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 - one mean (24)	**19** t inference for $\mu_1 - \mu_2$ (25) **Nov 19: Last Day to Withdraw**
24 No Class	**26** **Thanksgiving Holiday**
December **1** Paired Means (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 anothers work without citing the source is a form of academic misconduct. Even inadvertent or unintentional misuse or appropriation of anothers 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 others 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 are 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. *Answers will vary. (AWV)*

2. Now flip your coin 45 times and write down the result of each flip. *Answers will vary.*

   AWV.

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.

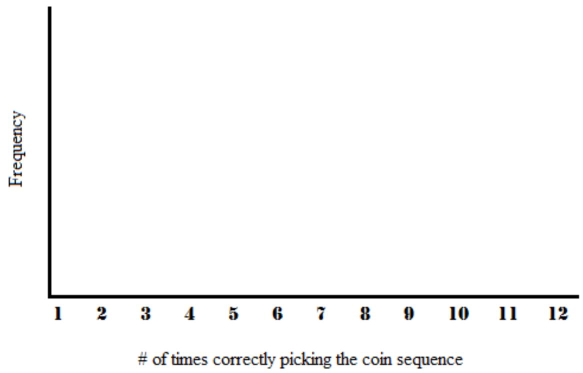
   Write both sequences on the board in shorthand with the number of consecutive heads or tails, so the sequence HHHTTHHT will be recorded as:
   3H
   2T
   2H
   T

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?



Answers will vary. AWV.

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? *Half the time.*
- 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. More than half.*

7. Which of these two explanations seems more plausible given the data? Explain your group's choice. *Knows something?*

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? *Answers will vary. Yes - instructor is 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. *Not just guessing?*

Further Exploration

10. What could your instructor do to make you more sure they can in fact detect fraud? *Do just as well with more such sequences.*

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. *20 of 24 is stronger evidence than 10 of 12.*

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 × 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

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

For the number of characters in the emails, we get

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

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 25^{th} percentile.

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

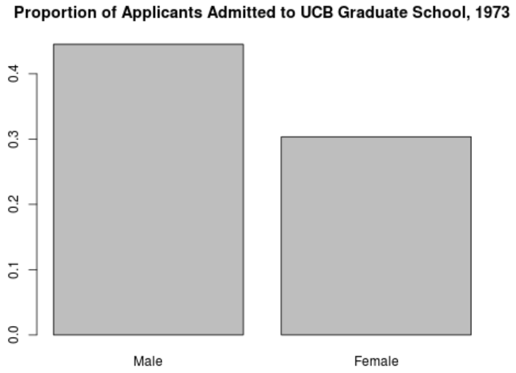
$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

which, for the data we are considering, is 13.125.
It is an important measure of **spread**.

Plotting Data

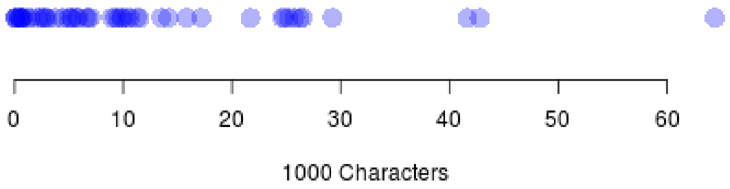
As with numeric summaries, the type of data determines the appropriate plot.

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.)

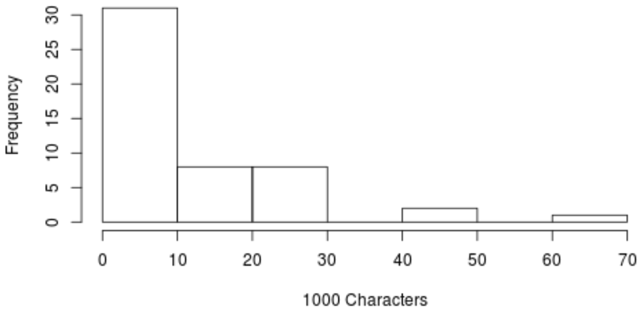


Quantitative variables are plotted with dot plots, histograms, density plots, and boxplots.

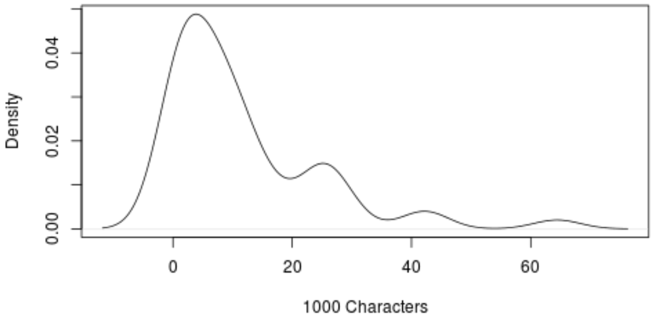
dot plots represent each point with a dot above the number line. This works well with small sample sizes. If the data are too close together to distinguish, we might stack them up to remove any overlap.



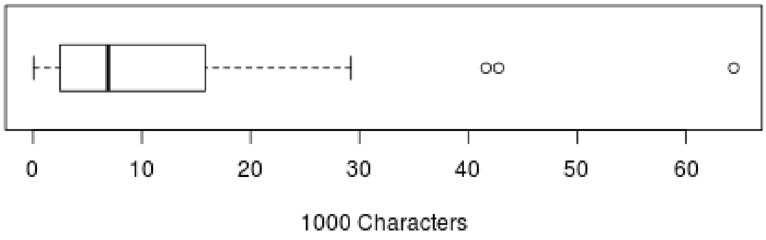
histograms divide the axis into "bins" and count the numbers of points falling into each bin. The height of each bin might show the count (frequency) of values in the bin or the proportion (relative frequency) for the bin. These plots work with moderate to large sized data sets. Choosing the best number 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 × IQR's of either Q_1 or Q_3 and flags points beyond 1.5 × 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?

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.

   B. It's numeric, not categorical.

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.

   D. Area code is categorical. Finding an average for the class makes sense for the others, but average area code is meaningless.

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?

   B, E, F, G

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

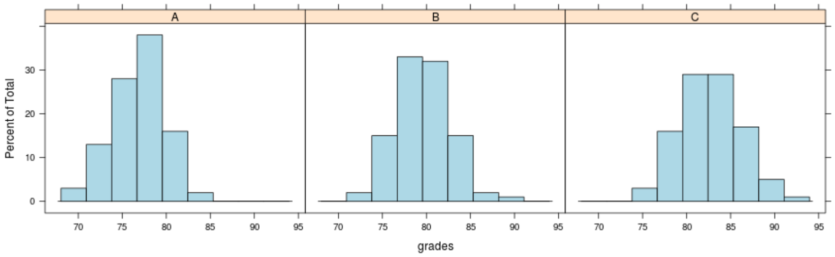
5. For which variables on the previous page, A through H would the **proportions** be informative?

   A, C, D, H

Comparing Distributions

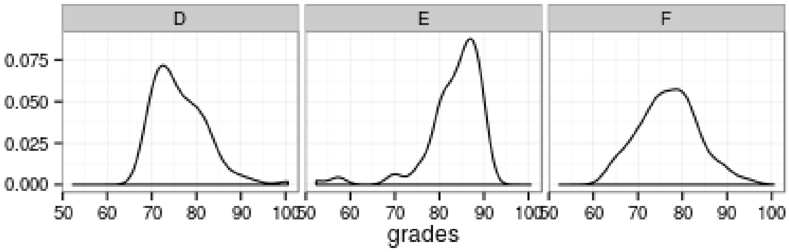
Now we'll focus on quantitative data.

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.



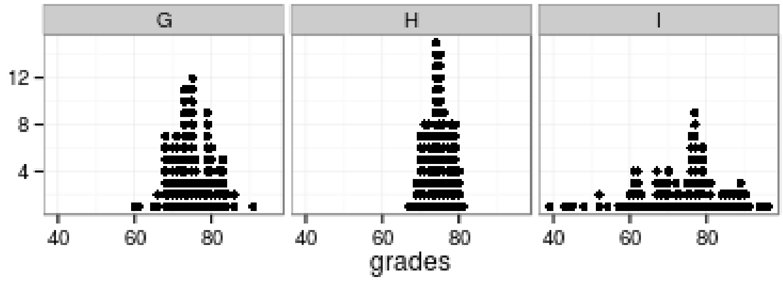
   Class C has the largest mean and median, so most will vote for it.

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.



Class H has more A's than C's, so it's the wise choice. I seems evenly split to high and low grades, while G seems to have lots of low grades.

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.



The big difference here is in spread. If you're an "average" student, then you would like E because almost everyone gets a C and there's little chance of flunking. If you are a good student, then F is more attractive since more people get A's in this class.

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.

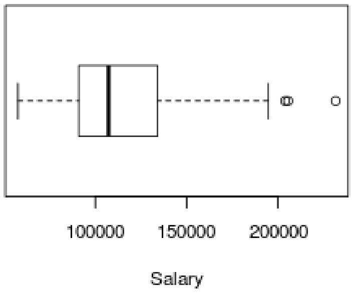
   (c) Comparing skew (could be left or right) to symmetric tells us which tail stretches out more. (Let's hope that there are more high grades than low ones.)

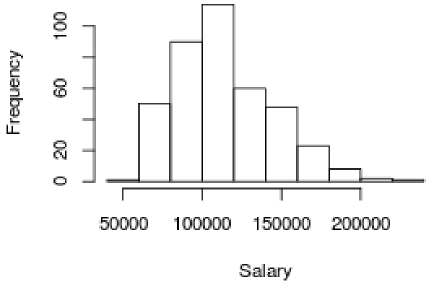
   In the three problems above, which comparison were you making? For each set of comparisons, fill in center, spread, or skew.
   (6) _____ *center* _____ (7) _____ *skewness* _____ (8) _____ *spread* _____

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

    Center is generally the easiest. One could argue that spread is hard because you have to read the scales carefully, plus it depends on your amount of ambition for a good grade. Skew is also hard because it require a close comparison of each tail. In this case, lots of A's are clearly preferred to an even spread or to more D's.

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.





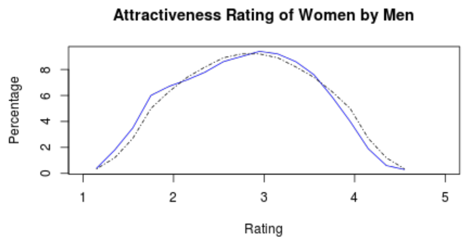
(a) Is salary skewed (if so which way?) or does it have a symmetric distribution?
right skewed

(b) Are any points flagged as outliers? If so, describe them.
Three points with salary over $200,000 are flagged as outliers.

(c) Give approximate values for the median and the first and third quartiles. Also compute the IQR.
Q1: $90k, Mdedian: $110k, Q3: $135K, IQR: $25k

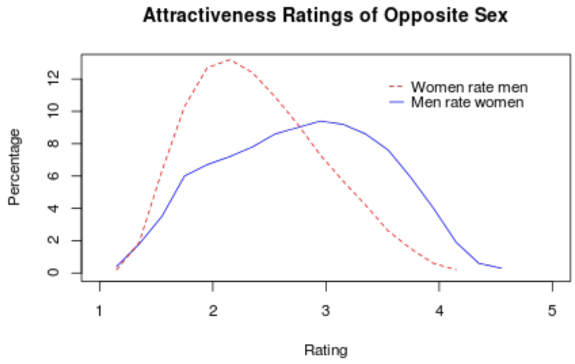
(d) For these data, which is a better summary: mean and standard deviation? or median and IQR? Why?
median and IQR – due to skewness.

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, (sometimes called a "hollow Histograms"). 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.



The solid line is slightly skewed to the right. If womens' looks are symmetrically distributed, then men are being a bit hard on them, pushing their scores a bit lower.

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.



Women have stricter standards in what they see as attractive? Or are women posting pictures that are showing themselves to better advantage than the pictures men choose?

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 shape 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 page 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:
$\overline{x}$ to represent the sample mean (or average value).
s to represent the population's standard deviation (spread).
$\widehat{p}$ to represent a sample proportion. (We often use a hat to represent a statistic.)
r for correlation between two quantitative variables in a sample.
$\widehat{\beta}_1$ 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 $\widehat{p}$ 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? *Answers will vary. Some will be more representative than others.*

3. Suppose we want to estimate the mean (average) length of all words in our population. Is that a parameter or a statistic? *parameter*

4. What is the average word length for your sample? *AWV*

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 Clear 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 Use Data. 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 1 and Sample Size to 10. Click Draw Samples. Write out the 10 word lengths in that sample. *AWV*

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. *AWV*

7. Click Generate 1 sample again and record the next mean. *AWV*

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.) *Should be close to 4.257 for mean, 0.60 for st.dev*

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. *Should see a fairly symmetric distribution about the center (mean close to 4.257). Mine goes 2.5 to 7*

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. *Hope to see some bias here. Discuss estimates close to 4.25.*

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. *Expect them to say: No, we got fooled into picking the larger words.*

12. For how many of us did the sample estimate exceed the population mean? What proportion of the class is this? *AWV, but more than half, I expect.*

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. *No, they are biased to generally be larger.*

14. Compare the applet plot from question 8 with the plot from 10. Which method is closer to being **unbiased**? Explain. *Random sampling should win the day here. It is unbiased.*

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 $\boxed{25}$. Draw 3000 random samples of 25 words, and write down the mean and standard deviation of the sample means (rightmost plot). *AWV. Std Dev should be smaller.*
16. Sketch the plot of the sample estimates based on the 3000 samples drawn. Make sure to label the axis appropriately. *AWV*
17. Does the sampling method still appear to be unbiased? Explain. *Yes, because the distribution is centered at the true mean.*
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? *Same in that both are centered at 4.257. Different in that the st.dev is larger for* $n = 10$ *(it is 0.60) than for* $n = 25$ *(0.38).*
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. *Sample size 25.*
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. *Yes, as sample size gets bigger, the st.dev goes down.*

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" $\boxed{\text{x4}}$ under the data box. How large a population do you now have? Do mean and SD change? *964. mean and SD stay the same.*
22. With sample size set to $\boxed{25}$, 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. *AWV*

23. Sketch the plot of the sample estimates based on the 1000 samples drawn. Label the axis appropriately. *Hope to see center and spread have not changed much.*

24. Does the sampling method still appear to be unbiased? Explain. *Yes. It's centered at about 4.25.*

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? *Means of the two distributions are essentially the same, so also is the st.dev.*

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? *No. The sample mean for 25 observations has roughly the same mean and st.dev as it did with 241 in the population.*

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? *We sampled over and over to see how variable our statistics will be. We compared SE's for different sample sizes and population sizes.*

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?

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 $\widehat{p}$ for the sample proportion.)

   $\widehat{p} = 14/16 = 0.875$

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?

   Close to 1/2 picking the helper.

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.

   We saw that relative to the class, the instructor's number correct was quite high.

4. How could you use coin flips to model a child's choice of toy? For 16 kids?

   For one kid, let Heads = "Helper", tails = "Hinderer". Count the proportion of heads in 16 flips of a fair coin.

5. In using the coin, what assumption are you making about the kids' preferences?

   That they are just picking one toy at random with no real preference for the helper or hinderer.

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

$$H_0 : \ p = 0.5.$$

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

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?

Yes, it is possible to get 14 or more heads. Anything is possible in a random sequence, but it is not very likely.

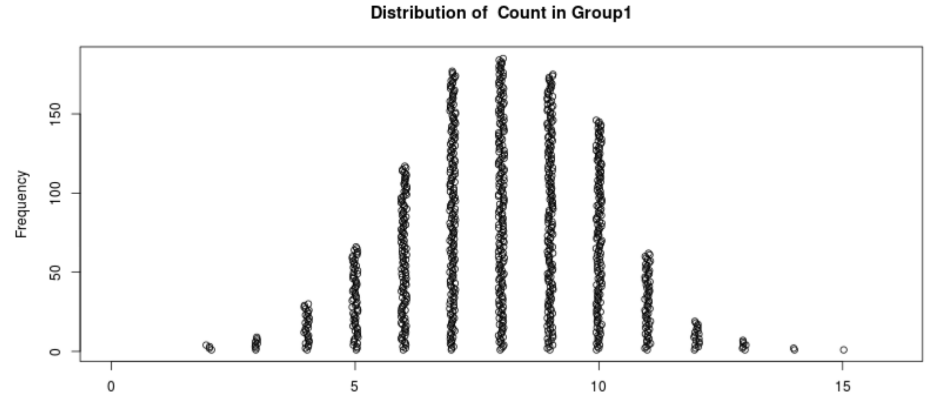
7. If infants had no real preference, would the observed result (14 of 16 choosing the helper) be very surprising or somewhat surprising, or not so surprising? How strong do you believe the evidence is against the null hypothesis?

   Pretty surprising, strong evidence against the null.

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 2, labels to hlp,hnd, Percentages to 50,50, Stop after Fixed number of spins, Stop after 16 spins, and click Run to see a simulation of 16 kids choosing helper or hinderer when the two are equally likely.

   12 in my first, 8 in my second

9. Set 1000 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?

    AWV, typically zero to 5.

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

    No, because it is unlikely that the observed result will pop up if kids are really picking at random.

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.

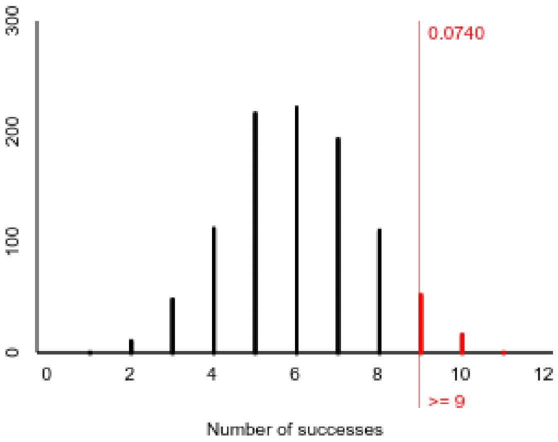


Figure 1: Simulation results obtained from the null model. The outcomes 9 and higher (74 out of 1000 trials) were as or more extreme as one instructor's number correct and indicate the strength of evidence of 0.074.

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

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.

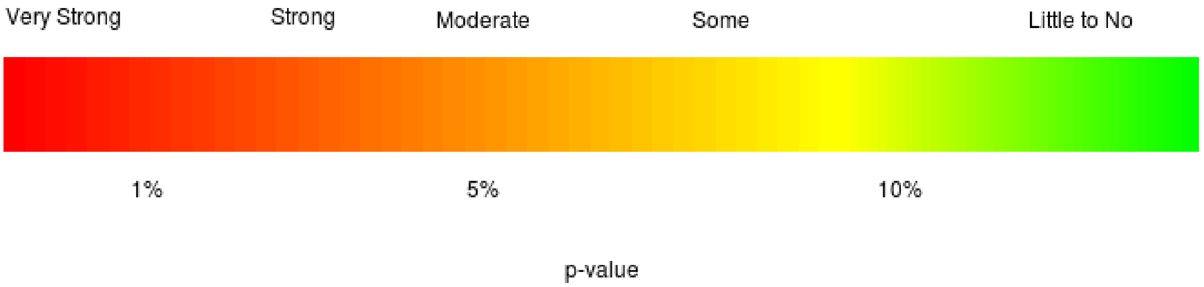


Figure 2: Numeric p–value and strength of evidence

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.

    1/1000 on my simulation which is "Very Strong" evidence against H_0.

13. What does this suggest about infants making their selections based only on random chance?

    It suggests they do not make their choices based on random chance and do make decisions based on social interactions.

14. Put the following steps into their proper order:

    (a) gather data *2*

    (b) formulate a hypothesis *1*

    (c) report strength of evidence *5*

    (d) simulate a distribution *3*

    (e) compare observed results to the distribution *4*

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.

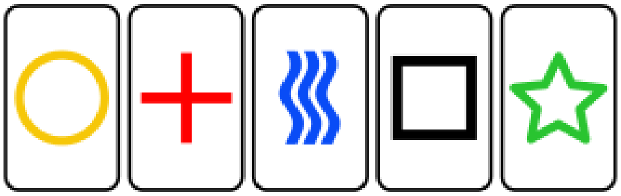
- The idea of p–value is very important in statistics. It will follow us all the way through the course. Stronger evidence means **smaller** p–value. Large p–values mean the data are not unusual under H_0.
- In any hypothesis test, we report p–values to the reader.
- In the remaining space add your own 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.

Can any person tell what someone else is looking at with no other form of communication?.

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?

   Right or Wrong

3. Your answer above gives the outcomes of the variable of interest in the study. Is this variable quantitative or categorical?

   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? $n =$

   25 guesses

5. Determine the observed statistic and use correct notation to denote it.

   $\widehat{p} = 0.36$

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?

   Yes, it is possible to do that well just by chance.

7. Do you think it is likely Linzmayer would have gotten 9 out of 25 correct if he was just guessing randomly each time?

   No.

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)?

   Rhine hoped to show Linzmayer actually had ESP (alternative), rather than that he was just guessing (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.

   $p = 1/5 = 0.20$

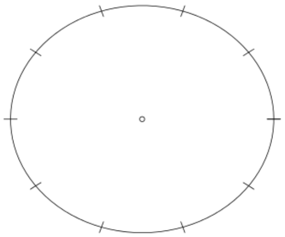
10. If Linzmayer is not just guessing and did have ESP, what values might the true proportion take on? Again, use correct notation to denote this range of values.

    $(0.20, 1.00)$ Is the observed statistic (9/25) in this interval?

    Yes

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: $p = 0.20$
    H_a: $p > 0.20$

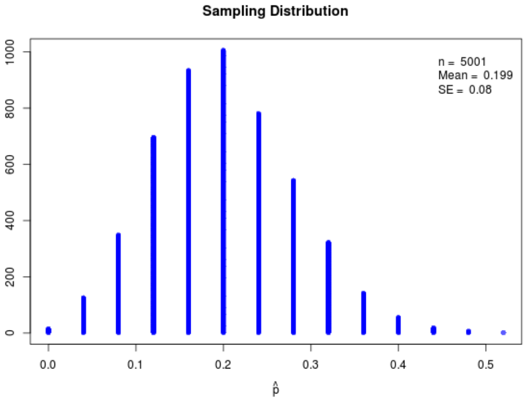
12. Think of a "spinner" on a game board. How would you subdivide and color it so that each spin would be equivalent to Linzmayer randomly guessing one card and getting it right/wrong with the null hypothesis probability. (Hint: you do not need 5 segments.) Sketch your spinner on the circle below and 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.

    $\widehat{p} = 0.36$

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. *My mean is 0.199, which is darn close to 0.20. They should agree because the simulations were created by assuing H_0 is true.*

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 Go and record the proportion of times this occurred.
    Would you consider this an unlikely result? $\widehat{p} = 0.36$, *I got 226 results of 5001 that were this extreme (p-value = 0.045), so doing this well is pretty unlikely.*

17. Go back to figure 2 to report the strength of evidence against H_0. Give the numeric and the verbal strength of evidence. *The p-value we observed is 0.045, which gives "moderately strong" evidence against H_0.*

Step 5: Formulate conclusions.
Based on this analysis, do you believe that Linzmayer was just guessing? Why or why not?

This is fairly strong evidence that he was not just guessing.

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

Yes, he could be cheating in some way.

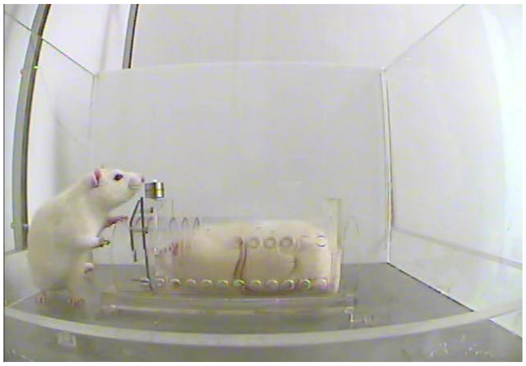
Another part of the scientific method is a reliance on replication. Other scientists tried to replicate this study and could not find another person like Linzmayer.

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.

   How large is the proportion of rats of this strain who show "compassion" to a trapped rat?

Step 2. Design a study and collect data.

2. What actions of the free rat will be recorded?

   Free the trapped rat or leave it caged.

3. Your answer above gives the outcomes of the variable of interest in the study. Is this variable quantitative or categorical?

   categorical

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

   p, the true proportion of rats of this strain who will free a fellow trapped rat.

Step 3. Explore and summarize the data.

5. What is the sample size in this study? $n =$ *30*

6. Determine the observed statistic and use correct notation to denote it.

   $\widehat{p} = 23/30 = 0.7667$

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.

   Not exactly, A few more or less than 23 would free the trapped rat.

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.)

Rats differ from each other. We've seen that not all 30 rats make the same choice, so if we randomly grab another 30 rats, we expect either fewer or more to release the captives, not always 23.

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.

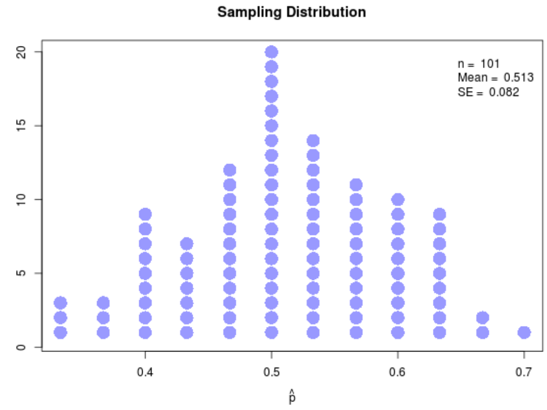
9. Reshuffling is slow, so we want to speed up the process by using the computer. Our goal is to see what other outcomes we might have gotten for different samples of 30 rat pairs. We will again use the One Categ. web app at `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/`. Select "Enter / Describe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

AWV

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



Where is the distribution centered?

approx 0.77

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

approx 0.078

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

The simulation is run assuming rats are freed with probability 0.77.

12. You should have 101 resampled values in your plot. If you have more than that, go back to Enter/Describe Data and make any small change in the spelling. Then come back to Estimate and click 100 just once.
Below the plot we have options for confidence limits for our interval estimate.

    (a) Click 80 and count: How many red points in the left tail? *10*
    How many reds in the right tail? *10*
    How many blue points in the middle? *81*

    (b) Click 90 and count: How many red points in the left tail? *5*
    How many reds in the right tail? *5*
    How many blue points in the middle? *91*

    (c) Click 95 and count: How many red points in the left tail? *3*
    How many reds in the right tail? *3*
    How many blue points in the middle? *95*

    (d) Explain how the confidence limit is related to the number of blue points.

    With 101 points, the percentage coverage matches the number of points left as blue in the middle.

    (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?

    The endpoints are the values at which colors change from red to blue or vice versa.

    (f) Predict: what will happen to the numbers in each tail for, say, a 90 % interval, if we double the number of resamples?

    Counts should double, percentages stay the same.

    (g) Click 100 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.)

    We resampled from 30 cards containing 23 "freed" and 7 "not freed" values 30 times. For each draw, record which type of card we get. Compute the proportion of "freed" rats in this resample.

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.

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

What Does "Confidence" Mean?

Mark Twain said:

> All you need in this life is ignorance and confidence, and then success is sure.

from quarterback Joe Namath:

> When you have confidence, you can have a lot of fun. And when you have fun, you can do amazing things.

and from scientist Marie Curie:

> Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained.

The above quotes (from brainyquote.com) refer to "self confidence" which is certainly important in any endeavor. In statistics, the word "confidence" is best summarized as **faith in the process** by which an estimate (in our case, an interval estimate) was created. A confidence interval carries information about the **location** of the parameter of interest, and tells us a lot about the **precision** of the estimate through the interval length.

In the news, interval estimates are often reported as a point value and a **margin of error**.

> 71% of Democrats and independents who lean to the Democratic Party say the Earth is warming due to human activity, compared with 27% among their Republican counterparts (a difference of 44 percentage points). This report shows that these differences hold even when taking into account the differing characteristics of Democrats and Republicans, such as their different age and racial profiles.

If you read the "fine print" from the Pew Research Center which conducted the poll, `http://www.pewinternet.org/2015/07/01/appendix-a-about-the-general-public-survey-2/` you find that the margin of error for a 95% confidence interval is listed as ± 5.1% for Republican/lean Republican and ±4.5% for Democrat/lean Democrat proportions.

Plus or Minus Confidence Intervals

In the web app used in previous activities, we clicked on a confidence level and the web app colored in the right number of dots as red to put our selected percentage of sampled proportions in the center (these stayed blue) and split the remainder into the two tails, turning these more extreme points red. We call this a "percentile" method because, for example, a 90% CI has lower endpoint of the 5th percentile and upper endpoint of the 95th percentile.

Another common way of building a 95% confidence interval is to take the estimated value and add and subtract twice the standard error of the statistic. A 95% confidence interval for p is then

$$\widehat{p} \pm 2SE(\widehat{p})$$

where $SE(\widehat{p})$ is a number coming from the plot on the web app. Why 2? Well, it's easy to remember, and with a symmetric distribution, 95% of the data will fall within 2 SD's (standard deviations) of the mean.

Margin of error is then the amount we add and subtract. In this case, it is twice $SE(\widehat{p})$. (Note: the parentheses do not mean multiplication, say of SE times $\widehat{p}$. They indicate that SE is a function of $\widehat{p}$, in the same way we use $\log(x)$ or $\sin(\theta)$.)

1. Go back to the rat data from Activity 6 where 23 rats opened the cage and 7 did not. Reenter the data in the One Categ part of the web app, and select Estimate.

   (a) Generate 5000 to 10,000 resamples and click 95%. Record the interval here:
   (0.60, 0.90)

   (b) Now write down the SE shown near the top right corner of the plot. (We will not use the mean of the plotted values).
   0.077

   (c) Add and subtract $2SE$ from the original proportion given in the box at left (**Do not** use the mean from the plot.) and write it in interval notation.
   $0.77 \pm 2 \times 0.077 = (0.63, 0.91)$

   (d) Compare the two intervals. Is one wider? Is there a shift?
   The percentile CI is shifted slightly to the left and is slightly wider.

Meaning of "Confidence"

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.

- Select a random sample from a population, measure each unit, and compute a statistic like $\widehat{p}$ from it.

- 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 from the One Categ menu.

The first slider on this page allows us to set the sample size – like the number of units or subjects in the experiment. Let's start with 40.
The second slider sets the true proportion of successes for each trial or spin (one trial). Let's set that at 0.75 or 75% which is close to the observed $\widehat{p}$ 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 wih 100 simulations of a 90% CI.

The upper plot shows 100 $\widehat{p}$'s – one from each of the 100 simulations.
The second plot shows the interval estimate we get from each $\widehat{p}$. These are stacked up to put smallest estimates on the bottom, largest on top. The vertical axis has no real meaning.

2. 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)? *lowest and highest CI's, resp.*

3. There is a light gray vertical line in the center of the lower plot. What is the value (on the x axis) for this plot and why is it marked? *It is the true parameter value: 0.75 if you followed the directions.*

4. What color are the intervals which do not cross the vertical line?
   How many are there?

   red, AWV about 10

5. What color are the intervals which cross over the vertical line?
   How many are there?

   green, AWV about 90

6. Change the confidence level to 95%. Does the upper plot change? Does the lower plot? Describe any changes. *Upper plot should not change. Each interval in the lower plot gets longer, so some that were red may turn green now.*

7. If you want an interval which is stronger for confidence (has a higher level), what will happen to its width? *it must be wider*

8. Go up to 1000 or more intervals, try each confidence level in turn and record the coverage rate (under plot 2) for each.

80	90	95	99

9. Now back to the Pew study we started with. Of the 2002 people they contacted, 737 were classified as Republican (or Independents voting Rep) voters and 959 as Democrats (or Indep leaning Dem).

   (a) What integer number is closest to 27% of the Republicans? Enter that value into the $\boxed{\text{One Categ}}$ "Enter Data" boxes as successes (relabel success as you see fit) and the balance of those 737 in the bottom box. Check that the proportion on the summary page is close to 0.27.

      i. What is your proportion of Republicans who think global warming is caused by human activity?
      199 "successes", 538 "Failures", $\widehat{p} = 0.27$

      ii. Go to the "Estimate" page and run several 1000 samples. What is the standard error?
      0.016

      iii. Find the "margin of error" for a 95% Confidence interval and write down the interval.
      ME = 0.032, 95% CI: 0.27$\pm 0.032 = (0.38, 0.302)$

      iv. Compare that interval to one you get by clicking the appropriate button in the app.
      almost identical: (0.237 , 0.303)

   (b) Repeat for the Democrats:

      i. Numbers of "successes" and "failures".
      700, 259

      ii. Margin of error and 95% CI related to it.
      0.028, (0.702, 0.758)

      iii. Percentile interval and comparison.
      (0.701, 0.758), again very close.

   (c) Explain what we mean by "confidence" in these intervals we created.
   We are 95% confident that the true

   (d) What can we say about the proportions of Republicans and the proportion Democrats on this issue? Is it conceivable that the overall proportion is the same? Explain.
   The intervals do not come close to overlapping, so we have to think that there is a strong difference of opinion between these two groups. I am "quite confident" of that.

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? *One could answer that winners of the award are their own small population, and we have a census of all Darwin Award winners. However, there are other idiots who have not yet won this competition, but seem to be working toward proving themselves. I would say that the sample observed represents a population of people who take risks which should be avoided.*

2. What is the parameter of interest? $p =$ *the proportion of all idiots who are male.*

3. What statistic do we obtain from the sample? Give proper notation, the statistic's value, and explain it in words. $\widehat{p} = \frac{282}{318} = 0.887$ *is the proportion of the sample which is male.*

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.

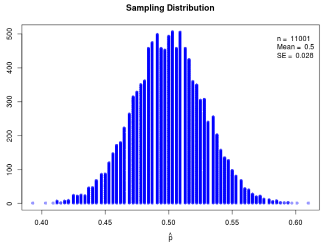
[2] Lendrem, B. A. D., Lendrem, D. W., Gray, A., & Isaacs, J. D. (2014). The Darwin Awards: sex differences in idiotic behaviour. BMJ, 349, g7094.

(a) State null and alternative hypotheses in symbols and words.
$H_0 : p = .5$. Half of all idiots are male. $H_a : p > .5$. More than half of all idiots are male.

(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? *Take an even number of cards (could be 318, but a smaller number will also work). Mark half of them male, half female. Shuffle and draw one. Record the gender. Return the card to the deck, repeat 317 more times and divide the total number of males by 318 to get one sample proportion.*

(c) Input the data under $\boxed{\text{One Categ}}$ in `http://shiny.math.montana.edu/jimrc/IntroStatShinyAp` and then select the $\boxed{\text{Test}}$ page. Do we need to change the "Null value" for p?

Click $\boxed{1000}$ several times to get a distribution of 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.

It's much bigger than any of the points I generated. The largest one I got was 0.585 or 186 males out of 318.

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

Extremely strong evidence, the p-value is less than 1 in 1000. The null hypothesis of only half males is not supported by these data. I conclude that there are more male idiots than female idiots.

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)?
   $\widehat{p} = 0.887$

   (b) In order to generate simulated data,

      i. How many individual "idiots" do we generate for one resample?
      318

      ii. Explain how you would mark 318 cards and use them to simulate the gender of one individual, and then another.
      Mark 26 "Female" and 282 "Male". Shuffle them and draw one at random. Replace the card, remix, and draw again for the second person.

iii. What probability of being male is used?
0.887

iv. After resampling 318 individuals, what number do you compute?
The proportion of the 318 new draws which are male.

(c) Use the web applet to create 1000 or more resamples from the original data.

i. Where is this distribution centered?
0.887

ii. What is the spread of the distribution of resampled proportions?
SE = 0.018

(d) Find a 95% confidence interval for the true proportion of idiots who are male.
$(0.852, 0.925)$

(e) Explain what the word "confidence" means for this confidence interval.
Our confidence is in the process, not in just one interval. If we repeat the process (gather a new random sample) over and over, 95% of the intervals we create will include the true parameter of interest.

6. Interpret this confidence interval.

We are 95% confident that the true percentage of idiots who are male is between 85.2% and 92.5%.

7. 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?

If H_0 is true, then the interval should contain 0.50. It does not, so the two inferences agree that one–half is not consistent with the data.

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 *The probability (proportion of simulations) of results as or more extreme as the observed result.*
- 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?

   Because only the observed result and nothing in the model actually changed, there is no reason to re-do the model. We just skip to step 4 and compare the observed result to the null distribution.

   (b) What strength of evidence against the null hypothesis does this new data provide?

   In my simulation of 1000, there were 238 trials with 10 or more picking the helper, which gives a strength of evidence of .238 = 23.8%

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

   With a strength of evidence of 9/1000 = .009 = .9%, I have strong evidence against the null model and can conclude that infants do in fact use social interactions to pick a toy.

(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? *The fraction is the same, but because the sample size is smaller, it is less unusual to see 7 of 8 picking helper than to see 14 of 16.*

(e) Explain how would you rerun the simulation for only 8 infants.

Change the number of trials to 8 instead of 16.

(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.

If they said: the same, then a response might be: The simulation showed my answer was wrong. There is less spread when the trial size was 16 than when it was 8. Due to the greater spread in trial size 8, there were more trials with 7 or more helpers chosen (approximately 40/1000) than there were trials with 14 or more helpers chosen out of 16 (approximately 1/1000).

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? $p =$ *the proportion of all couples who lean right when kissing.*

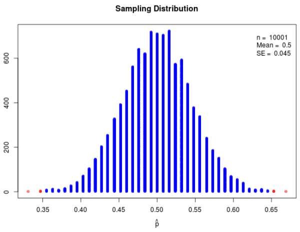
(b) What statistic do we obtain from the sample? Give proper notation, the statistic's value, and explain it in words. $\widehat{p} = \frac{80}{124} = 0.645$ *is the proportion of couples leaning right.*

(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.
$H_0 : p = .5$. *Half of all couples lean right when kissing.* $H_a : p \neq .5$. *The true proportion of couples leaning right when kissing is not one half.*

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? *Take an even number of cards (could be 124, but a smaller number will also work). Mark half of them right, half left. Shuffle and draw one. Record the lean. Return the card to the deck, repeat 123 more times and divide the total number of right's by 124 to get one sample proportion.*

iii. Use the $\boxed{\text{One Categ}}$ – $\boxed{\text{Test}}$ 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.
Only 8 of the 10000 points I generated are this extreme.

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?
Extremely strong evidence, the p-value is 0.0008 which is less than 1 in 1000. The null hypothesis of half leaning right is not consistent with these data. I conclude that more than half of kissing couples lean to the right.

(d) Now estimate the true population proportion.

i. What is our "point" estimate of the true proportion of couples who lean right?
$\widehat{p} = 0.645$

ii. In order to generate simulated data,

A. How many couples do we generate for one resample?
124

B. Explain how you would mark 124 cards and use them to simulate the lean of one couple, and then another.
Mark 44 "Left" and 80 "Right". Shuffle them and draw one at random and write down the lean on the selected card. Replace the card, remix, and draw again for the second person.

C. Each couple leans right with what probability?
0.645

D. After resampling 124 individuals, what number would you compute?
The proportion of the 124 new draws which are right.

iii. Use the web applet to create 1000 or more resamples from the original data.

A. Where is this distribution centered?
0.645

B. What number describes the spread of the distribution?
SE = 0.045

iv. Compute a 99% confidence interval.
$(0.532, 0.742)$

v. Explain what the word "confidence" means for this situation.
Our confidence is in the process, not in just one interval. If we repeat the process (gather a new random sample) over and over, 99% of the intervals we create will include the true parameter of interest.

(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?

If H_0 is true, then the interval should contain 0.50. It does not, so the two inferences agree that one–half is not consistent with the data.

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?
   This is tough. We want them to wrestle with questions of how much subjects knew before studying, how good a student they are, etc. (more in next question). Good responses: a post-test minus pre-test difference or we pick a topic which people don't generally know about before they study.

   (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? Focus in on one categorical and one quantitative variable.
   Some students are just smarter than others (IQ is quantitative). The subject area of the test they are studying for (anatomy versus philosophy?) – categorical. Years of musical training (quantitative). Surroundings (dorm room versus library) - categorical. Age (quantitative.)

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.
   SAT will generally be better because it allows us to use a comparable measure across all subjects, and it is immediately following the music treatment.

   (b) Is use of the four `sound` treatments an improvement over asking students how they study? Explain why or why not.
   The four treatment levels are generally be better because they take away many of the options people have when choosing music. We can then get a direct comparison of music versus no music and of lyrical versus instrumental music.

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?
   The second one from the article by Perham and Currie.
   Which was an observational study? Explain how you know this.
   The survey would be observational, because music levels are not assigned.

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?
   AWV

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 ($\overline{x}_T - \overline{x}_C$).

   (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/IntroStatShinyApps` and click Lurking Demo under One Quant. Select IQ. It gives you a bigger sample – 25 IQ's in each group. (newly generated at random from a symmetric distribution with mean 100 and SD = 15).

   (a) Write down the means for each group in the first shuffle and their difference.
   AWV

   (b) Write down the means for each group in the second shuffle and their difference.
   AWV

   (c) Compare your answers with another group's answers. Can you identify a pattern?
   centered about zero?

9. As we said above, we need to think about repeating the shuffling process over and over. Click 5000 and sketch the right-hand plot (above). Describe the center, spread, and shape of this distribution.

center *Near 0*

shape *Symmetric*

spread $SD \approx 3.7$

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?
No, all remain about the same.

11. Note that some differences are not close to zero. What are the largest and smallest values you obtained in 10,000 shuffles?

    ±12*?*

12. Does randomization always make mean IQs the same between the two treatment groups? Explain.

    Not every single trial though on average (over many shuffles) the groups are equivalent.

13. Does randomization tend to balance out mean IQ in the long run, after many trials? Explain.

    Yes, because the distribution of the difference between the two group means is centered at 0.

14. **Very Important Question:** In general, how similar are group mean IQs when we randomly assign people into two groups?
The difference in mean IQ between two randomized groups will be close to zero most of the time. There is no guarantee for any one sample, but in general randomization makes the groups very similar.

15. 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?
    categorical

    (b) Compute proprotion of "Y"s in the two groups of cards you had shuffled, and subtract. Write the proportions and the difference here.
    AWV

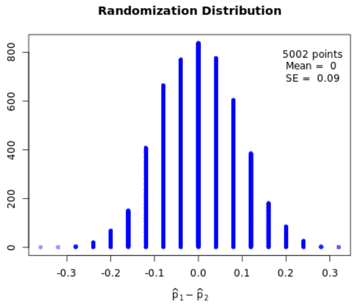
    (c) Again go to the web app and click Lurking Demo, but this time under One Categ. header. Change Success to Prep and Failure to No prep and enter the number of "Y"s, number of "N"s, and 6 in the treatment group and click Go. Run 5000 shuffles and record the mean and SE of the differences $\widehat{p}_1 = \widehat{p}_2$.
    AWV, but mean will be quite close to 0, SE about 0.3

    (d) You will have a few shuffles that give -1 or 1. How could that happen?
    One group got all the Prep people, the other group got none.

    (e) The plot gets more interesting with larger counts. Suppose we are randomly assigning 100 people to our two groups, and that 28 of them have taken SAT prep, 72 have not. Enter these values and 50 for size of the treatment group. Click Go. What proportions of "prep" in treatment and what differences do you get for the first two

randomizations?
.26, .28, differences: -0.04 and 0.00

(f) Run 5000 randomizations. Sketch your plot here.



(g) Compare with other groups. Do the pictures look the same? *Yes.*

center *Near 0 (0.00).*

shape *Symmetric*

spread *around 0.09*

16. 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.
Yes, the proportions might differ by as much as 0.3.

(b) Will the lurking SAT variable "usually" be poorly balanced across the two groups? Explain.
No. For most of the randomizations, we get a difference in proportion which is close to zero, which says that about the same proportion of treatment people as control people have taken SAT prep.

17. 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?
The difference in proportion with SAT prep between two randomized groups will be close to zero most of the time. There is no guarantee for any one sample, but in general randomization makes the groups very similar.

18. 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.
We can be pretty sure that the two randomized groups have similar proportions of people in them who took the SAT prep. Therefore, it's not a problem, and we can make out conclusions on the effects of music without worry about lurking variables.

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).

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

Bootstrap Confidence Interval for μ

We would like to know how much the "typical" MSU students spends on books each semester. Is this a question we can answer by testing? NO! We need an estimate, and as you now know, we like interval estimates because they include some information about uncertainty.

So far, the tools we have for working with a mean have dealt with testing some pre-specified value, not estimating an unknown parameter. We have a point estimate: a sample mean, $\overline{x}$, but we don't know how variable it is.

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 $\overline{x}$ is centered at the population mean, μ, and we know some things about its spread and shape. However, the sampling distribution of $\overline{x}$ 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 the applet:
`http://shiny.math.montana.edu/jimrc/IntroStatShinyApps` and select Bootstrap Demo under the One Quant menu.

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? *AWV. just one.*

(b) Click $\boxed{\text{1 Resample}}$ and watch what happens. Click $\boxed{\text{slower}}$ 1 or 2 times and watch it again. What is this button doing? *It selects 8 values from the sample with replacement, pulls each down to the next line, and leaves a colored spot on each one it grabbed. The resample then gets combined (averaged) to a single value and that is plotted on the dotplot scale.*

(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? *AWV.*

(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.

The second – sampling With Replacement is what we are doing on the computer. The first way always gives the same resample mean – they just change order. The second lets the resample mean vary.

(e) What statistic are we interested in (from the sample)? Compute it for the resample. *mean, $\overline{x}$, AWV*

(f) Click $\boxed{\text{100}}$ in the "Many Resamples" choices.

i. Explain what values are being plotted.
It takes 100 resamples, computes the mean of each, and plots the 100 resample means.

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. *One dot is the mean of one resample which was found by randomly selecting 8 values from the sample with replacement. We then average them together.*

(g) Click $\boxed{\text{500}}$ 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.
(16.4, 55.4) I see about 12 circles below and 12 above the interval.
Repeat twice more. Write down each confidence interval and guess how many points fall outside each. *AWV. I got (18.1, 54.9), (18.3, 54.9)*

(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?

The smaller numbers of resamples give more variability in CI. Centers don't change.

(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 $\boxed{\text{500}}$ again to obtain the interval.
going from 95% to 99% confidence intervals get longer
going from 95% to 90% confidence intervals get shorter

2. When we started, we saw the whole population of counts which has true mean $\mu = 34.5$ (\$345).

   (a) Look back at all the intervals you wrote down. Which ones contain the true value? *AWV. All of mine did.*

   (b) Click $\boxed{\text{Get New Sample}}$. 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 $\boxed{1000}$ again, write the interval and "missed" or "covered". *AWV. Mine 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?

   The sample. This is just like the simulation we did for proportions, but the method for computing the confidence interval is different.

3. With proportions we used $\widehat{p} \pm 2SE$ 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 $\overline{x} \pm t^*SE$ method. Also write the bootstrap intervals to compare.

   (a) Compute the mean of your sample (from the 8 values, not the "Mean" printed) $\overline{x} =$ *AWV, mine is 24.125*

   (b) a 90% CI for μ is (show work) *AWV,* $24.125 \pm 1.89 \times 5.25 = (14.2, 34.1)$
   Bootstrap: (15.4, 32.5)

   (c) a 95% CI for μ is (show work) *AWV,* $24.125 \pm 2.36 \times 5.39 = (11.4, 36.8)$
   Bootstrap: (13.8, 35)

   (d) a 99% CI for μ is (show work) *AWV,* $24.125 \pm 3.50 \times 5.31 = (5.5, 42.7)$
   Bootstrap: (11.3, 36.9)

4. Is there a pattern when you compare the two methods? Are bootstrap percentile methods always wider? shifted? relative to the $\overline{x} \pm t^*SE$ intervals?
   Bootstrap intervals are narrower. There is a tendency for the $\overline{x} \pm t^*SE$ *intervals to be too symmetric.*

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? *No, both are unbiased.*

   Will smaller sample size change the width?
   Yes, width should increase.

   Will larger sample size shift the center?
   No, both are unbiased.

   Will larger sample size change the width?

   Yes, width should shrink.

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?)

   Yes, feeding an infant 6 g per week of peanut protein or avoiding peanut protein.

2. What is the response variable in this study?

   Peanut Allergy at age 5 years.

3. Are the variables above quantitative or categorical?

   Both are 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	Avoiders	total
Allergic	5	35	40
Not Allergic	240	220	460
Total	245	255	500

5. Of the 245 subjects assigned to the eat peanuts, what proportion improved? We will label this $\widehat{p}_1$ 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 Activity 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.

   $\widehat{p}_1 = 5/245 = 0.02$

6. Of the 255 subjects assigned to the control condition, what proportion improved? We'll call this $\widehat{p}_2$, using a "2" for the control group.

   $\widehat{p}_2 = 35/255 = 0.137$

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. $\widehat{p}_1 - \widehat{p}_2 =$

   $0.02 - 0.137 = -0.117$

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 $\widehat{p}_T$ where T goes with "Total".

   $\widehat{p}_T = 40/500 = 0.08$

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.

   The data suggests that overall approximately 8% of participants developed peanut allergies, but the difference between the proportions who became allergic in the two treatment groups was -11.7%. This appears to be a large difference which supports the idea that eating peanuts early helps infants avoid peanut allergies.

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:

    $H_A : p_{treat} < p_{control}$

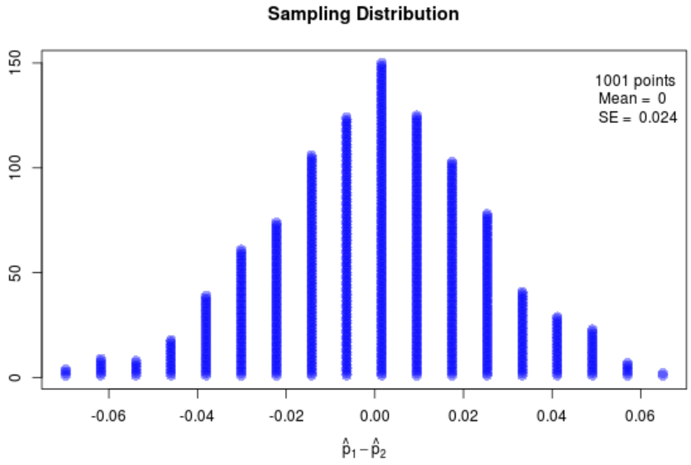
    Go to the web page: `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/` and select "Enter Data" under 2 Categ. Enter the numbers and labels from the table in 4. The proportions should agree with those above, but let's check:

    (a) The proportion of infants in Peanut group who became allergic:
    0.02

    (b) The proportion of infants in Control group who became allergic:
    0.137

    (c) The difference in proportions between the two groups:
    ±0.117

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.

    It is centered at 0 because the null hypothesis is that the treatment has no effect on allergy development in which case we should see the same proportion of allergic kids in both groups (or the difference in proportions should be 0). We are assuming H_0 is true.

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 Activity 6.)

    I got $0/10000 < 0.0001$

    Go back to $\boxed{\text{Enter Data}}$ and change labels slightly to clear the plot and then generate another 5000 random shuffles. How much does the strength of evidence change?

    I got $63/5000 = 0.0126$. *Little to no change.*

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

    strong

15. Based on the p–value, provide an answer to the research question. *With a p–value of 0.001, there is very strong evidence to reject the null hypothesis. We can conclude that in the group of infants from which these were drawn, eating peanut protein as an infant caused a reduction in prevalence of peaut allergy as evidenced by a lower proportion of allergic kids in the treatment group over the control group.*

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

    With a p–value of 0.25, there is little to no evidence against the null hypothesis. We cannot conclude that swimming with dolphins is therapeutic for patients suffering from depression.

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

    With a p–value of 0.73, there is no evidence against the null hypothesis. We cannot conclude that swimming with dolphins is therapeutic for patients suffering from depression.

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.

A randomization test for a difference in proportions with 5000 trials was used to test the null hypothesis that eating peanut protein has no effect on whether or not an infant will develop peanut allergy. In 500 participants randomly split into 245 treated and 255 control babies, 40 developed peanut allergies: 5 in the treatment group and 35 in the control group. This gave an observed difference in proportion of allergic kids between the treatment and control group of -0.117 (peanut proprotion minus control). This resulted in a p–value of < 0.0001*, which constitutes strong evidence against the null hypothesis. Since participants were randomly assigned to groups and are representative of all infants in the UK, we can conclude that in all UK babies, eatingpeanuts caused a lower proportion of infants to develop a peanut allergy as compared to a control group.*

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 study[3] 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.

	Gender	
Self Perception	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 $p_1 = p_2$ *(NO HATS!) The true proportions of people who over estimate their weight (versus underestimate it) is zero.*
   H_a $p_1 \neq p_2$ *(NO HATS!) The true proportions of people who over estimate their weight (versus underestimate it) are not equal.*

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. *Females:* $\widehat{p}_1 = 0.714$*, Males:* $\widehat{p}_2 = 0.145$*, Difference:* $\widehat{p}_1 - \widehat{p}_2 = 0.569$*,*

   (b) Switch to the "Test" page, and generate 1000 shuffles. Describe what the computer does for a single "shuffle". *It randomly shuffles the 60 successes and 79 failures into fake Female and Male categories (70 females, and 69 males).*

   (c) Why is the plot centered where it is centered? *Because the shuffling assumes* H_0 *is true, and that* $p_1 - p_2 = 0$

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

(d) Write down the p–value and strength of evidence. *In 10000 shuffles my largest difference was 0.30, so on the upper side I get less than 1 in 10000. This is a two-sided alternative, so we have to double the one-sided count. Still the p–value is less than 2 in 10000 and a difference of 0.569 is enormously strong evidence against the null hypothesis.*

(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.
  Men tend to underestimate and women to overestimate their weight relative to "ideal" 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_o 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)
Reject H_0.
State your conclusion about H_0 in the context of this situation. Be as specific as possible about the true proportions overestimating their weight. *We conclude that when people are wrong about their weight, a much larger proportion of women than men say they are over, rather than under weight.*

3. Now we will estimate the difference. You might want to look back a Activity 7 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? $\widehat{p}_1 - \widehat{p}_2 =$ 0.569

(b) The output shows results from 1 resample of the data. They have generated 70 "Female" responses using $\widehat{p}_1 = 50/70$ and 69 "Male" responses using $\widehat{p}_2 = 10/69$. What is the difference in proportions (over) for this one sample? *AWV. (0.40 to 0.74)*

(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? *Women: 0.714*
For each man being resampled? *Men: 0.145*

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. *For Women, write "Over" on 50 cards, "Under" on 20 (or 5 and 2) for 7 cards total). Shuffle, draw one and record success if it's "Over". Replace the card, shuffle, and draw the 2nd result. Continue til you get 70 results and count up the number of "Overs".*

(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). *I got mean*$= 0.569, SD = 0.068$

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

Level	Interval
80%	(0.483, 0.656)
90%	(0.454, 0.684)
95%	(0.427, 0.699)
99%	(0.396, 0.742)

(f) Also compute the approximate 95% CI using the $\pm 2SE$ method.
$0.57 \pm 2 * 0.069) = (0.43, 0.71)$

(g) Compare the two 95% intervals you created. *they should be quite close to each other.*

(h) Based on the 95% CI, is it plausible that $p_1 = p_2$? Explain your answer by referring to your CI. *The 95% CI for $p_1 - p_2$ does not contain zero, so it is not plausible that $p_1 = p_2$. There is strong evidence that women tend to overestimate their weight when they are wrong, and men tend to underestimate.*

(i) Interpret the meaning of this confidence interval in the context of this problem. What do we mean when we say "confidence"? *We are 95% confident that the difference in true proportions (women's proportion minus men's) of people who are not in the range* *of "ideal" weights, but mistakenly think they are over weight is in the interval (0.43, 0.71). Our confidence is in the process, which we know captures the true mean 95% of the time.*

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, $\widehat{p}_1 - \widehat{p}_2$.

- 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?

   AWV

2. What claims are made in the advertising of energy drinks?

   AWV

3. How do energy drinks interact with alcohol?

   AWV

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 Psycopharmacology* 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
Mean of RED+A is clearly lower (-8.3), Control (1.2)
Similarly, median for RED+A is -9.8; median for Control is 1.8
Spread (SD and IQR)
Control has smaller spread (SD = 3.9, IQR = 1.64) compared to RED+A (SD = 10.0, IQR = 10)
Shape
RED+A seems most symmetric (boxplot), Cotrol has some outliers.

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: $\overline{x}_1 - \overline{x}_2$.

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.)

Random variation

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?

   They should be similar but not exactly the same.

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 then RED+A and Control, why do the means change? (You are describing a model of how the data are generated)

   The scores within each group change so the means change. The person from RED+A no longer has to be in RED+A, etc. Who was in which treatment group is randomized.

9. Go back to the applet and select Test 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?

   They should! If it's reversed, swap the order in the data.

   (b) What are the means for control and RED+A in the reshuffled version? The difference?

   AWV, but difference should be closer to 0.

   (c) Explain how our shuffling the cards is like what the computer does to the data.

   The computer just randomized which group each score came from similar to shuffling the scores and replacing them into different groups.

   (d) Click 1000 three times. Where is the plot centered? Why is it centered there?

   Centered close to 0 because the null is that treatment has no effect (i.e. the means are the same).

   (e) Below the plot, keep more extreme and enter the **observed difference in means from the original data** in the last box. Click Go. What proportion of the samples are this extreme?

   0.008 in my sample.

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.)

    Genetic differences, difference in education levels, difference in amount of sleep the previous night, etc.

11. Lurking variables were discussed on Activity 10. When we randomly assign treatments, how should the groups compare on any lurking variable?

    The should be approximately equivalent in the long run (or on average).

12. Are you willing to conclude that the differences we see between the two groups are caused by REDA? Explain your reasoning.

    Most will say probably no because of the lurking variables or sample size (?) but we **can** *since we have random assignment.*

13. Write your interpretation of this interval.

    We are 95% confident that the true mean "change in RBANS score" is 2.7 to 15.99 points lower when people drink REDA than when they drink the control beverage. Be sure to indicate which one is higher (or lower).

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. Does it contain zero?

    (-16.65, -2.45) No.

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?
   birth weight, quantitative

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

   (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.
   Smoker: 6.306, nonsmoker: 7.084, difference: -.778

   (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?
   It does not seem like a difference of .78 is that large because of the spread of the plots (range from 1.7 to 8 for for smokers, 1.4 to 10 for nonsmokers).

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 kids in the study because we could not say that the sample of subjects was representative of a larger population, like 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 word "permutation" emphasizes that we could be assigning treatments, or just shuffling labels we observed 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.

   The mother's habit of smoking or not is not associated with weight of baby. Or baby's weights are the same, on average, for smoking and nonsmoking mothers.

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

   smoking: 6.31, nonsmoking: 7.08, difference: 0.778

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

   Answers will vary, I got 6.5

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

7.05

What is the difference in means between these two groups?

0.53

Evaluate the Results

8. Plot the differences in means from 1000 or more simulated trials. Sketch the plot below.

9. What does each dot in the plot represent?

   Each dot represents 1 re–randomized trial where the responses were kept the same but the habit for the responses was randomized (this represents what could happen under the null hypothesis that habit is not associated with birth weight). Clicking on a bar changes the plot of the Most Recent Shuffle showing you which re–randomized trial that bar represents. The placement of the point gives the difference in the mean birth weight between the two habits.

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

    Centered at 0 because this is showing us what could happen if the null model were true and the null hypothesis says there should be no difference in mean birth weight for smokers and nonsmokers.

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?

    The mean birth weight for babies whose mother smoked is lower than the mean birth weight for babies whose mothers did not smoke. Use less than -0.778

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.

    0.013

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

    Strong to very strong evidence against the null model.

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

    There is strong evidence against the null hypothesis that smoking habit is not associated with mean birth weight. We can conclude there is an association between these variables and that the mean birth weight is higher for babies with nonsmoking moms than for babies with smoking moms.

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?

    Yes because this is a random sample of all NC births, no because we didn't look at other states.

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.

    No because there was no random assignment of participants to a habit. (The researchers did not randomly assign smoking to some moms and nonsmoking to others. Possible alternative explanations are listed in #5.

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.

    A permutation (or randomization) test for a difference in proportions with 1000 shuffles was used to test the null hypothesis that birth weight is not associated with mothers smoking (or not). In a random sample of 200 births, the mean birth weight was 6.31 lbs for babies whose mother smoked and 7.08 lbs for babies whose mothers did not smoke. This gave an observed difference in means of 0.778 *(nonsmoking – smoking). This resulted in a p–value of 0.013, which constitutes moderate to strong evidence against the null hypothesis. Since births were randomly sampled from all North Carolina but were not randomly assigned to a habit, we can conclude there is an association between these variables and that the mean birth weight really is lower for all births to smoking mothers than to nonsmoking mothers.*

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.

    (0.063, 1.59) lbs

    (b) Write your interpretation of this interval.

    We are 99% confident that the true mean birth weight for babies born to nonsmoking Moms is .06 to 1.59 lbs greater than the true mean birth weight of babies born to smoking mothers. Be sure to indicate which one is higher (or lower).

    (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 vasoconstrictor 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.

   (a) *How high is the mean arsenic level for New Hampshire residents drinking from a private well?*

   (b) *Is there evidence that people drinking the ground water in New Hampshire are building up a hazardous level (over 0.15 on average) of arsenic concentration?*

2. Which research question should be answered using a hypothesis test and which should be answered using a confidence interval?

   The first should be answered with a CI because we are trying to estimate the parameter, and the second with a test (is there evidence?)

Step 2. Design a study and collect data.

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

   Arsenic concentration in toenails (ppm), quantitative.

[4] Ravenscroft, 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?

   notation: μ, is mean arsenic concentration in toenails for New Hampshire residents with a private well.

Step 3. Explore the data.
With quantitative data, we typically report and study the average value, or the mean.

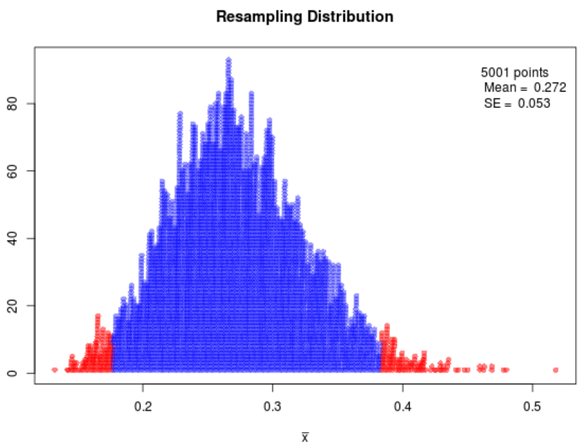
5. What is the sample size in this study? n = *19*
6. Calculate the observed statistic and use correct notation to denote it (check your answer with another group!). $\overline{x} = 0.272$.
7. Could your answer to 6 have happened if the arsenic concentrations in New Hampshire residents are not hazardous? *Yes, pretty much anything is possible if at least a few people have high arsenic levels.*
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?

   Not sure. Depends on how variable arsenic concentrations are!

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 11.

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. *Because we want to estimate the parameter (how high?), not compare it to some value.*
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.) *Write the original data on the cards, then sample with replacement 19 times.*
11. Use the $\boxed{\text{One Quant}}$ 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. *One dot was created by sampling with replacement from the original data 19 times. The mean level of arsenic concentration in the toenails of the resample is plotted.*

12. Create a 95% confidence interval using margin of error $ME = 2.11 \times SE$.

    $0.272 \pm 2.11(0.053) = (0.16, 0.38)$ *ppm*

13. Create a 95% confidence interval using the bootstrap Percentile Method.

    $(0.177, 0.383)$ *ppm*

14. How similar are the confidence intervals in 13 and 12?

    Pretty close, the percentile one is a bit higher on both ends than the 2.1SE interval.

15. Would you expect a 90% confidence interval to be wider or narrower? Explain, then give a 90% (percentile) confidence interval.

    Narrower, we need to capture fewer dots, so we can move in. (0.191, 0.365)

16. Interpret the 90% confidence interval from 15.

    We are 90% confident the true average arsenic level in toenails for new Hampshire residents with a private well is between 0.191 and 0.365 ppm.

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.
$H_0 : \mu = 0.15$*: The water is, on average safe, so we've just observed an unusually high sample.*

$H_a : \mu > 0.15$ *The water is, on average, dangerously high in arsenic.*

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

Right tailed.

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.)

    AWV, but if they have read/if you have lectured on this, they should know to shift the data to be centered on the null value, then resample.

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?

    0.15

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

    0.272 0.15 = 0.122 above the mean

    (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?

    0.15, because this is the null value.

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 have 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!

    Look for data entry errors.

21. What is the statistic from the first resample?

    AWV

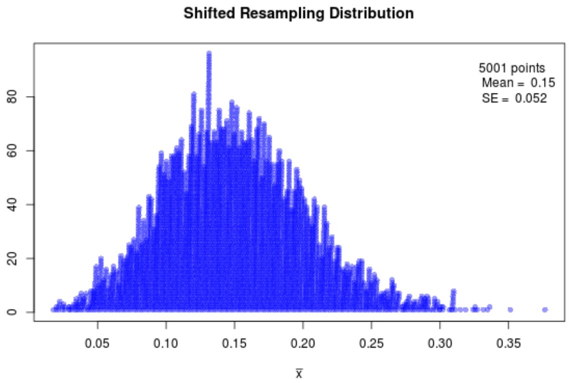
22. Where in the output did you find it?

    Below center tables.

23. Explain in the context of the problem what the one dot on the main plot represents.

    The (resample) mean arsenic concentration in toe nails of 19 New Hampshire residents who use a private well if the true arsenic concentration is 0.15 (not a hazardous level)

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?

    Again, centered at 0.15, because this is the null value.

    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.

    Pretty fair in the right tail. Appears unlikely to occur under the null hypothesis.

27. Just how unlikely is the observed result? Calculate your p-value using the web app and the appropriate direction and cutoff value.

    Right tail p-value = 0.016

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

    AWV

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.

    *0.016*3000 = 48*

    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?

    We have strong evidence that the true mean arsenic level in toenails of NH residents drinking from wells is greater than the hazardous threshold of 0.15 ppm.

30. Can you extend your results to all of New Hampshire residents? All New Hampshire residents with a private well? Explain your reasoning.

    Our evidence does not extend to all NH residents drinking from wells because we don't know that this is a representative sample. The data does not include people on municipal water systems, so it certainly does not extend to all NH residents.

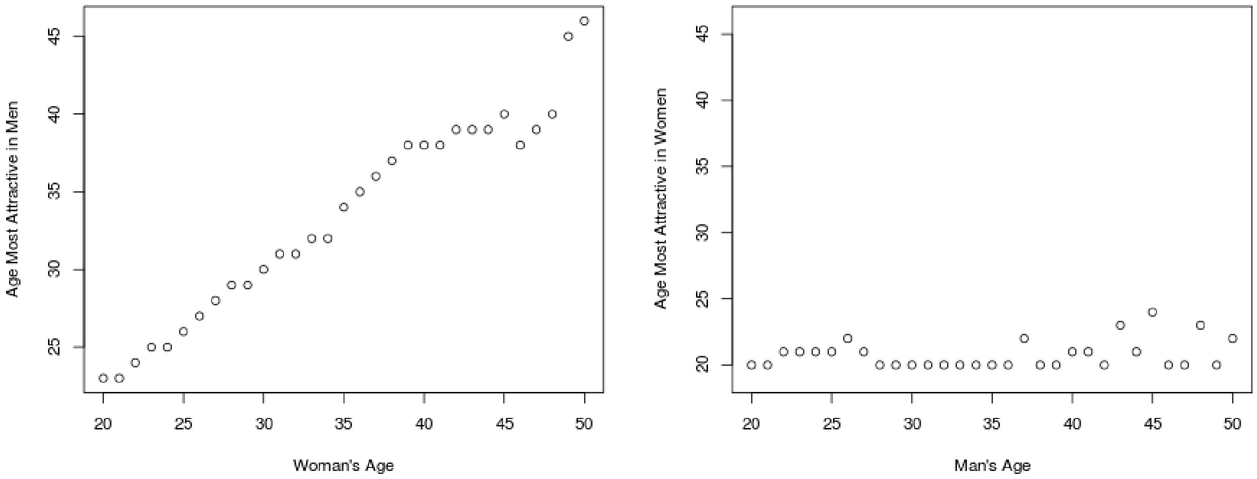
Take Home Messages

- 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 $\overline{x}$ to be centered at μ_o. Then we resampled it as if for a bootstrap confidence interval, and located the observed statistic ($\overline{x}$) 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?

   I'm hoping for some talk about fitting a line or about slope or correlation.

2. What statistical summaries differ in the two plots?

   Means of the y variable certainly differ, but again I hope they think of slope or correlation.

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?

   m for slope = rise over run?, b for intercept = y value when $x = 0$. Other answers are possible.

   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, $\hat{y}$.

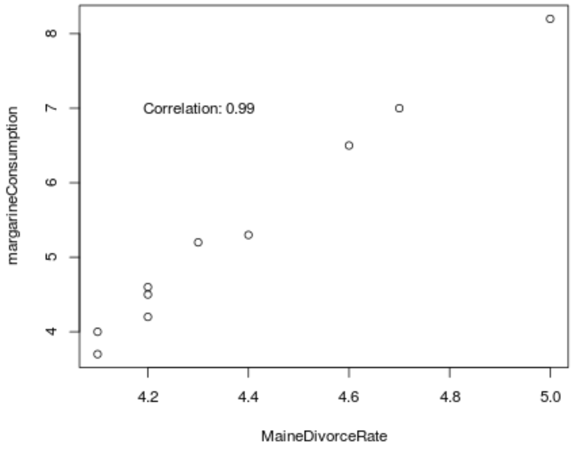
   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

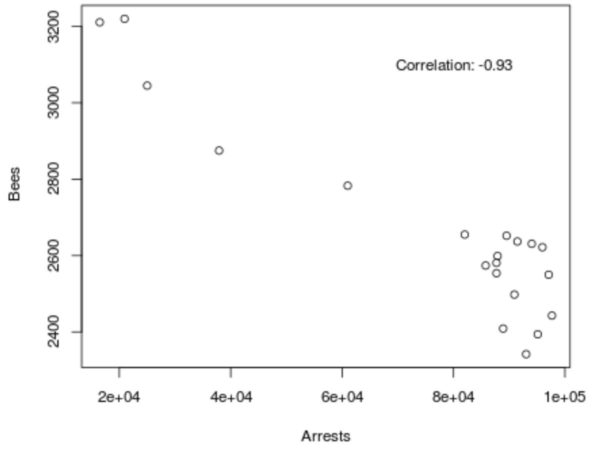
4. Take a guess at the slopes for the two plots above.

   Women: less than one – perhaps .75? Men: slightly bigger than zero, like 1/30?

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.

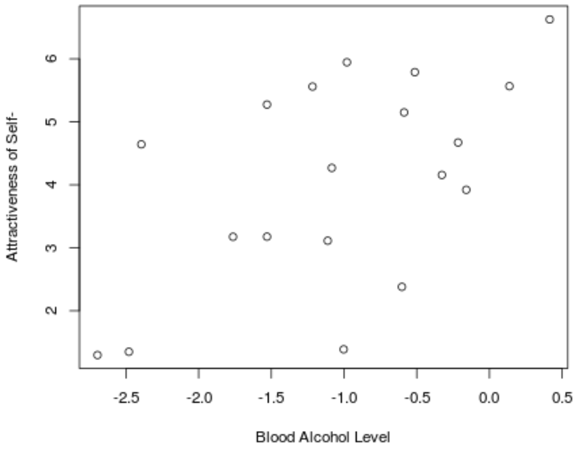
   Just to show you found the site, what variables are in the first plot, and what is their correlation?

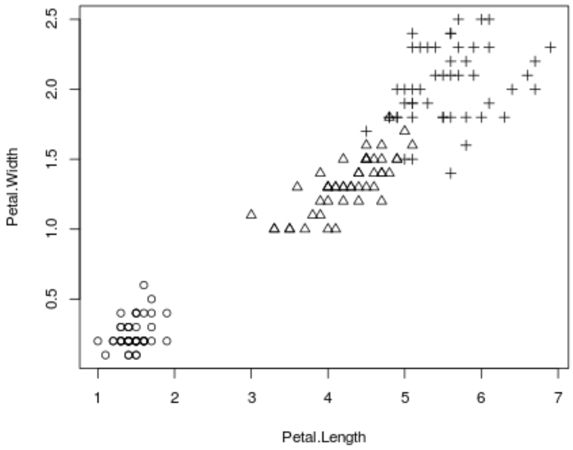
   US spending on Science, Space, Technology has correlation 0.99 with Suicides by hanging, suffocation, and drowning.

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

   Because there is no reason that the two variables should be correlated.

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





0.56 - attractiveness and BAL, and 0.96 - petals

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.

   0.98 and 0.28

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 New Samples. (Click Reset 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.

Linear, positive, moderately strong. $r = .71$

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:

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

I got intercept 40.5, slope = 0.94

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

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

$\widehat{y} = 38.3 + 1.03x$

(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?

(50, 56.4) has $r = 0.001$

(b) Now what is the slope of the regression line?

Close to 0 as well, I got a weird underflow problem, but it's about 0.0009

(c) Is correlation resistant to outliers? Is slope? Explain. *No. One outlier completely changed correlation and slope.*

14. Click $\boxed{\text{Revert}}$ 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.

(30, 74)

(b) Compute it's predicted value using the equation given. Also compute the residual for the one furthest below the line.

(30, 74) has the largest residual of $74 - (38.3 + 1.03 \times 30) = 74 - 69.2 = 4.8$. *Smallest comes from (24, 56) with predicted value* $38.3 + 1.03 \times 24 = 63.02$ *so its residual is* $56 - 63.02 = -7.02$

(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).

235 Any other line will have larger SSE.

Take Home Messages:

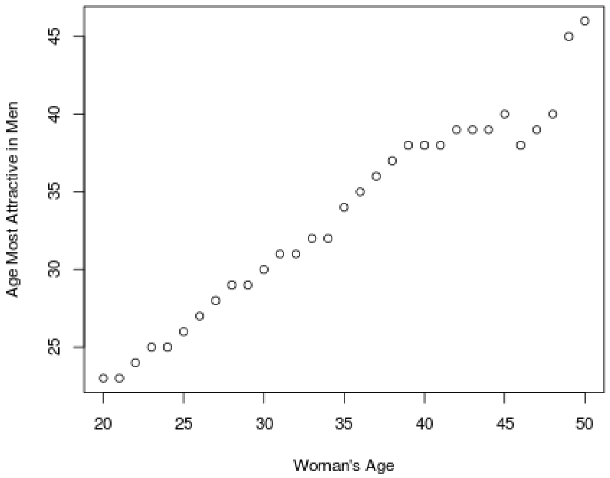
- It's not right to speak of the correlation between gender (or any categorical variable) and age, or between two categorical variables.
- It only works for linear relationships. We can have very strong nonlinear association with correlation near zero.
- Positive relationships mean large values in one 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.

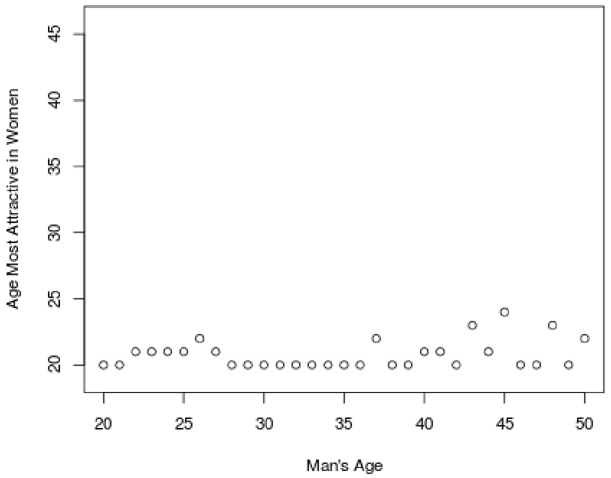
  $$\hat{\beta}_1 = r \times \frac{s_y}{s_x}$$

  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:

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

   Plug 36.5 in as the x value and compute $\hat{y} = 34.47$.

2. What would you guess men aged 49.5 would say is the most attractive age of women?

   Plug 49.5 in as the x value and compute $\hat{y} = 21.27$.

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 $\overline{y} = 20.78$ 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.

   I hope they can see some validity in both regression line and flat line.

   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 $\boxed{\text{Two Quant.}}$ 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: $most\widehat{AttrW}omen = 20.77 + 0\times\ mansAge$

(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?
20.77

(c) What does that say about the relationship between x and y?
y does not depend on x

(d) What will be the true slope for y based on x if there is no relationship? Use correct notation.
$\beta_1 = 0$

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.
$H_o : \beta_1 = 0$
$H_a \; \beta_1 > 0$

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.

   Y's are shuffled. If H_0 is true, then there is no linear association between x and y, so the shuffled data could have occurred just by chance.

7. Is the least squares line in the lower plot flatter or steeper than the one in the upper plot? Is $\hat{\beta}_1$ closer or further from zero?

   AWV. generally flatter

8. Take at least 1000 shuffles and compute the p-value. Explain which shuffles you are counting.

   In 10000 shuffles, I had 549 with slope greater than 0.0343, so my p–value is 0.055

9. State your decision and conclusion using $\alpha = 0.05$.

   At the $\alpha = 0.05$ significance level we do not reject H_0, We have only moderate evidence that the true slope between men's age and the age of woman they find most attractive is greater than zero. In other words: There is a weak positive association between the two variables which is not significant at the 0.05 level.

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 $H_a : \; \rho > 0$?.

    $r = 0.287$ *and the p-value is the same.*

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.
    $H_o : \beta_1 = 0$
    $H_a \; \beta_1 > 0$ *We don't expect a negative relationship.*

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

$\widehat{mostAttrMen} = 9.02 + 0.6972 \; x \; womansAge$

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

Use the same woman's ages from 20 to 50. Shuffle the ages of the men each woman finds most attractive (y)

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

I get 0. In 10000 shuffles, none was as extreme as the slope we observed. The p-value is the probability of seeing a slope this far above 0 if, in fact, woman's age (x) and most attractive man's age (y) are really unrelated (true $\beta_1 = 0$).

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

At the $\alpha = 0.05$ significance level we reject H_0, We have super strong evidence that the true slope between women's age and the age of man they find most attractive is greater than zero. In other words: There is a very strong positive association between the two variables which is significant at the 0.05 level.

(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 $H_a: \; \rho > 0$?.

$r = 0.982$ *and the p-value is the same.*

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

    No. The data are averaged across each age, and we can't say that the people using this dating site represent a larger population of singles.

13. Was some treatment randomly assigned?

    No. The "explanatory" ages (x values) were observed for each person, not assigned.

14. What is the scope of inference?

    We can only say there is (for women, is not for men) evidence of association in these samples of singles.

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

Take Home Messages:

- A slope of zero is very "special" in that is 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 $H_0 : \quad \rho = 0$. 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:

	Decision:	
H_0 is:	Reject H_0	Do not reject H_0
true	*Type I Error*	Correct
false	Correct	*Type II error*

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.

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

   0.484

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

   13

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?

   0.484 Smaller, down to 1.4

   What is your effect size?

   $2 - 0 = 2/1.4 = 1.429$

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?

   Larger, 2.9

   What is your effect size?

   $(2.9 - 0)/2 = 2.9/2 = 1.45$

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

   Larger effect size in the latter.

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

   Larger effect for same SD means more power, lower SD for same effect means more power. In general, larger effect and lower SD means more power.

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

   .971

   Change it to 0.10. What is the power?

   .992

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

    It would decrease (power and significance level change in the same direction).

11. Suppose that we are planning to do a study of how energy drinks effect RBAN scores similar to the study we read about in Activity 14. 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?

    33

    If we choose $\alpha = .01$, how large a sample is needed? *50*

12. Now suppose that we are using a memory test used to study sleep deprivation. Historical data provides an estimate of SD = 13. We want to use $\alpha = .05$ and need to detect an decrease in mean score (when people are sleep deprived) of 6 with 80% power. How large a sample is needed?

    31

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

    Type II = 1–power so we want more than 90% power. Need 43 people.

13. Suppose we do another study on energy drinks with alcohol using Control and REDA. This time 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?

    To conclude that there is a difference in mean coordination between the two treatments when, in fact, REDA has no effect on coordination scores.

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

    To fail to find a difference in mean coordination between the two treatments when, in fact, REDA lowers coordination scores.

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 assuming 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 and Explanatory variables
- Random Assignment (why do we do it?)
- Random Sampling
- Lurking Variables
- Causal inference (versus 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 is called β. 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	Exp	Music or Quiet	SAT score
Book Cost	Obs	None	Cost of Textbooks
Peanut Allergies	Exp	Eat peanut protein or not	Allergic to peanuts at age 5
Nonideal Weight	Obs	Male/Female	Over/Under ideal weight
Energy Drinks	Exp	REDA/Control	change in RBANS
Birth Weight	Obs	Smoking/non	baby weight
Arsenic	Obs	None	arsenic level
Attraction	Obs	Age of interviewee	Most attract age in opp sex

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.

   One fewer allergic kid in the treatment group and one more in control make this even stronger evidence against the null hypothesis. The difference in proportions becomes -0.125, and in 5000 randomization trials, I never got one sample with this large a difference in sample proportions, so p–value is $< 1/5000 = .0002$ *Our conclusion is the same.*

   (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.

   A good guess is that the same proportion in the smaller study provides weaker evidence. It does. When I run 5000 trials with 100 kids, the differences I got 16 values < -0.117 *for a p-value of* $< .003$.

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.

   Question: Is the proportion of teens in the US with hearing loss still 14.9%, or has it increased?
   Explanatory variable: year of survey
   Response: Some hearing loss.

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

(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.

We can infer association back to the populations of teenagers (2004 and 1991), but it is not an experiment, so we cannot make causal inference.

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

Sample proportions. We cannot take a census to find the true population proportions.

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

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]

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

(a) What percent of women in this study have breast cancer?

160/10000 = .016 = 1.6%

(b) 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?

Being told she has cancer

(c) Estimate that error rate using these data.

700/9840 = 0.071 = 7.1%

(d) If a woman really has cancer, what would an error in the test be saying? Is that a false positive or a false negative?

That she has no cancer, a false negative.

(e) Estimate that error rate using these data.

90/160 = .563 = 56.3% *That seems poor!*

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.

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

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

I rate death from a cancer which should have been detected as more critical than the anxiety of a false positive, so I think false negatives are more important.

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.

Randomized experiment because the treatment (talk or count) was assigned randomly.

(b) Calculate the difference in the means.

2.625

(c) Write out the null model for this study.

Mean blood pressure is the same for people talking or counting.

(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.

Running 5000 trials of a randomization test, I got a p–value of .101 which gives only weak evidence against the null hypothesis of equal means.

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 preloaded as "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) to get that ordering.

$\widehat{distance} = 17.37 + 0.4795xage$

(b) How fast is this measurement changing?

It increases by an estimated .48 mm for each year.

(c) What hypotheses are we testing?

$H_o : \beta_1 = 0$

$H_a\ \beta_1 > 0$ *We don't expect a negative relationship.*

(d) Compute the p-value for the hypothesis test.
0.0009 or 9 in 10,000 for me

(e) Give the scope of inference.
Association in the sample.

Unit 3

Theoretical Distributions

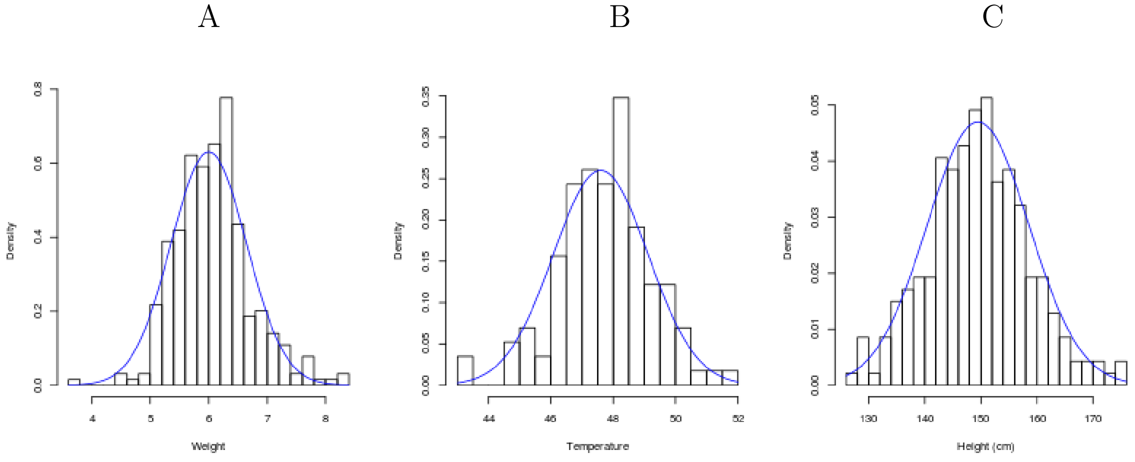
Whatever your field of study, you will need to read research articles which present statistical inference like p-values and or confidence intervals. We hope you now understand how those are properly used. In particular:

- Always report p-values. Smaller p-values provide stronger evidence against the null hypothesis.
- Confidence intervals are a list of "plausible" values. Our confidence is in the process used to build intervals, not in one interval.

Many reports you read will refer to tests based on normal distributions rather than randomization and simulation. Before we had modern computers and the web apps we've been using, people had to use tables to look up probabilities to make confidence intervals and compute p-values. We'll now look into these methods as a "short cut" alternative to simulations. We encourage you to take more statistics, for example regression is a powerful technique used to describe relationships between quantitative variables. We are happy to visit with you about possibilities for more statistics (Stat 217 is a great second course). Part of the motivation for this lesson (and Unit 3 in general) is to make it easier to continue your statistical adventures.

Shapes of Data Distributions

Look at these three different data sets:



A) Weights (g) of newly born lab rat pups. B) Mean annual temperatures ($°F$) in Ann Arbor, Michigan. C) Heights (cm) of 14 year old boys in Oxford, England.

1. In what way are these distributions similar and different?

   Means and spreads all differ, but the shapes are quite similar – sort of bell-shaped.

Many distributions we look at have a shape similar to those above. Most of the data lies close to the mean, and the left and right sides are symmetric. We call it "bell-shaped" and the best example is called the "Normal" distribution.

Important fact:
Statistics vary from sample to sample, and the pattern is predictable. For many statistics, the pattern of the sampling distribution resembles a normal distribution with a bell-shaped curve. Studying the normal distribution will allow us to find probabilities for statistical inference which do not require running simulations.

Normal Distributions all have the same shape.
They differ only in mean (μ) and spread (σ). We write $X \sim N(\mu, \sigma)$ to say that random variable X is normally distributed with center μ and spread (officially: standard deviation) σ. This distribution has two **parameters**, μ and σ.

Definition: Random Variable is a number which comes from some random process, like randomly selecting one unit from a population and measuring it.

The **Standard Normal Distribution** has center $\mu = 0$ and spread $\sigma = 1$. We can "standardize" any normal distribution to make it have center 0 and $\sigma = 1$ by subtracting μ and dividing by σ. If a random measurement, X, has a N(μ, σ) distribution, then

$$Z = (X - \mu)/\sigma$$

has a N(0,1) distribution. We use the standardized versions to say how many standard deviations (σ's) an observation is from the mean (μ).

2. When you get back results from standardized tests, they give a score and tell you your "percentile rank", the percentage of test takers who score at your level or below. The exam scores have an approximately normal distribution. For example, with ACT, $\mu = 21$ and $\sigma = 5$.

   (a) What is the standardized $z = (x - \mu)/\sigma$ score for Amber who got 27 on the ACT?

   $$\frac{25 - 21}{5} = 1.20$$

   (b) What is Amber's percentile rank? Select Normal Distribution under One Categ. on the `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps/` app and plug her standardized score into the first line of this web app. Explain what the number means in terms of other test takers.

   .885 or 88.5% of students taking the ACT are at this level or lower.

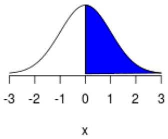
   (c) Bert took the SAT instead of the ACT, and SAT scores are normally distributed with mean $\mu = 1500$ and spread $\sigma = 300$. Bert's score was 1720. What is Bert's standardized score?

   $$z = \frac{1720 - 1500}{300} = 0.733$$

(d) What is Bert's percentile rank? Compare how well he did on the SAT with how well Amber did on the ACT.

.768 or 76.8% of students taking the SAT are at this level or lower. Amber did better relative to others taking the ACT than Bert did relative to SAT takers.

Definition: Probability: the proportion of times – in the long run – that a particular outcome occurs. For example, the probability of drawing a "heart" from a well-shuffled deck of cards is 0.25. Probabilities must be between 0 and 1.



Normal Probabilities are areas under the normal curve. Area under the entire curve is 1. What area is shaded for this normal distribution?

We can also go from a percent to a percentile (or, from a probability between 0 and 1 to a **quantile**) by back–solving the "Z" formula for X like this:

$$\text{Start with } Z = \frac{X - \mu}{\sigma} \text{ and solve for } X \text{ to get } X = Z\sigma + \mu$$

What SAT score is needed to be in the top 10% of SAT takers?
In the same web app, put 0.10 in the second box and click upper (or 0.90 and click lower). That returns a Z value of 1.282, so $X = 1.282 \times 300 + 1500 = 1884.5$. SAT scores are always multiples of ten, so a score of 1890 puts a person into the top 10%, or above the 90^{th} percentile.

3. **Your Turn:** Suppose birth weights of full term babies have a $N(3000, 700)$ (in grams) distribution. Show your work, that is, how you standardize each value, or "unstandardize" to get a percentile.

(a) How heavy is a baby whose weight is at the 98^{th} percentile? The 5^{th} percentile?

Standardized percentiles are 2.054 and -1.645. Converting to birth weights(times 700 + 3000): 4438g and 1848.6g.

(b) What is the probability that a randomly chosen baby will weigh under 3500g? Under 2500g?

$Z = \pm 500/700 = \pm 0.714$, *Probabilities: 0.762, 0.237*

(c) What proportion of birth weights are within one standard deviation of the mean? Within 2 SD's? within 3 SD's?

0.683, 0.955, 0.997

Note: The last question gives us a useful rule of thumb which we call the empirical rule. The middle value (probability of being within 2 SD's) is usually rounded to 95%.

t Distributions

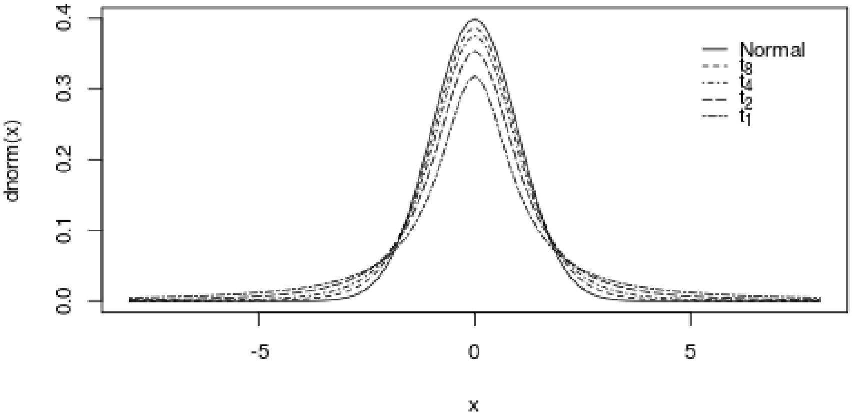
To standardize a normal distribution, we need to know its true mean, μ, true standard deviation, σ. Problem: we never know the true parameter values. We can work around the unknown mean pretty easily, but not knowing σ causes a bit of a problem.

Discuss: Look back to Activity 2 – estimates of spread. What can we plug in to estimate the unknown population spread, σ?

s, the spread in the sample

In the early 1900's, Mr. Gossett was working for Guinness Brewery. He discovered that using an estimated instead of "known" σ changes the distribution from regular normal to something called a t-distribution which is more spread out. Furthermore, there is not just one t-distribution, but many depending on your sample size.

This make sense because as sample size gets big, s becomes a better and better estimate of σ. Here is a picture of several t-distributions compared to a standard normal distribution.



We can standardize like this:

$$t = \frac{X - \overline{x}}{s} \quad \text{or go the other way:} \quad x = \overline{x} + t^* \times s$$

and use the same web app to compute the probabilities and quantiles for t-distributions just as we did with normal distributions. The one additional fact needed is that for a one-sample mean, we use $n-1$ (sample size minus 1) as the **"degrees of freedom"** which define the t distribution needed. When you select t under One Quant. or Two Quant, you'll be able to set the degrees of freedom.

Example:
Heights for adult men in the US are close to normally distributed. We'll take 70 inches to be

the mean height. From a simple random sample of 20 men, we compute a sample standard deviation of $s = 3.3$ inches.

6. You will use a t-distribution with what degrees of freedom?

   19

7. Use the appropriate t distribution to determine how unusual it is to see a man who is over 76 inches tall. Show your standardized value and the probability.

   t = 1.818, $P(t_{19} > 2) = 0.042$

8. Under 68 inches tall?

   t = -1.212, $P(t_{19} < -1.212) = 0.12$

9. Between 65 inches and 75 inches? (You have to enter each standardized value, selecting "Lower", and subtract to get the area in between.)

   t values are 1.515 and -1.515. subtract 0.073 from 0.927, or use the $\boxed{\textit{Center}}$ *choice to get 0.854.*

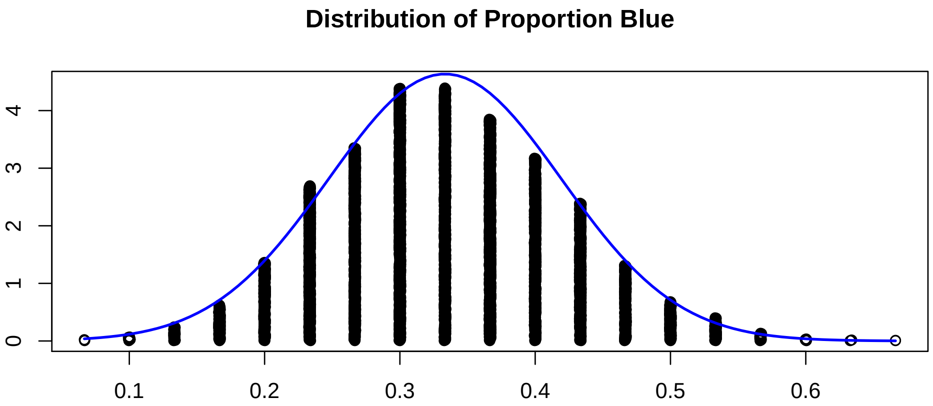
Take Home Message

- Theoretical distributions are a shortcut method to obtain probabilities for p-values and confidence intervals.
- Normal and t distribution probabilities are areas under a curve. We can get them from statistical computer packages.
- Both normal and t distributions are symmetric and bell–shaped. The t-distributions are more spread out because we have to estimate the standard deviation.
- When σ is known use a normal distribution. Otherwise, use a t-distribution with $n - 1$ degrees of freedom for just one sample. (This changes for 2 or more samples).
- To look up normal or t probabilities, we have to standardize by subtracting μ and dividing by σ (for normal) or s (for t). The standardized score tells us how many standard deviations we are from the center.
- You need to be able to go either way: find a probability from a Z or t score, or find the percentile from the probability.
- Empirical Rule for data with a normal distribution:

  - 68% of the values fall within one SD of the mean.
  - 95% of the values fall within 2 SD of the mean, and
  - 99.7% of the data fall within 3 SD of the mean.

Proportions

If we mix 40 blue balls and 80 gold balls together in a jar and draw out 30 at random with replacement, the sampling distribution for the proportion of blue balls in a sample of size 30 looks like this:



Each of 5000 dots came from a computer simulation in which we sampled 30 draws from the container at random with replacement, and computed the proportion of blue balls in the simulated sample. The curve shown on top of the dots is the density of the normal distribution with the same center and spread as the distribution of dots.

With modern computers, we can easily simulate random draws five or ten thousand times, and/or we can easily obtain probabilities from the normal distribution. Both are valid tools, and each method can give us the p-values and confidence intervals that we need. We did the simulation approach first because it allowed us to bypass two weeks of more theoretical probability and quickly get to the meat of this course – statistical inference. Now it's time to see how the other approach works, and we'll continue to emphasize the interpretation of p-values and confidence intervals.

We can use spinners and coin flips to simulate the distribution of a proportion. From that distribution we use the standard deviation of the resampled points to get the standard error of $\widehat{p}$. We added two standard errors to – and subtracted two standard errors from – the point estimate to build our confidence interval. In order to use the normal distribution instead of a simulation, we need this formula for standard error of the sample proportion:

$$SE(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$$

We can then multiply $SE(\widehat{p})$ by the appropriate z^* value from the normal distribution to get whatever confidence level we need. Setting margin of error to $2SE$ was a good approximation for a 95% confidence interval, but you'll see in this web app that a more precise value is 1.96 SE's on each side.

Notation: We have two ways to describe the same measure of spread. For any distribution, we can compute the spread, or standard deviation of the sampled values. When we talk about the sampling distribution of a statistic, we can refer to the standard deviation of the sampling distribution because it is a distribution, so it has some "spread". However, we prefer to talk about the "Standard Error" (or SE) of the statistic. We'll be able to show the difference better when we look at the mean of a continuous variable, so we'll come back to this.

When we write $SE(estimate)$ as in $SE(\widehat{p})$ or $SE(\overline{x})$, we **do not** mean to multiply SE by the estimate. This is notation to say SE is a function which we apply to the estimate. We read it as "standard error of ..." just like when we take log of x we write $\log(x)$.

Confidence Intervals

The general form is

$$\text{estimate} \pm z^* SE(\text{estimate})$$

or for proportions:

$$\widehat{p} \pm z^* \sqrt{\widehat{p}(1-\widehat{p})/n}$$

To build confidence intervals, we use the same z^* values over and over again. Go to the web app `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`, select Normal Distribution under One Categ and put confidence levels 0.80, 0.90, ..., 0.99 into the probability box (one at a time). Change Lower to Center to get the confidence limits. Write them into this table.

Confidence level:	80%	90%	95%	98%	99%
z^* cutoff	1.282	1.645	1.96	2.326	2.575

Let's try it out:

1. In a survey of 2142 American adults, 1221 gave the opinion that college is a "poor to fair" value. We want to estimate the proportion of all American adults with that opinion using a 99% confidence interval.

   (a) Compute $\widehat{p}$ for these data.
   0.57

   (b) Compute the standard error of the estimate.
   $\sqrt{0.57 \times 0.43/2142} = 0.0107$

   (c) Find the margin of error and compute a 99% CI using the multiplier you found above.
   $ME = 2.575 \times 0.0107 = 0.02755,$ *CI:* $0.57 \pm 0.0276 = (0.542, 0.598)$

(d) If we use resampling to create a 99% bootstrap percentile confidence interval, it is $(0.542, 0.597)$ and the SE in the plot is 0.011. Which interval is narrower?

Bootstrap is a hair narrower.

How similar is the standard error from 1b to the bootstrap SE?

Very close. Off just by round off error?

(e) Interpret the interval in the context of this problem. What do we mean by the word "confidence"?

We are 99% confident that the true proportion of US adults who thought that college was a poor to fair investment is within the interval (0.542, 0.597). Our confidence is in the process: when we use this method over and over take a sample and compute a 99% CI from it, in the long run, 99% of those intervals will contain the true parameter.

Assumptions

To do any statistical analysis, we must make some assumptions. We should always check to see if the data indicate a violation of an assumption. For the methods used before today we need:

- A population of size at least $10n$. (If a sample is a really large part of the population, our methods over-estimate sampling variation.)
- Representative sample.
- Independent trials (one outcome has no effect on another).

Using normal distributions adds another assumption:

- Large enough sample size to expect at least 10 successes and at least 10 failures. If you are building a confidence interval, just make sure the counts are over 10. If you are conducting a hypothesis test, we need $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

2. In an earlier assignment you checked to see if a roulette wheel was fair. Were these assumptions met? The observer saw 3 "Greens" in 30 spins, and the chance of green for a fair wheel is 2/38.

   (a) Is population size at least $10n$ ($n = 30$)? Hint: How many spins are possible? *An almost infinite number, so assumption is met.*

   (b) Representative sample?

   One spin should be like another, so I don't question this assumption.

(c) Independent trials?
If there is no cheating, this assumption is met.

(d) At least 10 successes? at least 10 failures?
Not met. 27 "Failures" and only 3 "Successes".

3. In the Unit 1 Review you estimated the probability a kissing couple leans to the right from data in which 80 of 124 couples did lean to the right. Let's check assumptions.

- Is population size at least $10n$ ($n = 124$)?
All couples, so assumption is met.
- Representative sample?
Sort of random observations, so I hope this assumption is met.
- Independent trials?
Couple should act independently, so this assumption is met.
- At least 10 successes? at least 10 failures?
Yes, $80 > 10$ *and* $44 > 10$

(a) Now we'll build a 99% confidence interval for the true proportion of couples leaning right when kissing using normal procedures.

i. What is the sample statistic?
0.645

ii. What is the standard error of $\widehat{p}$ for these data?
$\sqrt{0.645 * 0.355/124} = 0.0430$

iii. From the table about two pages back, what z^* values goes with 99% confidence?
2.576

(b) Build the interval and interpret its meaning.
$(0.534, 0.756)$. *We are 99% confident that the true proportion of couples who lean right when kissing is between 53.4 and 75.6%.*

Hypothesis testing

Reminder: when we do hypothesis testing, we give the benefit of the doubt to: ____________.

Our assumption of "innocent until proven guilty" changes the formula for standard error. We plug in the null hypothesis value instead of the sample proportion, so when hypothesis testing: $SE(\widehat{p}) = \sqrt{p_0(1-p_0)/n}$. Secondly, instead of counting points as or more extreme than the observed statistic, we will use the normal distribution to compute the probabilities. To do that, we need to convert the distance between p_0 and $\widehat{p}$ to "standard deviation" units by dividing by $SE(\widehat{p})$. The complete formula is:

$$z = \frac{\widehat{p} - p_0}{SE(\widehat{p})} = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

We will illustrate with the "Kissing on the Right" data. As we did before, we will not assume which direction the alternative will take, but the null is that couples are equally likely to lean left or right.

4. State null and alternative hypotheses in symbols and words.
   $H_0 : p = .5$. *Half of all couples lean right when kissing.* H_a :

   $p \neq .5$. *The true proportion of couples leaning right when kissing is not one half.*

5. Compute the $SE(\widehat{p})$ using the null hypothesis value for p.

   $SE(\widehat{p}) = \sqrt{.5(.5)/124} = 0.0449$

6. Build the z statistic by dividing the difference between $\widehat{p}$ and p_0 by $SE(\widehat{p})$.

   $z = \frac{.645-.5}{0.0499} = 3.23$

7. Put the standardized z value into the web app `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`, One Categ – Normal Distribution and ask for the Extremes. What part of the plot is shaded? What is the (total) p-value and how strong is this evidence against H_0?

   $2 \times .001 = .002$ *This is very strong evidence against the null.*

8. Report the strength of evidence. At the $\alpha = .01$ level of significance, what do you decide to do with H_0? What is your conclusion?

   We reject H_0 at the 1% significance level. We conclude that the true proportion of couples who lean right when kissing is over 0.50.

Rock, Paper, Scissors

A student played a game of "Rock, Paper, Scissors" with each of 120 inexperienced players and found that 55 of his opponents first chose "Rock". We want to test to see if the three options are equally likely and to build a confidence interval for the true proportion to pick "Rock".

9. Check the assumptions.

   There is a huge population of potential first time players, so we have less than 10% of the population. We'd like to know how the opponents were selected. We can't assume they are representative, and if they watched some of his earlier games, they might be influenced in their choices. We cannot assume independence from the data provided. We do have large enough sample size to use normality because $55 > 10$ *and* $120 - 55 = 65 > 0$ *(for CI) and* $np_0 = 120 \times 1/3 = 40 > 10$ *(for hypothesis test).*

10. Although you should have identified some possible problems with the assumptions, we will go ahead with the normal theory based inference. Compute $\widehat{p}$ and its SE for a confidence interval.

    $\widehat{p} = 55/120 = 0.458$ *with SE* $= \sqrt{.458(.542)/120} = 0.0455$

11. Build a 90% confidence interval for the true mean proportion of first time players who pick "Rock".

    $0.458 \pm 1.645 \times 0.0455 = (0.383, 0.533)$

12. Now switch to thinking of a hypothesis test using "random guessing" as the null model. What are the null and alternative hypotheses?

    $H_0 : p = 1/3$, *versus* $H_a : p \neq 1/3$

13. Compute the SE under H_0 and the test statistic.

    $SE = \sqrt{1/3 \times 2/3 \times 1/120} = 0.0430$, $z = \frac{0.455 - 0.333}{0.043} = 2.84$

14. What is the strength of evidence (from the web app `http://shiny.math.montana.edu/prob`) against H_0?

    Very strong. The p-value is 2×0.002.

15. Write a short report on the hypothesis test. Include the 5 main points:
    Type of test, null hypothesis, observed result, strength of evidence, and scope of inference.

Take Home Message

- To use any statistical method, our assumptions must be met. We need representative samples, independent trials, and a sample size less than one tenth of the population size. To use the normal probabilities we also need at least ten successes and ten failures (for a CI) or $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.
- The only change to our writeups is that we describe the test as a "z" test instead of a permutation test, and the confidence interval is also "z based".
- Write your questions and summary here.

Smell of Baking Bread

Are we better people in a pleasant environment?

A study from the *Journal of Social Psychology*[7] reports on a study of people's behavior under two different conditions. The researchers gave this description of their methods:

> "The participants were 200 men and 200 women (between the ages of approximately 20 and 50) chosen at random while they were walking in a large shopping mall. The participant was tested while walking near areas containing pleasant ambient odors (e.g.: bakeries, pastries) or not (e.g. clothing stores). Four young women (M = 20.3 years) and four young men (M = 21.3 years) served as confederates in this study. They were dressed in clothing typically worn by people of this age (jeans/T-shirt/boat shoes). The confederate chose a participant walking in his/her direction while standing in front of a store apparently looking for something in his/her bag. The confederate was carefully instructed to approach men and women walking alone, apparently aged from 20 to 50, and to avoid children, adolescent, and elderly people. The confederate was also instructed to avoid people who stopped near a store. Once a participant was identified, the confederate began walking in the same direction as the participant about three meters ahead. The confederate held a handbag and accidentally lost a glove. The confederate continued, apparently not aware of his/her loss. Two observers placed approximately 50 meters ahead noted the reaction of the passer-by, his/her gender, and estimated, approximately, his/her age. Responses were recorded if the subject warned the confederate within 10 seconds after losing the object. If not, the confederate acted as if he/she was searching for something in his/her hand-bag, looked around in surprise, and returned to pick up the object without looking at the participant."[8]

Assume that when the confederate saw a person who fit the qualifications, a coin was flipped. If it came up Heads, the subject was picked to be in the study, if Tails, they were skipped.

Be prepared to answer questions about

1. The questions researchers wanted to answer.

   Does the smell of baking bread influence people to be more altruistic?

2. What were the subjects?

   400 people in a mall

3. Were treatments applied at random?

   Sort of

[7] Nicolas Guéguen, 2012. The Sweet Smell of . . . Implicit Helping: Effects of Pleasant Ambient Fragrance on Spontaneous Help in Shopping Malls . *Journal of Social Psychology* **152**:4, 397-400

[8] ibid

Normal Inference - Difference in Proportions

When we did a hypothesis test to see if the difference in two true proportions was zero, for example when evaluating the effectiveness of peanut protein, we shuffled cards and relabeled the two groups many times. Now we'll use the normal distribution instead.

To make the switch from simulations to a theoretical distribution, we need, just as for a single proportion, a formula for the standard error of our statistic. In the last activity our statistic was $\widehat{p}$ and our formula was $SE(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$. To compare two groups, our statistic is $\widehat{p}_1 - \widehat{p}_2$ and we need a formula for $SE(\widehat{p}_1 - \widehat{p}_2)$. As with a single proportion, the form of this standard error depends on whether we are doing a hypothesis test (assuming some H_0 is true) or building a confidence interval. We'll start with the hypothesis test which is typically testing to see if the two groups have the same true proportion, that is: $H_0: \ p_1 = p_2$.

- If H_0 is true, the two groups are really the same, and we should combine them to get a better estimate of the overall proportion of successes. We'll call estimate $\widehat{p}_T$ where the 'T' stands for "**marginal**" because it's based on totals which appear in the outer margin of a table. We find it by combining all successes from both groups, then dividing by the sum of the sample sizes.

  $$\widehat{p}_T = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

  Recall: we used the same combined estimate when simulating draws from H_0 earlier.

The **hypothesis testing** formula for standard error of the difference in sample proportions is:

$$SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_T(1-\widehat{p}_T)}{n_1} + \frac{\widehat{p}_T(1-\widehat{p}_T)}{n_2}} = \sqrt{\widehat{p}_T(1-\widehat{p}_T)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Are we better people in a pleasant environment?

For class today, you read an article abstract about a study involving the smell of baking bread. Answer these questions about it:

1. Was a random mechanism used to select the person studied? Explain.
   Yes, if the coin flip was performed, it would randomly select about half of the people walking by.

2. What was the treatment and how was it applied to a subject?
   The "treatment" was the location: either near a bakery or near another type of store. It was "set", but not for each subject. Choice of treatment filtered out the potential subjects. It was not applied at random.

3. Does this study fit the definition of an experiment, or is it observational? Explain.
   I'd say it's observational, since a given passerby probably was a possible subject for just one of the treatments, not both.

4. Name three or more possible lurking variables.
   Reason for visiting the mall (need bread? or need clothing?). Gender. Socio-economic status. Tendency to lose things.

5. What is the scope of inference for this study?
   We can only infer association within this sample because the subjects where only haphazardly selected, and treatments were not randomly applied.

6. What are null and alternative hypotheses for this study? Assume that researchers were not willing to state ahead of time whether a good smell makes people "better" or "worse".
   $H_0 : p_1 = p_2$
   $H_a : p_1 \neq p_2$

   Check you answers just above with other groups at your table. Do we all agree about the direction of the alternative?

7. Compute the following proportions:
   Bakery group: $\widehat{p}_1 = 154/200 = 0.752$
   Clothing store: $\widehat{p}_2 = 105/200 = 0.525$
   Overall: $\widehat{p}_T = 259/400 = 0.648$

8. When testing one proportion, we created a z statistic with $z = \frac{\widehat{p}-p_0}{SE(\widehat{p})}$. In general, we use

   $$z = \frac{\text{statistic - null value}}{SE(\text{statistic})}$$

   Now our statistic is $\widehat{p}_1 - \widehat{p}_2$.

   - What value do we expect it to have if H_0 is true? 0
   - What is the standard error of the statistic under H_0?
     $SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{0.648 \times 0.352(\frac{1}{200} + \frac{1}{200})} = \sqrt{.002281} = 0.04776$
   - Compute our z statistic. $z = \frac{0.648-0.5025}{0.04776} = \frac{0.1455}{0.04776} = 3.047$

9. Use the web app to find the probability. How strong is the evidence against H_0?
   p-value $= 2 \times (0.001) = 0.002$ *This is very, very strong evidence refuting the null hypothesis that people act just as helpful in the two situations. In fact, the people close to the bakery were much more helpful than those by the clothing store.*

10. Write up the results as a statistical report on your own paper. (Suggestion: finish the activity, then come back to this.)

Confidence Interval for the Difference in True Proportions

Next we want to get an interval estimate of the difference in true proportions. We'll use the same data to ask: "How much more helpful are people near a bakery than near a clothing store?"

Again, looking back to a single proportion we used $\widehat{p} \pm z^* SE(\widehat{p})$ which is a special case of the general rule:

$$\text{estimate} \pm \text{multiplier} \times SE(\text{estimate})$$

All we need to do is to find the $SE(\widehat{p}_1 - \widehat{p}_2)$. We **do not assume the two are equal**, so no $\widehat{p}_T$ is needed. The formula is:

$$SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

You might ask why there is a plus sign between the two terms inside the square root, but a minus sign in the estimator. It's because each sample proportion has some variability, and $\widehat{p}_1 - \widehat{p}_2$ can vary due to changes in $\widehat{p}_1$ or in $\widehat{p}_2$. Subtracting would imply that having a second sample makes the difference **less** variable, when really it makes our statistic **more** variable.

OK, we are now ready to build a confidence interval.

12. Use $\widehat{p}_1$ and $\widehat{p}_2$ to compute the standard error of the difference in sample proportions. $SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{0.648 \times (1-0.648)/200 + 0.5025 \times (1-0.5025)/200} = \sqrt{0.0011405 + .0012500} = \sqrt{0.002391} = 0.04889$
13. In this case, a 90% confidence interval is needed. Refer back to your table from last class, or use the web app to find z^*. Find the margin of error and build the interval. $\text{ME} = 1.645 \times 0.04889 = 0.0804$ $90\%\ \text{CI} = 0.648 - 0.5025 \pm 0.0804 = 0.1455 \pm 0.0804 = (0.065, 0.226)$
14. Interpret the CI in the context of this research question. *We are 90% confident that the true proportion of helpful people near a bakery is 6.5 to 22.6% higher than near a clothing store.*

Assumptions?

We need basically the same assumptions when working with two samples as with one sample proportion. The first three apply to any method of doing hypothesis tests or confidence intervals. The last is particular for normal-theory based methods with proportions.

- The size of each sample must be less than a tenth the size of its population.
- Each sample must be representative of its population.
- **Independent** responses. Definition: responses are independent if knowing one response does not help us guess the value of the other. Sampling multiple people from the same household gives **dependent** responses.
- To use normality for a confidence interval: at least 10 successes and 10 failures in each group. To use normality for hypothesis testing, take the smaller of n_1 and n_2 times the smaller of $\widehat{p}_T$ or $(1 - \widehat{p}_T)$ and this value should be at least 5.

Take Home Messages

- To do hypothesis testing we needed the "overall" estimated success proportion – forgetting about the two groups.
- Our estimate of spread, the SE, changes depending on whether we assumed $p_1 = p_2$, as in hypothesis testing, or not (confidence intervals). Know both versions.
- The general form of a standardized statistic for hypothesis testing is:

  $$z = \frac{\text{statistic - null value}}{SE(\text{statistic})}$$

  in this case that is

  $$z = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{SE(\widehat{p}_1 - \widehat{p}_2)} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_T(1-\widehat{p}_T)}{n_1} + \frac{\widehat{p}_T(1-\widehat{p}_T)}{n_2}}}$$

- To build a confidence interval, we do not assume $p_1 = p_2$.
- The general form of a CI is

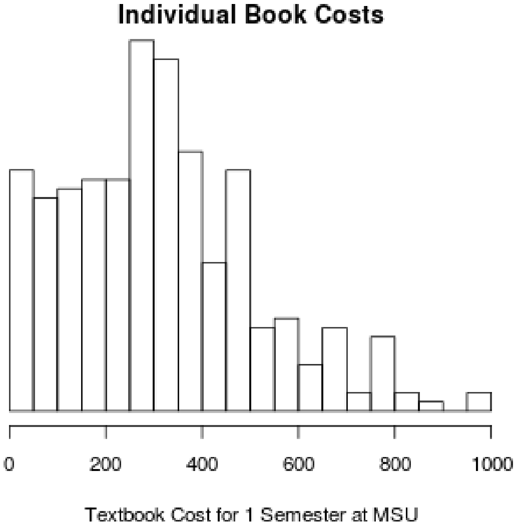
  $$\text{estimate} \pm \text{multiplier} \times SE(\text{estimate})$$

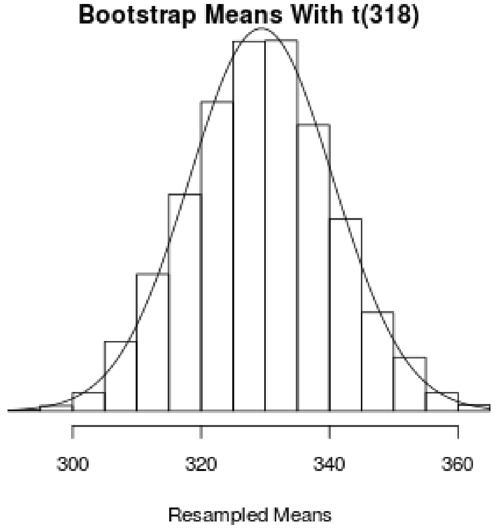
  which in this case is

  $$\widehat{p}_1 - \widehat{p}_2 \pm z^* SE(\widehat{p}_1 - \widehat{p}_2) = \widehat{p}_1 - \widehat{p}_2 \pm z^* \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

t-Distributions - Inference for One Mean

In a previous semester, we asked STAT216 students to estimate how much they were spending on textbooks during that semester. We had sample size 319 with mean $\overline{x} = 329.44$ and spread $s = 202.93$ (both in dollars). Here is a histogram of the individual data and another of 5000 resampled means.





The second plot is overlaid with a t_{318} density curve. Note how it gives a very similar distribution to the bootstrap resampled means, even though the original data is not symmetric.

We needed many bootstrap samples to get the "standard deviation" of the resampling distribution. Now we will use a formula for the standard error (SE) of the sample mean, $\overline{x}$ based on sample size and the standard deviation of the raw data.

$$SE(\overline{x}) = \frac{s}{\sqrt{n}}$$

In the left-hand plot above, the spread of the individual points is $s = 202.93$. The resampled means have spread of 11.20 which is quite close to $SE(\overline{x}) = 202.93/\sqrt{319} = 11.36$.

- **Standard Deviation** has several meanings:
  - the spread of some distribution, especially: the spread of individual measurements.
  - Population standard deviation with Greek letter: σ,
  - Sample standard deviation, s from the formula: $s = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$
  - Standard deviation of a statistic, for instance $SD(\overline{x}) = \sigma/\sqrt{n}$, is the true spread of the statistic.
- **Standard error** is the estimated standard deviation of a statistic. Below is a table of the standard deviations and standard errors we are using. Note how the SE just plugs an estimated value in to the SD formula.

Statistic:	Standard Deviation	Standard Error
$\widehat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
$\widehat{p}_1 - \widehat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$
$\overline{x}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
$\overline{x}_1 - \overline{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Each of these SD's and SE's has sample size (square rooted) in the denominator. As sample size gets big, SD and SE get smaller. More information leads to less variation.

Add to this the fact that sample mean is an **unbiased** estimator of the true population mean, and sample proportion is an **unbiased** estimator of the true population proportion, and we see the power of using statistics to estimate parameters: larger sample sizes provide more information, less variation, and we can close in on the true parameter values. Here are two big ideas in statistics:

Law of Large Numbers

In the long run, the sample mean $\overline{x}$ will get closer and closer to the population mean, μ, and the sample proportion $\widehat{p}$ will get closer and closer to the true proportion of successes, p. You can define "close" as you wish – the more precision needed, the greater the sample size.

The **LLN** addresses the width, or spread of the sampling distribution. The second important idea addresses the **shape** of the sampling distribution, and was more of a surprise when it was discovered about 200 years ago (and proven, in general only 100 years ago).

Central Limit Theorem

As the sample size gets big, the shape of the sampling distribution of sample mean, $\overline{x}$ gets closer and closer to a normal distribution.

The **CLT** explains why the normal distribution is so useful. Even if we start with skewed data, like in the case of book costs, when we average together 100 or more random values, the distribution of the sample mean, $\overline{x}$, will approach a normal distribution. It also applies to proportions because if we record success as 1 and failure as 0, then $\widehat{p}$ is just the sample mean of the zeroes and ones.

Assumptions for t-Procedures

Just as with proportions, all means methods require

- The size of each sample must be less than a tenth the size of its population.
- Each sample must be representative of its population.
- Independent samples and independent subjects within each sample.

In addition, we do need to examine the sample size **and** shape of the sample data to determine if we can use t procedures.

- For small sample sizes, we need distributions to be close to normally distributed: symmetric with no outliers.
- If sample size is 30 or more, we can use t procedures unless the data are heavily skewed.
- If sample size is over 100, the Central Limit Theorem lets us use t distributions even for heavily skewed data.

True or False:

1. ______ The Law of Large Numbers says that the distribution of $\overline{x}$ gets close to normal as n gets big.
2. ______ As degrees of freedom get big, t-distributions get closer and closer to the standard normal distribution.
3. ______ The Central Limit Theorem gives us the shape of the distribution of $\overline{x}$ for large n.
4. ______ With larger sample size, we have better information about population parameters.
5. ______ Statistics from larger samples are less biased than those from smaller samples.

Confidence Interval for μ

With text book costs, we had:

$$n = 319, \qquad \overline{x} = 329.44, \text{ and } \qquad s = 202.93$$

With one sample, we use $n-1$ as the "degrees of freedom" for the t distribution.

6. In the web app `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps` select $\boxed{\text{t distribution}}$ under $\boxed{\text{One Quant}}$. Change the degrees of freedom to $\boxed{318}$, then enter the probabilities for the $\boxed{\text{center}}$ to find the t^*_{318} multipliers to use in the formula for confidence interval:

$$\text{estimate} \pm t^*_{df} SE(\text{estimate})$$

Confidence level:	80%	90%	95%	98%	99%
t^*_{318} cutoff	1.284	1.65	1.967	2.338	2.591

7. How different are these from the values you found for the z distribution last week?
Just a touch wider
8. Find the margin of error and build a 90% confidence interval for the true book cost.
$\text{ME} = 1.65{\times}202.93\sqrt{319} = 18.747$ the 90% CI is $329.44{\pm}18.747 = (310.69, 348.19)$
9. Interpret the interval.
We are 90% confident that the true mean amount spent on textbooks by some group of students is between \$310.69 and \$348.19.

10. To what group of students does this inference apply?
    Trick question: This was not a random sample of MSU students. It really just applies to the students in the sample.

Hypothesis Test

We do not have a hypothesized value to use for true mean textbook cost, so let's look at a different situation to do a hypothesis test on one mean. An article in the *Journal of American Medical Association* in 1992 provided data to challenge the long held belief that "normal" human body temperature is 98.6^{o}F.[9] Research before this study was published led the researchers to believe that "normal" for most people is lower than 98.6^o.

6. What are the null and alternative hypotheses?
   $H_0: \quad \mu = 98.6$
   $H_a: \quad \mu < 98.6$

   Check these with another group, because being off on direction will mess us up later.

7. Go to the same web app we've been using and choose the pre-loaded $\boxed{\text{bodytemp}}$ data under $\boxed{\text{One Quant}}$ which contains these values:

   ```
   temp
   97.3 97.3 97.7 97.8 98.4 99.8 96.7 98.1 98.7 97.5
   97.9 98.1 97.8 98.5 98.8 98.7 99.4 97.8 98.6 98.7
   ```

   Obtain the mean and standard deviation (s). Use 3 decimal place accuracy throughout.
   $\overline{x} = 98.18$, $s = 0.747$

8. Compute $SE(\overline{x})$.
   $0.747/\sqrt{20} = 0.167$

9. How many standard errors is $\overline{x}$ from μ_0? Compute the test statistic.
   $\frac{98.180-98.6}{0.167} = -2.514$

10. Which t distribution should we use?
    t with 19 df

11. Pull up the t-distribution web app and put the t-statistic into the top box and set the degrees of freedom you found just above. Check the direction of your alternative hypothesis and give the p-value for the test.
    0.011

12. How strong is the evidence against H_0? State your conclusion at the $\alpha = 0.04$ significance level.
    Very strong. We reject H_0 and conclude that true mean "normal" body temperature is less than 98.6^o *F*

[9] Mackowiak, P. A., Wasserman, S. S., & Levine, M. M. (1992). A critical appraisal of 98.6 F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *JAMA*, 268(12), 1578-1580.

13. Based on your p-value, will a 96% confidence interval for true mean body temperature contain 98.6?
No. It is not a plausible value based on this sample.

14. Build the 96% CI. Be sure to show which t^* you use. $t^*_{19} = 2.2047$, *96% CI* $= 98.18 \pm 2.2047 \times 0.167 = (97.81, 98.55)$

15. Write up a report on your hypothesis test of "Normal" body temperature. You may assume that these temperatures are from a random sample of US men and women aged 20 to 40 years old. (With one only group we do not make comparisons in the report.)

Take Home Message

- Larger sample sizes are generally better than smaller ones – as long as they are representative. If we are using a biased sampling method, no amount of increasing sample size will fix the bias problem.
- As sample size, n gets big:
  - statistics get closer to their respective parameters.
  - distributions of means get closer to normal in shape.
  - t-distributions get closer to Z (N(0,1)) because degrees of freedom are related to sample size.
- Generic confidence interval:

$$\text{estimate} \pm \text{multiplier} SE(\text{estimate})$$

  for a single mean:

$$\overline{x} \pm t^*_{n-1} SE(\overline{x})$$

- Test statistic:

$$t = \frac{\overline{x} - \mu_0}{SE(\overline{x})}; \quad SE(\overline{x}) = \frac{s}{\sqrt{n}}$$

  Compare to a t_{n-1} distribution to get p-values.
- What questions do you have? Write them here.

More Energy Drinks

Back on Activity 14 we compared an energy drink with alcohol, REDA, to a control. Our conclusion that "change in RBANS" was lower in the REDA group did not allow us to say whether that was due to the alcohol or the stimulant in REDA. The two explanatory variables were "confounded" meaning that we can't separate the effects. The researchers knew this would be a problem, so they included a third group in the study: 9 randomly selected women got RED, an energy drink with no alcohol. Today we'll compare RED and control means using confidence intervals and hypothesis tests based on a t distribution.

The research question is:

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

Higher RBANs scores indicate better memory skills.

Go to the usual website and use the pre-loaded `REDvsControl` dataset under One of Each.

1. Obtain "Actual" Mean and Std dev for each group.

   Means: control: 1.219, RED: -2.44
   Std.dev.: control: 3.919, RED: 6.425

2. Using t-based methods requires that we either have "near normal" data or large sample sizes. Do you see extreme outliers in the plots? Are the data skewed?

   These look OK to me, though the spreads do not look equal.

We'll start by comparing RED to Control using a hypothesis test.

3. What are the null and alternative hypotheses when comparing RED to Control? Use notation, and write them out with words. Do not assume they knew ahead of time which mean would be larger.

   $H_0: \ \mu_1 = \mu_2$ *The mean change in RBANS is the same for treatment (RED) and Control.*

   $H_a: \mu_1 \neq \mu_2$ *The mean change in RBANS is not the same for treatment (RED) and control.*

   Check with another group to be sure we have the correct direction for H_a.

4. Let group 1 be Control and group 2 be RED. Compute $SE(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

   $\sqrt{\frac{3.92^2}{9} + \frac{6.43^2}{9}} = \sqrt{1.707 + 4.594} = 2.5102.$

5. As with a single mean, our test statistic is called t and is found by dividing the difference between sample statistic and its value under H_0 by its SE. Find t.

   $t = \frac{\overline{x}_1 - \overline{x}_2 - \mu_1 - \mu_2}{SE(\overline{x}_1 - \overline{x}_2)} = \frac{1.219-(-2.44)}{2.5102} = 1.458$

6. We need to figure out which degrees of freedom for our t test. When working with two means, we take the smaller sample size and subtract one from it. What is that for the Energy Drinks study? (This is a conservative approach. Other books or software programs might use a different number.)

   8

7. Go to the t distribution part of the applet and put your t test statistic in the top box, set its degrees of freedom, and select the right direction for the alternative. What is our p-value? State your decision and conclusion about these two drinks.

   0.182. At any of the commonly used α levels, we fail to reject the null hypothesis that mean change in RBANS score is the same for control and RED groups. There is not sufficient evidence to say that they differ.

8. Next we want to compare RED drinkers to REDA drinkers (group 2 to group 3).

   (a) Your null and alternative hypotheses are the same, but we need to use subscripts 2 and 3 instead of 1 and 2. Copy them here.
   $H_0: \mu_2 = \mu_3$ *The mean change in RBANS is the same for RED and REDA.*
   $H_a: \mu_2 \neq \mu_3$ *The mean change in RBANS is not the same for RED and REDA.*

   (b) Compute the difference in means.
   $-2.44 - (-8.329) = 5.889$

   (c) Compute the SE of the estimator.
   $\sqrt{\frac{6.43^2}{9} + \frac{10.02^2}{9}} = \sqrt{4.594 + 11.1566} = 3.969$

   (d) Compute the t test statistic.
   $5.889/3.969 = 1.484$

   (e) What degrees of freedom will you use? Find the p-value.
   8 df. The p-value is $2 \times 0.088 = 0.172$

   (f) What is your decision? your conclusion?
   Again, we do not have much evidence against the null hypothesis that the two groups have the same mean. We fail to reject H_0 at all commonly used α levels.

Confidence Interval

Reminder. The general form of a t-based CI is:

$$\text{estimate} \pm t^*_{df} \times SE(\text{estimate})$$

9. Finally, we'll go back to the original comparison between REDA and Control means. However, we'll set this one up as a confidence interval instead of a t-test.

   (a) Compute the difference in means between REDA and control.
   $1.219 - (-8.329) = 9.548$

   (b) Compute $SE(\overline{x}_1 - \overline{x}_3)$
   $\sqrt{\frac{3.919^2}{9} + \frac{10.02^2}{9}} = 3.587$

   (c) What degrees of freedom do we need?
   Use the web app to find the t^* multiplier for a 95% confidence interval.
   8 df. $t^* = 2.306$
   8 df. $t^* = 2.306$

(d) Find the margin of error and construct the 95% CI. Does it contain zero? If testing $H_0 : \mu_1 = \mu_3$ versus a two-sided alternative at the $\alpha = 0.05$ level, would you reject H_0? Explain.
ME $= 2.306 \times 3.587 = 8.272$, *95% CI:* $9.548 \pm 8.272 = (1.28, 17.82)$ *Zero is not in the interval, so it is not a "plausible value" for the difference in means. We would reject the null in favor of the alternative at* $\alpha = 0.05$.

10. How do you explain the fact that two of the three t-tests we've done gave large p-values and another gave a small p-value? Is that inconsistent?

    The confidence interval shows that we have fairly strong evidence that the mean for control change in RBANs is larger than the mean for REDA change in RBANS. We do not have very large sample sizes in this study, so it is not surprising that two comparisons found that the middle mean (RED) was not very different from either extreme (Control or REDA), yet the two extreme values were far enough apart to detect a fairly strong difference.

11. Could the difference in REDA and Control means be just due to random chance? Explain.

    Yes. Our conclusions do not change just because we've used a different method to compute the p-value. It's still possible that one group was higher just due to random assignment of treatment.

    Yes. Our conclusions do not change just because we've used a different method to compute the p-value. It's still possible that one group was higher just due to random assignment of treatment.

12. Can we make causal inference about the effects of energy drinks and alcohol?

    Yes. Because treatments were randomly assigned, it is very unlikely that we would see such large differences just by chance, so the observed differences (or lack thereof) are attributable to the treatments.

13. Write up the hypothesis test results as a report. Include all three comparisons we've made.

Take Home Message

- Interpretation of results does not change just because we switched from permutation testing to t-tests. We still ask "Was this a random sample of subjects?" to obtain inference back to a population, and we still ask "Were treatments assigned at random?" to conclude that change in one variable caused changes in the other.
- The t-test approach uses formulas for standard errors, while the permutation test relies on repeated permutations or resampling to get the same information about the spread of the sampling distribution under H_0. Both work!
- We can use t-tests to compare two means, but do have to be careful with small sample sizes. We should not use t-tests with sample sizes less than 15 unless the data look symmetric with no outliers.
- With large sample sizes, t methods are robust to outliers and skewness.
- When sample sizes are small, only *large* differences will be flagged as having "strong evidence". We'll look at this later in more detail.
- What questions do you have? Write them here.

Concussions in the News

1. Read the abstract from this article.

   Petraglia AL, Plog BA, Dayawansa S, Chen M, Dashnaw ML, Czerniecka K, Walker CT, Viterise T, Hyrien O, Iliff JJ, Deane R, Nedergaard M, Huang JH. (2014). "The spectrum of neurobehavioral consequences after repetitive mild traumatic brain injury: a novel mouse model of chronic traumatic encephalopathy." **J Neurotrauma** Jul 1; **31**(13):1211-24. `http://www.ncbi.nlm.nih.gov/pubmed/24766454`

   > Abstract
   > There has been an increased focus on the neurological consequences of repetitive mild traumatic brain injury (TBI), particularly neurodegenerative syndromes, such as chronic traumatic encephalopathy (CTE); however, no animal model exists that captures the behavioral spectrum of this phenomenon. We sought to develop an animal model of CTE. Our novel model is a modification and fusion of two of the most popular models of TBI and allows for controlled closed-head impacts to unanesthetized mice. Two-hundred and eighty 12-week-old mice were divided into control, single mild TBI (mTBI), and repetitive mTBI groups. Repetitive mTBI mice received six concussive impacts daily for 7 days. Behavior was assessed at various time points. Neurological Severity Score (NSS) was computed and vestibulomotor function tested with the wire grip test (WGT). Cognitive function was assessed with the Morris water maze (MWM), anxiety/risk-taking behavior with the elevated plus maze, and depression-like behavior with the forced swim/tail suspension tests. Sleep electroencephalogram/electromyography studies were performed at 1 month. NSS was elevated, compared to controls, in both TBI groups and improved over time. Repetitive mTBI mice demonstrated transient vestibulomotor deficits on WGT. Repetitive mTBI mice also demonstrated deficits in MWM testing. Both mTBI groups demonstrated increased anxiety at 2 weeks, but repetitive mTBI mice developed increased risk-taking behaviors at 1 month that persist at 6 months. Repetitive mTBI mice exhibit depression-like behavior at 1 month. Both groups demonstrate sleep disturbances. We describe the neurological consequenses of repetitive mTBI in a novel mouse model, which resemble several of the neuropsychiatric behaviors observed clinically in patients sustaining repetitive mild head injury.

   Be prepared to answer questions about

   (a) The questions researchers wanted to answer.
   What happens to mice after repeated brain injury?

   (b) Who/what were the subjects?
   mice

   (c) Were treatments applied at random?
   Yes

2. And read this abstract:

   Lin, Ramadan, Stern, Box, Nowinski, Ross, Mountford. (2015). "Changes in the neurochemistry of athletes with repetitive brain trauma: preliminary results using localized correlated spectroscopy." **Alzheimers Research & Therapy**. 2015 Mar 15;7(1):13 `http://www.ncbi.nlm.nih.gov/pubmed/25780390`

   > Abstract
   > INTRODUCTION:
   > The goal was to identify which neurochemicals differ in professional athletes with repetitive brain trauma (RBT) when compared to healthy controls using a relatively new technology, in vivo Localized COrrelated SpectroscopY (L-COSY).

> METHODS:
> To achieve this, L-COSY was used to examine five former professional male athletes with 11 to 28 years of exposure to contact sports. Each athlete who had had multiple symptomatic concussions and repetitive sub concussive trauma during their career was assessed by an experienced neuropsychologist. All athletes had clinical symptoms including headaches, memory loss, confusion, impaired judgment, impulse control problems, aggression, and depression. Five healthy men, age and weight matched to the athlete cohort and with no history of brain trauma, were recruited as controls. Data were collected from the posterior cingulate gyrus using a 3 T clinical magnetic resonance scanner equipped with a 32 channel head coil.
> RESULTS:
> The variation of the method was calculated by repeated examination of a healthy control and phantom and found to be 10% and 5%, respectively, or less. The L-COSY measured large and statistically significant differences (P <=0.05), between healthy controls and those athletes with RBT. Men with RBT showed higher levels of glutamine/glutamate (31%), choline (65%), fucosylated molecules (60%) and phenylalanine (46%). The results were evaluated and the sample size of five found to achieve a significance level P=0.05. Differences in N-acetyl aspartate and myoinositol between RBT and controls were small and were not statistically significance.
> CONCLUSIONS:
> A study of a small cohort of professional athletes, with a history of RBT and symptoms of chronic traumatic encephalopathy when compared with healthy controls using 2D L-COSY, showed elevations in brain glutamate/glutamine and choline as recorded previously for early traumatic brain injury. For the first time increases in phenylalanine and fucose are recorded in the brains of athletes with RBT. Larger studies utilizing the L-COSY method may offer an in-life method of diagnosis and personalized approach for monitoring the acute effects of mild traumatic brain injury and the chronic effects of RBT.

Be prepared to answer questions about

(a) The questions researchers wanted to answer.
Do brain chemicals differ in athletes exposed to brain trauma?

(b) Who/what were the subjects?
5 athletes and 5 non-athlete men.

(c) Were treatments applied at random?
No

3. Here's a quote from a news report in 2013:

> "We need to figure out what's making some people more vulnerable than others," says Michelle Mielke, an Alzheimer's researcher at the Mayo Clinic who led the study. It was published online Thursday in the journal Neurology.
>
> "Just because you have a head trauma doesn't mean you're going to develop memory problems or significant amyloid levels," Mielke told Shots. And it doesn't mean you're going to get Alzheimer's. "But it does suggest to us that there's a mechanism with head trauma that does increase your risk."
>
> Mielke and her colleagues did PET scans on 589 people who are participating in a long-term study of aging and memory. That's a lot of people for a brain imaging study, which makes it more likely that the findings are accurate.

`http://www.npr.org/blogs/health/2013/12/27/257552665/concussions-may-increase-alzheimers-risk-but-only-for-some`

Be prepared to answer questions about

(a) The questions researchers wanted to answer.
Are athletes exposed to brain trauma more at risk of Altzeimer's disease?

(b) What were the subjects?
589 people involved in a long term brain study

(c) Were treatments applied at random?
No

Concussions in the News

- Chris Borland, age 24, retired from the 49ers after one year of NFL play, because he feared his brain might be permanently injured by concussions.
- In 2013 the NFL agreed to pay $765 million to fund exams, research and for compensation of former players.
- High school coaches are now advised to make their players sit out if they sustain a blow to the head.
- The Congressional Brain Injury Task Force is looking into the prevalence of brain trauma in troops deployed in recent wars and at how VA hospitals are treating the condition.

Applying What We've Learned

Some have argued that blows to the head – for instance on the high school football field – are getting too much attention. They point out that of many kids taking similar hits, only a few show lasting decrease in cognitive abilities, and suggest that the effects are as much due to genetics as to the blow to the head.

1. Your group is asked to design a study to determine how large a risk the "hits" taken by high school athletes are to their future cognitive abilities. Write down a plan for your study. Include:
    - Who will your subjects be? How will you find them? How many will you need?
    - What will you measure? (Brain scans like MRI? Cognitive tests of memory and reasoning? Some measure of emotional states like anger?) Include the timing of measurements.
    - Will your study be observational or an experiment?
    - Is your study ethical in its treatment of subjects?

2. Trade papers with another group and read their study over carefully. Address the bullet points above (gently – remember someone is doing this to your proposal as well). Provide suggestions as to what might be improved.
    - Subjects?
    - Measurements?
    - Observational? Experiment?
    - Ethical issues?

3. You should have read the abstract of this article for today:

   Petraglia AL, Plog BA, Dayawansa S, Chen M, Dashnaw ML, Czerniecka K, Walker CT, Viterise T, Hyrien O, Iliff JJ, Deane R, Nedergaard M, Huang JH. (2014). "The spectrum of neurobehavioral consequences after repetitive mild traumatic brain injury: a novel mouse model of chronic traumatic encephalopathy." **J Neurotrauma** Jul 1; **31**(13):1211-24. `http://www.ncbi.nlm.nih.gov/pubmed/24766454`

   (a) What are the advantages of this study design?
   Replication, control, random assignment, very similar subjects.

   (b) Disadvantages?
   Does not extend to humans directly – only if we assume some linkage.

   (c) Does the design allow us to make causal inferences?
   Yes

   (d) Inferences to high school students?

4. And the abstract for this article:

   Lin, Ramadan, Stern, Box, Nowinski, Ross, Mountford. (2015). "Changes in the neurochemistry of athletes with repetitive brain trauma: preliminary results using localized correlated spectroscopy." **Alzheimers Research & Therapy**. 2015 Mar 15;7(1):13 `http://www.ncbi.nlm.nih.gov/pubmed/25780390`

   (a) What are the advantages of this study design?
   Ethically– they didn't hurt anyone.

   (b) Disadvantages?
   No random assignment. Small samples.

   (c) Does the design allow us to make causal inferences?

   (d) Inferences to high school students?

5. And this quote:

   > Mielke and her colleagues did PET scans on 589 people who are participating in a long-term study of aging and memory. That's a lot of people for a brain imaging study, which makes it more likely that the findings are accurate.

   `http://www.npr.org/blogs/health/2013/12/27/257552665/concussions-may-increase-alzheimers-risk-but-only-for-some`

   Explain the last sentence. In what sense does a large sample increase accuracy?

   Larger sample sizes give smaller variance estimates to $\overline{x}$ and $|phat$. That tends to shrink p–values and increase power, so we are less likely to make Type I or Type II errors. Larger samples do not decrease bias, so we'd still like to have random samples from the population.

Take Home Message

- In reading a research article, carefully read what they say about selection of subjects (or units of study) so we know how far we can extend our inference.
- Similarly, evaluate the assignment of any treatments (at random, we hope) so that we know if causal inference is appropriate.

- In designing a study, we'd like to get a large sample, but expense might prevent that.
- You might see names of tests we've not covered, but the idea of a p-value is the same for any test, and the null hypothesis is generally "No change", or "No effect".
- What's not clear? Write your questions here.

Unit 3 Wrapup

Vocabulary

We have used these methods:

- Z-test for single proportion
- Z-based CI for one proportion
- Z-test for difference in two proportions
- Z-based CI for difference in two proportions
- t-test for a single mean
- t-based CI for a single mean
- t-test for difference in two means
- t-based CI for a difference in two means
- t-test for mean difference in paired data
- t-based CI for mean difference in paired data

Know when to use z versus t, and know what degrees of freedom to use with the t's. Know the assumptions needed to use these methods.

Important ideas:

- With random events we cannot predict a single outcome, but we can know patterns we see with many repetitions.
- Interpretation of confidence intervals – they work in the "long run".
- Strength of evidence in Hypothesis Tests
- Assumptions to use z procedures on proportions
- Assumptions to use t procedures on means

Practice Problems

1. What does the Law of Large Numbers (LLN) tell us about the distribution of the sample mean $\overline{x}$ as n gets big?
   It gets tighter and tighter around the true mean, μ.
2. What does the Central Limit Theorem (CLT) tell us about the distribution of the sample mean $\overline{x}$ as n gets big?

   It becomes normally distributed.
3. Which of LLN and the CLT apply to sample proportions as n gets big?
   Both, $\widehat{p}$ is just a mean of zero = failure and 1 = success random variates.
4. How do we determine what to use as degrees of freedom for a t distribution when we are conducting a one-sample t test?

   a two-sample t-test?

1 sample: $n - 1$, *2 sample:* $\min(n_1, n_2) - 1$.

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

5. Rating Chain Restaurants[10]

   The July 2006 issue of Consumer Reports included ratings of 103 chain restaurants. The ratings were based on surveys that Consumer Reports readers sent in after eating at one of the restaurants. The article said, "The survey is based on 148,599 visits to full-service restaurant chains between April 2004 and April 2005, and reflects the experiences of our readers, not necessarily those of the general population."

   (a) Do you think that the sample here was chosen randomly from the population of Consumer Report readers? Explain. *No, it relies on voluntary response.*

   (b) Why do the authors of the article make this disclaimer about not necessarily representing the general population? *Because Consumer Reports readers are different in some ways – perhaps education level, perhaps, from the general population.*

   (c) To what group of people would you feel comfortable generalizing the results of this study? Explain. *Readers of Consumer Reports with enough time to answer the survey.*

6. Emotional Support[11]

   Shere Hite undertook a study of womens attitudes toward relationships, love, and sex by distributing 100,000 questionnaires in womens groups. Of the 4500 women who returned the questionnaires, 96% said that they gave more emotional support than they received from their husbands or boyfriends.

   (a) Comment on whether Hite's sampling method is likely to be biased in a particular direction. Specifically, do you think that the 96% figure overestimates or underestimates the proportion who give more support in the population of all American women? (In a voluntary response survey, those who do respond tend to have stronger and generally more negative opinions.) *Again, it relies on voluntary response so it's likely to be biased against men.*

   (b) ABC News/Washington Post poll surveyed a random sample of 767 women, finding that 44% claimed to give more emotional support than they received. Which polls result do you think are more representative of the population of all American women? Explain. *The ABC poll because it used random selection.*

7. Balsa Wood

   Student researchers investigated whether balsa wood is less elastic after it has been immersed in water. They took 44 pieces of balsa wood and randomly

[10] Rossman, A. J., Chance, B. L., & Lock, R. H., (2009). *Workshop Statistics: Discovery with Data and Fathom* (3rd ed.). Emeryville, CA: Key College Publishing.

[11] Hite, S. (1976). *The Hite Report: A nationwide survey of female sexuality.* London: Bloomsbury.

assigned half to be immersed in water and the other half not to be immersed in water. They measured the elasticity by seeing how far (in inches) the piece of wood would project a dime into the air. Use the data file located on D2L.

(a) Before opening the data file, which applet should be used to create an interval estimate for the difference in elasticity (single proportion, single mean, two-sample proportion, two-sample mean)? Explain. *StatKey two-sample means*

(b) The observed difference in mean elasticity between the two groups is 4.16 inches. Explain why a confidence interval is a better summary of the data than just this difference in sample means. *The point estimate does not include any information about how good an estimate it is. The confidence interval has a length, so we can see that it is not exact, but has built in variability.*

(c) Produce a 95% bootstrap interval for estimating the actual size of the treatment effect of immersing balsa wood in water. Describe the process by which you produce this interval, and also interpret what the interval means in the context of this study.

8. MicroSort® Study

The Genetics and IVF Institute is currently studying methods to change the odds of having a girl or boy2. MicroSort® is a method used to sort sperm with X- and Y-chromosomes. The method is currently going through clinical trials. Women who plan to get pregnant and prefer to have a girl can go through a process called X-Sort®. As of 2008, 945 have participated and 879 have given birth to girls[12]

(a) Compute a 95% interval to estimate the percentage of girl births for women that undergo X-Sort® using a bootstrap method.

(b) Interpret your interval.

(c) Compute a 95% interval estimate using a theoretical distribution.

i. Should you use a z or a t multiplier?

ii. What is the multiplier that should be used?

iii. What is the standard error of the sample proportion?

iv. What is the margin of error of the interval estimate?

v. Give the interval estimate.

vi. Compare the interval estimate using the theoretical distribution to the interval estimate using Bootstrap method. Are they similar in center? Width?

(d) Suppose more data has been collected since 2008. If the number of women had increased to 3000 but the observed percent of girls stayed the same, what would you expect to happen to your interval? *It should get narrower, because SE will get smaller as n gets bigger.*

[12] Genetics & IVF Institute, (2011). MicroSort. Genetics & IVF Institute. Retrieved from `http://www.microsort.net/`.

(e) Test out your conjecture by creating a new interval using a sample size of 3000. Report your new interval estimate. Was your expectation in question 13 correct?
StatKey two-sample means

(f) How many resamples (trials) did you run in your bootstrap simulation?

(g) What is the difference between sample size and number of trials?

9. Anorexia nervosa is an eating disorder characterized by the irrational fear of gaining weight and a distorted body self-perception. Several psycho-therapy methods have been tested as ways to treat individuals suffering from anorexia. The data set available on D2L gives the results of a study using cognitive behavioral therapy (CBT) and the pre-and post-treatment weights of 29 patients, along with the change in weight. You can get the descriptive statistics you need from the Rossman-Chance web app.

   (a) Build a 99% confidence interval for the difference in true means. Should we use the single or two-sample mean formula for SE? Explain, then give the confidence interval.

   (b) Would a 95% confidence interval be wider or narrower? Explain. *Narrower. It has less probability in the center, so that moves the cutoffs in closer to 0.*

   (c) Based on the confidence intervals, do the patients seem to improve (where improvement is based on increasing weight)? What significance level is being used?

   (d) Using the theoretical distributions, do a hypothesis test to answer this question. Write the null and alternative hypotheses, calculate the standard error and the test statistic, and find the p-value. Does the hypothesis test agree with the results for the confidence interval? Explain.

10. As mentioned in the last class, sports related concussions are a big concern. One study investigated whether concussions are more common among male or female soccer players. The study took a random sample of college soccer players from 1997 1999. Of 75,082 exposures (practices or games) for female soccer players, 158 resulted in a concussion while 75,734 exposures for men resulted in 101 concussions. Does this show a gender difference in concussions rates among collegiate soccer players?

   (a) Write the null and alternative hypotheses.

   (b) We would like to use a theoretical distribution to conduct this test.

      i. List the four assumptions for a hypothesis test (Activity 20). Is each one met?

      ii. Calculate the standard error of the difference in sample proportions.

      iii. Calculate the z test statistic.

      iv. Find the p-value.

      v. Write-up your results using all five components required. Use a 10% significance level to make your decision.

(c) Instead of performing a two-proportion z-test, create a confidence interval to estimate the difference in the concussion rates between male and female soccer players.

i. What confidence level should be used? *90%*

ii. Give and interpret the interval.

iii. Does the interval agree with your conclusion from the hypothesis test? Explain.

iv. Write-up the results of your confidence interval (including method used (and number of trials if appropriate), interval estimate, interpretation of the interval estimate, and conclusion regarding the null hypothesis).

Are You Two Together?

In the "Energy Drinks" study we've looked at several times, each participant took a test before the experiment really began. She returned about 10 days later, and was randomly assigned a treatment (RED, REDA or Control) and was again given a very similar test. The response variable we looked at was her "change in RBANS", meaning we subtracted the first test score from the second.

1. The "Repeatable Battery for the Assessment of Neuropsychological Status" outputs scores for immediate memory, visuospatial/construction, language, attention, and delayed memory. Discuss: What attributes of a person would make her score higher or lower? Write down two or three of your best guesses.

   Intellect, sleepiness, memory, motivation.

2. Write down an estimate of how much a person's attribute will change in 10 days for each of the attributes above. Think of how much one person changes relative to differences between different people (take $\sigma = 4$ to be the spread from person to person).

   Intellect and memory do not change much from day to day motivation could change a fair amount, and sleepiness could change a lot.

   The "Repeatable" part of RBANS means that it comes in several different versions which are all supposed to give the same scores to the same people. If the researchers had used exactly the same questions, subjects might have learned from the first attempt and done better the second time.

3. For each of the following situations, write "paired" if there is one sample measured twice or "two samples" otherwise. Ask your self: "Does it make sense to take differences and analyze those? (paired) Or do we compute means from each of two groups?" Another clue: If sample sizes might be different, it's not paired.

   (a) To study the effect of exercise on brain activity researchers recruited sets of identical twins in middle age. One twin was randomly assigned to engage in regular exercise and the other didn't exercise.

   (b) To see if curriculum changes are effective, researchers took a sample of 100 eighth graders' standardized test scores from this year and compared them to a sample of 100 scores of last year's eighth graders (on the same exam).

   (c) In a study to determine whether the color red increases how attractive men find women, one group of men rated the attractiveness of a woman after seeing her picture on a red background and another group of men rated the same woman after seeing her picture on a white background.

   (d) To measure the effectiveness of a new teaching method for math in elementary school, each student in a class getting the new instructional method is matched with a student in a separate class on IQ, family income, math ability level the previous year, reading level, and all demographic characteristics. At the end of the year, math ability levels are measured.

   (e) Each student in an intro stat class walked 100 yards twice. Once with arms down at his/her sides, another time while "pumping" arms up and down with each stride. The order of "pumping" or "not" was randomized for each student. Pulse (beats per second) was measured after each walk.

Being Careful With Wording:
With paired data, we are making inference about the "true mean of the differences." This is different from the wording we used with two independent samples. There we looked at the "difference in true means" – which makes sense because we had two populations to compare. With paired data, we have a single sample of differences. The observations subtracted to get the differences were not independent because they came from the same subject, but the sample of all differences can be independent as long as one subject (or unit) did not influence another.

Now we'll analyze some paired data.

Tears and Testosterone

Do pheromones (subconscious chemical signals) in female tears **change** testosterone levels in men?
Cotton pads had either real female tears or a salt solution that had been dripped down the same woman's face. Fifty men (a convenience sample) had a pad attached to their upper lip twice, once with tears and once without, in random order. Response variable: testosterone level measured in picograms/milliliter, pg/ml.

Take differences: Saline T–level minus Tears T–level.
The mean of the differences is $\overline{x}_D = -21.7$, and the spread of the differences is $s_D = 46.5$ pg/ml.

4. Test: "Is the mean difference 0?"

   (a) Check the assumptions. If the differences are fairly symmetrically distributed, can we use t-procedures? Explain.
   Yes. $n = 50 > 30$ *so we're OK. We have random ordering, and subjects' responses should be independent.*

   (b) State hypotheses in terms of parameter μ_D, the true mean of differences.
   $H_0 : \mu_D = 0$ *versus* $H_a : \mu_D \neq 0$
   STOP. Check the direction of the alternative with another group.

   (c) Compute t statistic as for testing $H_0 : \ \mu_D = 0$.
   $t^* = \frac{-21.7}{46.5/\sqrt{50}} = -3.30$

   (d) Do we use Normal or t distribution? If t how many df?
   t with 49 df

   (e) Look up the p-value in a web app and give the strength of evidence.
   $2 \times .0009 = 0.0018$ *This is very strong evidence against the null hypothesis.*

   (f) At the $\alpha = .02$ level, what is your decision?
   Reject H_0.

   (g) Explain what we've learned in context. (Give scope of inference. If outside observers would say the men studied were representative of all US men, how could the scope be extended?)
   We've found causal evidence that exposure to womens' tears does reduce testosterone levels in men. (p-value = 0.002). If this is a representative

sample, then we can extend the inference back to the population of US adult males.

5. Build a 90% CI for the true mean difference, μ_D.

   (a) Find the correct multiplier.
   $t^*_{49} = 1.677$

   (b) Compute the margin of error and build the interval.

   $ME = 1.677 \times 46.5/\sqrt{50} = 1.677 \times 6.58 = 11.03;$ CI: $-21.7 \pm 11.03 = (-32.7, -10.7) pg/m$

   (c) Interpret the interval in context. What do you mean by "confidence"?
   We are 90% confident that the true mean difference in testosterone levels for men without and with exposure to womens' tears is in the interval (-32.7, -10.7) pg/ml. This says that the tears really did cause a decrease in T–levels in this group of men. Our confidence is in the process by which we built the interval. When the procedure is used over and over, 90% of intervals built this way (in the long run) will include their true mean.

6. Designing a study: Researchers at the Western Transportation Institute (just south of the football stadium) use a driving simulator to test driver distractions (among other things). It is the front end of a car with large projection screens. Suppose they want to assess how distracting it is to read text messages on a phone while driving. The response they measure will be response time when a child suddenly runs out onto the road in front of the car.

   (a) How would they set this up to use paired measurements?
   Each driver does the same course twice – once while reading a text and again without the phone.

   (b) How could randomization be used?
   Randomly (flip a coin?) select whether each driver gets the text reading first or second.

   (c) Alternately, they could just test half their subjects while reading a text and half without the texting. Which study design do you recommend? Explain why.
   Pairing seems like a good idea because some people have faster reaction times than others. By subtracting two reaction times you get rid of other effects such as age or sleepiness.

7. A study of the effects of drinking diet soda versus sugar–sweetened soda on weights of 32 adults used a "cross–over" design in which each adult was given diet soda to drink for one six week period, and given sugar–sweetened soda to drink in another six week period. The order of assignment was randomized for each participant. The response measurement was "weight gain" over the six weeks.

   (a) It the parameter of interest the difference in true means or the true mean difference? Explain.

True mean difference. This is a case of one sample measured twice. We subtract weight gains (say sugar minus diet) so that our observations are independent of each other.

(b) How would you process the data in order to analyze it?
Take difference in weight gain (diet minus sugar) for each person.

(c) What distribution would you use to find p-values or a confidence interval multiplier?
t_{31}

Take Home Message

- Taking two measurements on the same subjects is quite different from taking two samples or assigning two treatments. You need to read carefully to see how data were collected. We do not have independent measures if they are on the same units.
- With two independent samples the parameter of interest was "difference in true means." With a single sample measured twice, the parameter is "true mean difference".
- To analyze paired measures, take differences first (before averaging) and use the one-sample t procedures.
- Pairing is a good strategy for reducing variability.
- The Energy Drinks study has a lot going on. They took differences first, to get "change in RBANS", but they also had three independent groups to compare: Control, RED, and REDA.
- What questions do you have? Write them here.