explanatory and the response variables in this study.

- (b) Just as with the peanut protein therapy and sleep deprivation studies, this study made use of randomness in collecting the data. But the use of randomness was quite different in this study. Discuss what type of conclusions can be made from each type of study and why you can make those conclusions for one study but not the other.

- (c) Are the percentages reported above (14.9% and 19.5%) population values or sample values? Explain.

- (d) Write out the null model for this analysis.

⁵ Shargorodsky et. al., 2010. *Journal of the American Medical Association*

4. Mammography Study

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. In addition, routine screening is recommended for women between the ages of 50 and 74, but controversy exists regarding the benefits of beginning mammography screening at age 40. The reason for this controversy stems from the large number of false positives. Data consistent with mammography screening yields the following table:⁶

Truth:	Mammogram Results:		Total
	Positive	Negative	
Cancer	70	90	160
No Cancer	700	9140	9840
Total	770	9230	10000

- What percent of women in this study have breast cancer?
- If the null hypothesis is that a woman is cancer free, what would an erroneous test result be? Is that a false positive or a false negative?
- Estimate that error rate using these data.
- If a woman really has cancer, what would an error in the test be saying? Is that a false positive or a false negative?
- Estimate that error rate using these data.

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.

- Given the consequence of a false test result, is the false negative or false positive a larger problem in this case? Explain.

⁶*Annals of Internal Medicine* November 2009;151:738-747

5. Blood Pressure Study

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 1 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	103	109	107	110	111	106	112	100
Counting	98	108	108	101	109	106	102	105

- (a) Do the data in this study come from a randomized experiment or an observational study? Explain.
- (b) Calculate the difference in the means.
- (c) Write out the null model for this study.
- (d) Use our web app to do the appropriate test to determine if a difference this large could reasonably occur just by chance. Comment on the strength of evidence against the null model.

6. Investigators at the UNC Dental School followed the growth of 11 girls from age 8 until age 14. Every two years they measured a particular distance in the mouth via xray ((in mm) . Assume that they want to test “Is the rate of growth zero?”. The data are called “Dental” under “Two Quant”. Note: ages are fixed by design, not randomly assigned.

(a) Find the estimated least squares line. Note: be sure that “age” is the explanatory variable in your plot. You may need to click Swap Variables (X goes to Y) get that ordering.

(b) How fast is this measurement changing?

(c) What hypotheses are we testing?

$H_o :$

H_a

(d) Compute the p-value for the hypothesis test.

(e) Give the scope of inference.