Stat 216 Course Pack Spring 2016

Activities and Notes

Photo by Kelly Gorham

Dr. Jim Robison-Cox

Department of Mathematical Sciences
Montana State University

Contents

1 Stat 216 Intro and Syllabus Summer 2016

People

- Your Instructor: (Write contact info here)

- Student Success Coordinator: Melinda Yager
  email: melinda.yager@montana.edu Office: Wilson 2-259 406-994-5344

Course Materials

You need to buy the Stat 216 Course Pack the MSU Bookstore. It will not work to use an old one from a friend.

Other materials, such as readings and assignments 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.

Recommendation: In D2L you can click on your name, go to your account settings, select the "Email" tab, and set **Forwarding Options** to send D2L mail to an account which you check more regularly. We strongly recommend that you do this. We might need to send out updates, and forwarding means you will not have to login in to D2L to get them.

We will 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.

Course Description

Stat 216 is designed to engage you in statistics using a simulation approach to inference via web apps. Small group discussion activities and daily assignments have been shown by the research to be effective. Upon completion of this course, you should understand 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. You will need to prepare for class every day and to turn in assignments twice per week.

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

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. We will cover tests and estimation in the contexts of: one proportion, one mean, two proportions, two means, and a regression slope.
- Understand when we might make causal inference from a sample to a population.
- 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.

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` or if those are unavailable use the site: `https://jimrc.shinyapps.io/Sp-IntRoStats`.
  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. You will need at least one laptop within your group each day.
- **Appropriate Use**: We need to use web apps, but it is NOT OK to use other websites during class. **You may not I-chat or text with friends or use web sites other than those we direct you to during class.** Our class time is really limited. We need to use it for group work and for instructors to give intros, wrapups, and reviews. Students who use technology inappropriately will lose attendance or RAT points for the day, and will have to leave the room if they cannot stop such behavior.
- **Turn OFF your cell phone and put it away**.

Math Learning Center in 1-112 Wilson Hall is a very important resource providing help on Stat 216 topics. It is open every day, into the evenings on MTWR, and closes early on Friday.

Assessment

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

- **Assignments: 25%** 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 a D2L Dropbox 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.
- **Midterm Exam 30%** Taken individually, not in groups. You may bring a one handwritten sheet of notes.
- **Final Exam 35%**.
  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%). In addition to attending, it's critically important that each student participates in class. Lack of participation can result in the same penalty as absence.

  Online students are expected to spend an equivalent amount of time in the course "Chat Room".

  Preparation: The in-class activities and out-of-class assigned readings and videos are the primary source of information for this course. Take them seriously, work through them with care. As a way to provide further emphasis to the activities and readings, most classes will include 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.

Planning Ahead: In our experience, it takes 6 to 9 hours per week outside of class to achieve a good grade in Stat 216. By "good" we mean at least a C because a grade of C- or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day's class.
Summer merits a special warning – each week is like three weeks of a spring or fall semester. You really need to spend time with this material – at least 20 hours per week.

The Math Sciences office cannot accept assignments and cannot provide information about grades (you can check on D2L – they can't).

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

Discussing assignments with others (in your group for example) is a good way to learn. Giving others answers is not doing them a favor, because then they aren't learning the material. Copying from others is cheating, and will not be tolerated. 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

University policy is explicit that the adviser and instructor must approve requests to withdraw from a course with a grade of "W". Students who stop attending and stop doing the work are not automatically dropped. Taking a "W" does not hurt your GPA but it is a sign that you are not making progress toward your degree, and could affect your financial aide or student loans.

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. These 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.

1.1 Martian Alphabet

How well can humans distinguish one “Martian” letter from another? In today’s activity, we’ll find out. When shown the two Martian letters, kiki and bumba, write down whether you think bumba is on the left or the right.

When your instructor tells you which is correct, write down whether you got it right or wrong.

1. If humans really don’t know Martian and are just guessing, what are the chances of getting it right?

2. We will assume that humans are just guessing. Discuss with your group: How can the three of you use coins and the “just guessing” assumption to mimic an the number of people in a group of three who would get the right answer just by chance?

3. We will now gather some data. Each of you will flip a coin 3 times and record the number of Tails. Sketch a plot of the numbers of Tails everyone got. The number of Tails will represent the number of right guesses of the Martian letters in three attempts.

4. Our class of thirty-some students might not give a clear picture of the distribution. Your instructor will use a web app to get several 1000 trials. Sketch the distribution here.

5. Now return to the ’bumba’ results and count the CORRECT bumba results in your group. Is your group particularly good or bad at Martian? How do you tell?

6. Let's collect more data, because just 3 people do not provide much information. We want to combine 3 or 4 groups (as instructed) to have 9 or 12 of your responses. What will change from # 3 above?

   (a) Each flip a coin ______ times to see what would happen under the "just guessing" scenario.

   (b) Change the spinner app to get the right distribution.

   (c) Sketch the distribution. Your instructor will pick 9 or 12 students to see how unusual are their 'bumba' answers are relative to the "just guessing" spinner results. Where does their number correct fall?

7. Finally, we'll use data from the whole class.

   (a) How do we change the spinner app to get the correct distribution? Sketch it here.

   (b) How unusual are the classes answers relative to the "just guessing" spinner results?

8. Is it possible that we could see results this extreme just by chance?

9. Does this activity provide strong evidence that we were not just guessing at random? If so, what do you think is going on here?

Take Home Messages

- In this course we will learn how to evaluate a claim by comparing observed results (classes guesses) 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.
- Purchase a copy of the course pack.
- Log in to this course on D2L. Set message forwarding to an account you read daily.
- View videos 1a through 1e posted on the Videos Link of D2L.
- Read the Syllabus and "Readings 1" for the next class. You will be quizzed over them.

Reference for "Martian alphabet" is a TED talk given by Vilayanur Ramachandran in 2007. The synesthesia part begins at roughly 17:30 minutes. `http://www.ted.com/talks/vilayanur_ramachandran_on_your_mind`

2 Reading 1 – Descriptive Statistics

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

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.

2.1 Data summaries vary with data type

Categorical variables are summarized with tables like this:

category	count	proportion
html	13	0.26
text	37	0.74

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

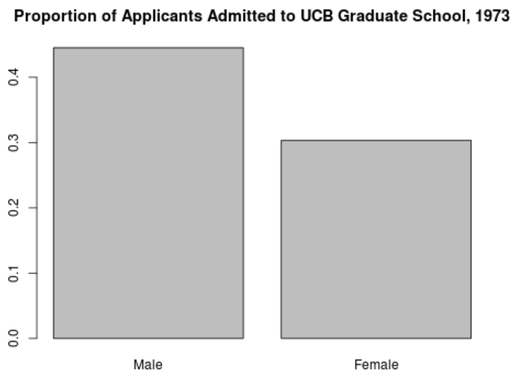
$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

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

2.2 Plotting Data

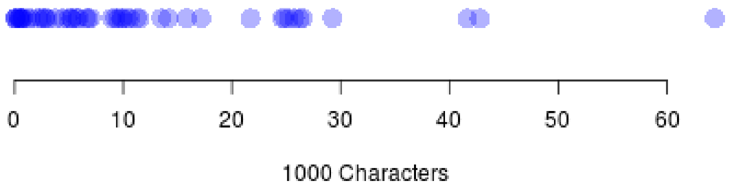
As with numeric summaries, the type of data determines the appropriate plot.

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.)

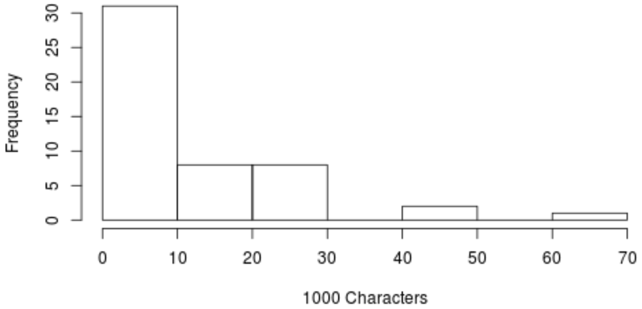


Quantitative variables are plotted with dot plots, histograms, density plots, and boxplots.

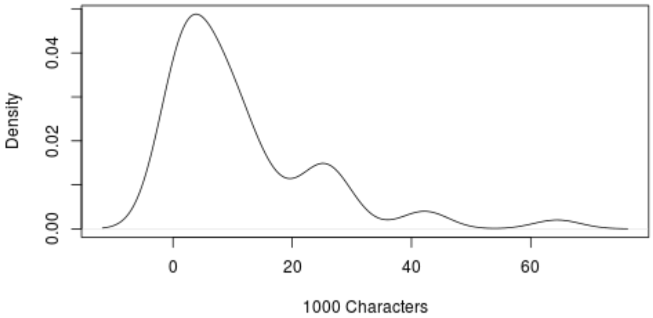
dot plots represent each point with a dot above the number line. This works well with small sample sizes. If the data are too close together to distinguish, we might stack them up to remove any overlap.



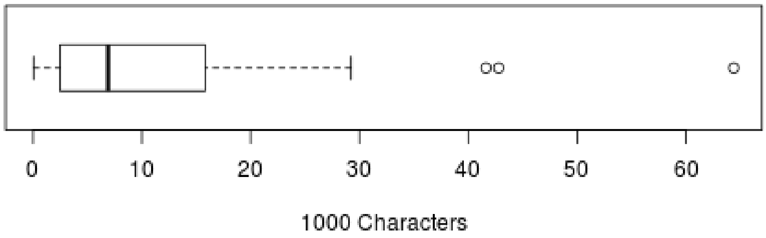
histograms divide the axis into "bins" and count the numbers of points falling into each bin. The height of each bin might show the count (frequency) of values in the bin or the proportion (relative frequency) for the bin. These plots work with moderate to large sized data sets. Choosing the best number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

Important Points

- From the Syllabus (p 1-5) What portion of your grade comes from D2Quizzes?
  from D2Boxes?
  from Attendance, Preparation, Participation?
- What is your goal for a grade in this class?

  Will you be able to spend 9 hours per week (outside of class) to achieve that goal?
- Who in your group will bring a laptop to the next class?

From pages 10–13:

- What are the two main types of data mentioned in this reading?
- What plots are used to display each type of data?
- How do we summarize each type of data?

3 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.

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.

One thing we need to be comfortable with is summarizing data. As you read in the reading for today, we first have to identify the type of variable, then decide how to summarize it. You've read about two main types of data:

Quantitative takes numeric values which we can average.

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

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

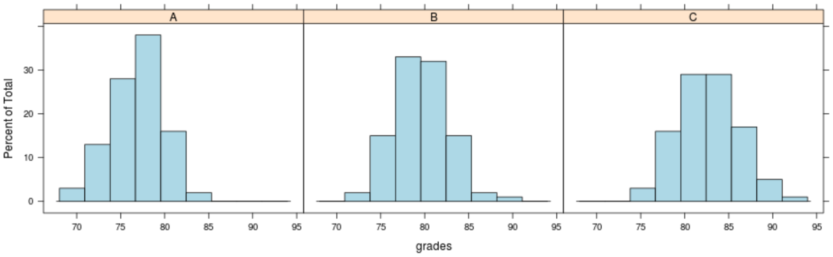
We also need to summarize categorical data, so we use proportions: the number in a certain category divided by the total number.

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

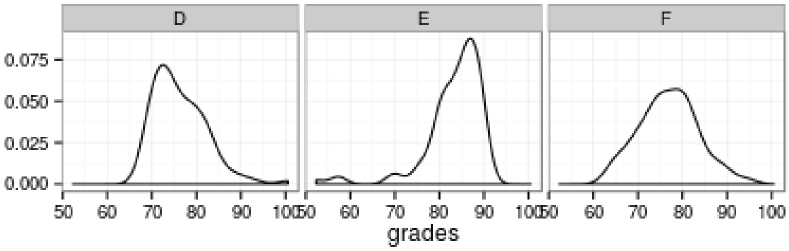
3.1 Comparing Distributions

Now we'll focus on quantitative data.

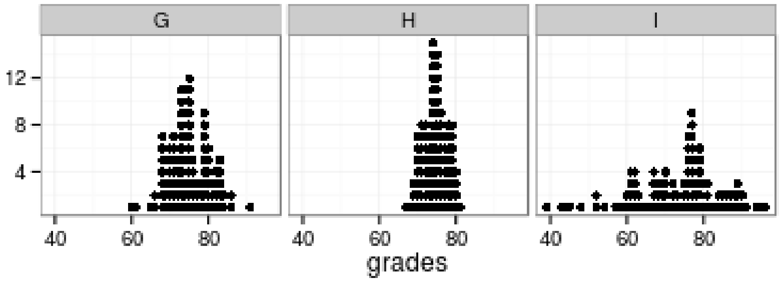
6. Suppose you are choosing which professors' class to enroll in. You have three choices, and have data on the grade distribution for each shown as histograms. Which class seems to have the best grade distribution? Explain.



7. Here are density plots of another set of three distributions of exams scores. Which do you prefer? Explain why.



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



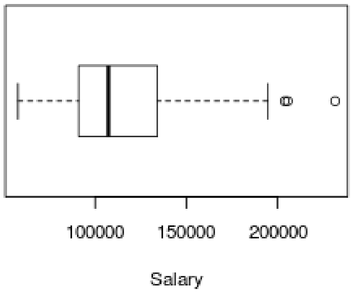
9. When comparing distributions there are several things to consider:

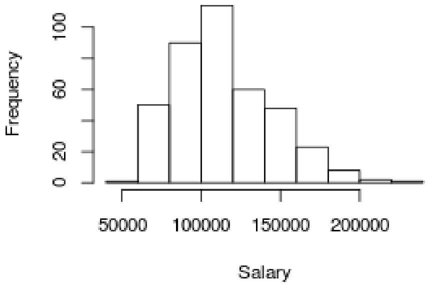
   - Comparing location or center (measured by mean or median) tells us which class did best “on average”.
   - Comparing spread (interquartile range or standard deviation) tells us which class is generally closest to its mean.
   - Comparing skew (could be left or right) to symmetric tells us which tail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

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

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

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





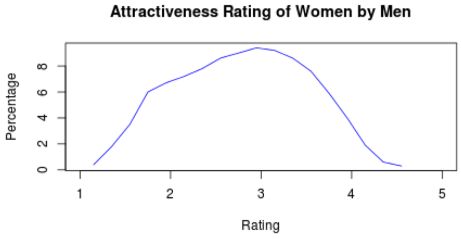
(a) Is salary skewed (if so which way?) or does it have a symmetric distribution?

(b) Are any points flagged as outliers? If so, describe them.

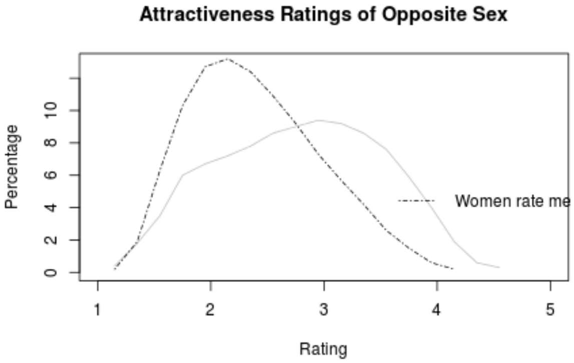
(c) Give approximate values for the median and the first and third quartiles. Also compute the IQR.

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

12. In Christian Rudder's book *Dataclysm* (2014) he shows plots of how men rate the attractiveness of women (data from the online dating site OKcupid) on a scale of 1 to 5 – the solid line in this plot. Y axis is the percentage of women who get this ranking. The line connects what would be the centers at the top of each bar of a histogram, (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.



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



Take Home Message:

- To learn about the world, we collect data. Two main types:
  - Categorical – summarize with proportions
  - Quantitative – describe center (mean or median) 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.

Assignments

- A template for a “Box” assignment is posted on D2L. Your completed assignment must be exported as a pdf file and uploaded to the D2L dropbox folder for D2Box # 1.
- Read Reading 2 for the next class.
- View Video # 2 listed in the videos link.

4 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.

Case: One row of data pertaining to one unit.

Variable: A quantity of interest which is measured or observed on units.

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 (pronounced mew) to represent the population mean.
σ (pronounced sigma) to represent the population's standard deviation (spread).
p (just plain pea) to represent a population proportion.
ρ (the Greek letter "rho" which sounds just like row) for correlation between two quantitative variables in a population.
β_1 (read it as beta-one) 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:
$\overline{x}$ (read it as ex–bar) to represent the sample mean (or average value).
s to represent the sample's standard deviation (spread).
$\widehat{p}$ (read it as pea–hat) to represent a sample proportion. (We often use a hat to represent a statistic.)
r for correlation between two quantitative variables in a sample.
$\widehat{\beta}_1$ (beta–hat one) slope of the "best fitting" line between two quantitative variables in a sample.

In this Unit 1, we will focus on parameter p and will use 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**. Bias creeps in in several ways:

- Asking a leading question can bias results.

- Missing a part of a population can bias results. For example, it's very hard to sample the part of the US residents who have no home and no phone.
- When a web page or a newspaper asks for peoples' opinions, it is typically the people with strong opinions who take the time to respond.

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." Non-responses cause bias in the results if those who fail to respond differ (in their response) from those who do respond.

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** which combines simple random samples from each level into one big 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.

Important Points

- Know that we gather data from the **sample** to learn about the **population**.
  - A number describing a population is called a
  - A number describing a sample is called a
- Why is a representative sample important?
- How can we be sure we are getting a representative sample?

5 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. (This is different from how we usually collect data. In practice we never have the entire population. Here we have created a small population to learn how well our methods work.)

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. The professor 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?

   Note: Do this quickly. Our goal will be to use the sample to estimate average word length in the entire text, but do not try to study the text too closely. Two minutes should be plenty of time to select 10 words.

2. Did you use any strategy to select words at random?

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

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

STOP!

Give your sample means to your instructor.

5. To evaluate a method of estimation, we need to know the true parameter and we need to run our method lots of times. That's why we chose a small population which we know has mean word length of 4.29 letters. (Where does 4.29 appear in the web app?). 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 everyone's sample mean we will look at another method:

5.1 Simple Random Sampling

(a) Point your browser to `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`. Bookmark this page, as we'll come back here often.
Click on One Quant. because we are dealing with one quantitative variable – word length – and drop down to Sampling Demo.

(b) The joke text should appear in the gray box. You can drag across this text and delete it if you want to paste other text into the box, but leave it there now.
Click Use This Text. You should see a plot of all word lengths with summary information. This is our population of 242 words.

(c) Set Sample Size to 10 and click Draw one Sample. Write out the 10 words and their lengths.

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.

7. Click Draw one Sample again and record the next mean.

8. Click the "More Samples:" choices to obtain at least 3000 more samples. Record the mean and standard deviation of all the sample means. (See upper right of the plot.)

9. If the sampling method is unbiased, the estimates of the population average (one from each sample of size 10) should be centered around the population average word length of 4.29. Does this appear to be the case?
Copy the plot here and describe what you see.

10. Click on the leftmost blue dot. The "Sample Words" change to show you the sample with the smallest average. How many one-letter words are in this sample? Copy the sample and its mean here:

11. Click on the rightmost blue dot. What is your longest word? Copy its mean here:

12. **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.

13. The actual population mean word length based on all 242 words is 4.29 letters. Where does this value fall in the above plot? Were most of the sample estimates around the population mean? Explain.

14. For how many of us did the sample estimate exceed the population mean? What proportion of the class is this?

15. Based on your answer to question 14, are "by eye" sample estimates just as likely to be above the population average as to be below the population average? Explain.

16. Compare the applet plot from question 9 with the plot from 12. Which method is closer to being **unbiased**? Explain.

5.2 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. Larger samples do provide more information about our population (but they do not fix a problem with bias).

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

17. Back in the web app, change "Sample Size" from 10 to $\boxed{25}$. Draw at least 3000 random samples of 25 words, and write down the mean and standard deviation of the sample means.

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

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

20. 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?

21. Compare the spreads of the plots in 9 and 18. You should see that in one plot all sample means are closer to the population mean than in the other. Which is it? Explain.

22. 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.

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

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

Take Home Messages

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

Assignment

- For D2L QUizzes, remember: you can save and come back, but once you hit "submit" you cannot change any answers.
- Reading 3 on Helper–Hinderer research.
- View Helper, Hinderer, and "Ethics for Babies" posted as 3a – 3c and video # 4 in the videos link before the next class.

6 Ethical Instincts of Babies?

Researchers at Yale University were interested in how soon in human development children become aware of (and start to favor) activities that help rather than hinder others.
Title: "Social evaluation by preverbal infants"
Authors: J. Kiley Hamlin, Karen Wynn & Paul Bloom
Journal: *Nature* 450, 557-559 (22 November 2007)
Abstract

> The capacity to evaluate other people is essential for navigating the social world. Humans must be able to assess the actions and intentions of the people around them, and make accurate decisions about who is friend and who is foe, who is an appropriate social partner and who is not. Indeed, all social animals benefit from the capacity to identify individuals that may help them, and to distinguish these individuals from others that may harm them. Human adults evaluate people rapidly and automatically on the basis of both behavior and physical features, but the origins and development of this capacity are not well understood. Here we show that 6- and 10-month-old infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive: infants prefer an individual who helps another to one who hinders another, prefer a helping individual to a neutral individual, and prefer a neutral individual to a hindering individual. These findings constitute evidence that preverbal infants assess individuals on the basis of their behavior towards others. This capacity may serve as the foundation for moral thought and action, and its early developmental emergence supports the view that social evaluation is a biological adaptation.
>
> The following were randomized across subjects: (1) color/shape of helper and hinderer; (2) order of helping and hindering events; (3) order of choice and looking time measures; and (4) positions of helper and hinderer.

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.

Definition: The p-value is the probability of observing a results at least as the result we have observed if the null hypothesis is true.

We quantify the strength of evidence 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 “Martian Alphabet” in Figure 1. A group of 12 humans had 9 correct matches and 3 incorrect. The simulation assumed $H_0 : p = 0.5$, and counted the number of heads in 12 flips of a fair coin. (Head ↔ Correct). 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 group's score. The chance is $74/1000 = 0.074$ of getting a result this extreme when H_0 is true. The p-value of 0.074 is 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.

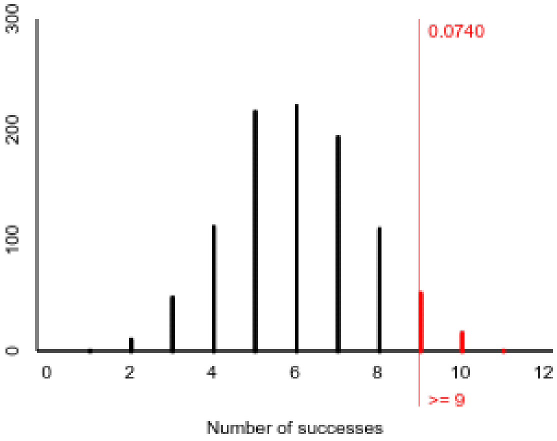


Figure 1: Simulation results obtained from the null model. The outcomes 9 and higher (74 out of 1000 trials) were as or more extreme as one group's number correct (of 12) and indicate the strength of evidence = 0.074.

For this group of 12, we would say that there is some evidence that they can read Martian, but while an event which can happen 7% of the time is fairly rare, it may not be totally convincing. A p-value of 0.07 is not really tiny, but is a “cautionary” yellow light.

Important Points

- From the abstract, what was the research question?
- What response was recorded? What type of variable is the response?
- How was randomness utilized?
- Would an outcome of 10 or of 8 provide stronger evidence against the null than our observed outcome of 9?
- Do smaller or larger p-values provide strong evidence against the null hypothesis?

7 Helper – Hinderer

Do young children know the difference between helpful and unhelpful behavior? You read about a study in *Nature*[1] which reported results from a simple study of infants which was intended to check young kids' feelings about helpful and non-helpful behavior. The research question is:

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

Data: Of the 16 infants age 6 to 10 months, 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.)

2. Suppose the infants really had no preference for one toy or the other, and the puppet show had no effect. What sort of results (numbers) would you expect to see?

3. Think back to our "Martian Alphabet" activity on the first day of class. What sort of evidence made us think that humans could decipher Martian script? Note: it depended not just on how many people in the class got it right, but also on the "background" distribution from the coin flips or the spinner.

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

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

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

6. In statistical language the idea of "no preference" is called the **null hypothesis** and it is written in terms of the population proportion, p = the true proportion of infants who chose the helper toy, as
$$H_0: \ p = 0.5.$$
We also have an **alternative hypothesis**, labeled H_a, which tells us the direction the researchers would like to conclude is true. For this situation, they think there might be a preference toward the helper, so they would like to conclude that H_0 is wrong, and that really
$$H_a: \ p > 0.5 \text{ is true.}$$
Under H_0, is it possible that 14 out of 16 infants could have chosen the helper toy just by chance?

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

8. **Carry Out the Simulation**
To see that happen, use the `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps` web app. Under the One Categ menu select Spinner. Set the Number of categories to 2, Labels to help, hinder, Percentages to 50,50, Stop after Fixed number of spins, Last spin: 16, and click Run to see a simulation of 16 kids choosing helper or hinderer when the two are equally likely. Record the number of "helpers", click Run again, and write down that number as well.

9. Set 1000 or more 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. Note: the direction we go from the observed 14 is toward higher values because the alternative, H_a was defined as $p > 0.5$ with the inequality pointing to

the right. How many of yours are this extreme? Circle these dots on your plot above. Check with the other groups nearby. Do we all agree?

11. Do you think that babies are just randomly choosing one of the toys? Explain.

You read about p-value or "Strength of evidence" in the reading for today. To help interpret strength of evidence, we offer this picture:

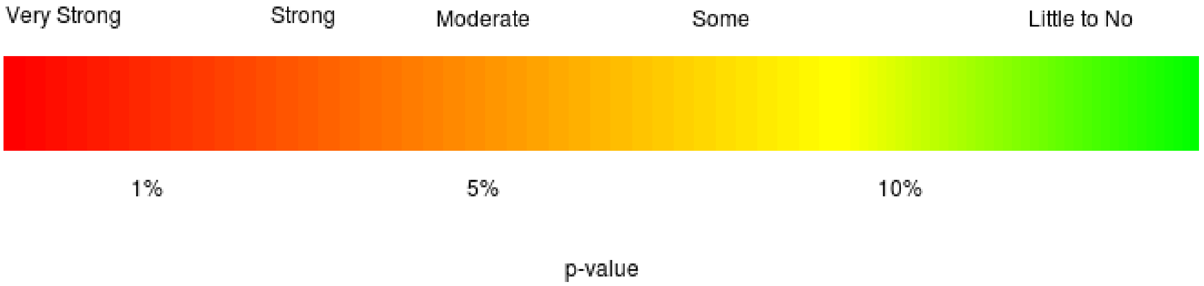


Figure 2: Numeric p–value and strength of evidence

The important point is that **smaller** p–values (in red) provide **stronger** evidence against H_0 and then H_0 gets a "red light". Red indicates that we don't believe it. We will soon talk about actually rejecting H_0 when the evidence against it is strong. Notation to watch: strong evidence is always against the null. We never have strong evidence in favor of the null.

12. Use your plot from above to quantify the strength of evidence for the observed result of 14 out of 16 infants choosing the helper toy. Give the numeric p–value and a verbal description of the evidence it provides.

13. Explain in your own words why **smaller** p-values provide **stronger** evidence against H_0.

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

15. Summarize how the p-value is computed.

16. Put the following steps into their proper order:
    (a) report strength of evidence
    (b) gather data
    (c) formulate a hypothesis
    (d) simulate a distribution
    (e) compare observed results to the distribution

17. Suppose another researcher had done similar study before this one and thinks that the proportion of all infants favoring helper is really 0.75. Change the spinner app to reflect this new hypothesis, compute a new p-value, and report the strength of evidence against $p = 0.75$.

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.

Assignment

- The last page of this course pack is a review table. You should tear it out and fill it in as we go. You will be able to bring it with you to exams. You can now fill in the top five boxes in column 1.
- Read the next two pages before the next class.

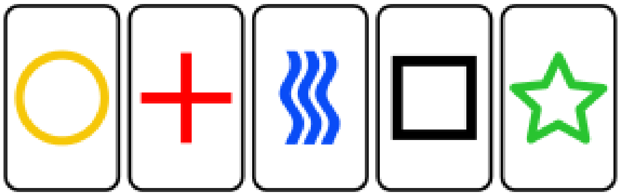
- Watch video # 5 on randomization distributions and hypothesis testing before class. Review # 4 as well.

- Make your own summary of the lesson.
  Thinking back about the most important ideas of this lesson help cement them in your head and help you avoid cramming right before the exam. Writing them here will make studying much easier.

8 Extra Sensory Perception

In the next classroom activity, we will look at an experiment conducted to see if a person could read another's mind.

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 use Rhine's data, but we want you to know that research into ESP (also called the "psi" effect) has continued.

Go to this blog and read the Skeptic's report on recent ESP research.
`https://skeptoid.com/episodes/4348`

Pay particular attention to how the researchers designed the experiment to remove all possible forms of communication between the individuals.

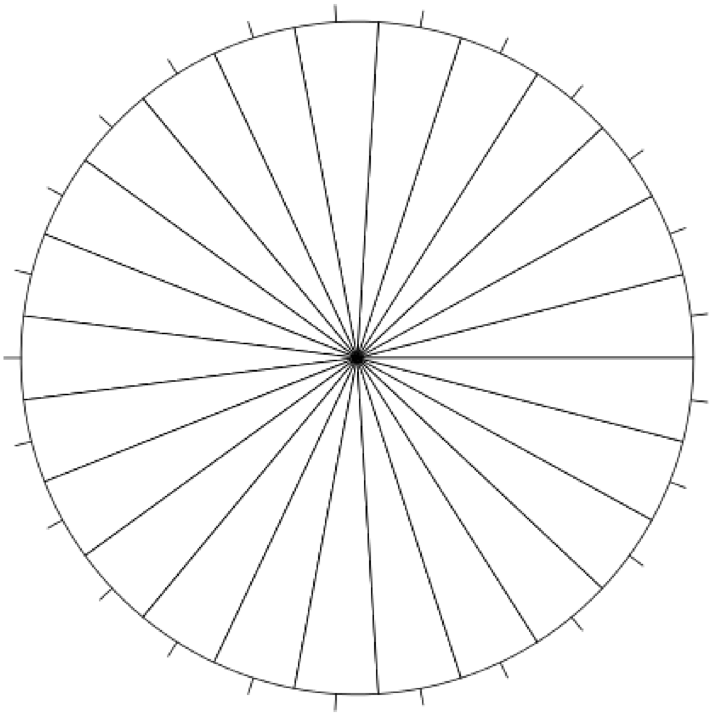
What is the "file drawer" effect?

What does the author find refreshing and unique about researchers studing the ganzfeld effect?

In the next activity, we will use a spinner to generate random outcomes.

To prepare for that, We'd like you to use the circle below (divided into 25 equal parts) to build a "Wheel of Fortune" (just like on the game show) with the following chance of each outcome:

Outcome	Chance
$2500	.04
$1000	.08
$900	.12
$800	.08
$700	.12
$650	.08
$600	.08
$550	.08
$350	.08
$100	.04
$1 million	.02
Bankrupt	.10
free play	.04
Lose a turn	.04



In the game show, they mix the outcomes up and give them different colors. It's fine if you want to do that, but we really want you to practice getting the right proportions, so putting, for example, all the $900 wedges together is fine.

Important Points

1. How could the ganzfeld experiment go wrong if the scientists were not very careful?

2. What was the chance the subject would – just by chance – pick the right object (or video) when given the choices?

3. On the real "Wheel of Fortune" do all outcomes have the same chance of getting picked by the pointer when the wheel stops?

9 Can Humans Sense Each Others' Thoughts?

We will investigate the data from Adam Linzmayer who correctly identified 9 of 25 Zener cards. Do these data show that Linzmayer had extrasensory perception? More broadly, Rhine wanted to know if anyone can catch a signal sent from another's mind using no other form of communication.

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

Step 2. Describe the study design and report the data collected.
Linzmayer was tested 25 times, and correctly identified the card 9 times in one trial.

2. What was recorded for each guess?

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

Step 3. Explore the data.
With categorical data, we report the number of "successes" or the proportion of successes as the "statistic" gathered from the sample.

4. What is the sample size in this study? $n =$

   Hint: it is not the number of people tested (just Adam).

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

6. Could Linzmayer have gotten 9 out of 25 correct even if he really didn't have ESP and so was randomly guessing between the five card types?

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

Step 4. Draw inferences beyond the data.
Two things could have happened:

- He got over one third correct just by random chance – no special knowledge.
- He is doing something other than merely guessing and perhaps has ESP.

8. Of the two possibilities listed above, which was Rhine trying to demonstrate (the alternative hypothesis) and which corresponds to "nothing happening" (the null hypothesis)?

9. What is the value of the **true parameter** if Linzmayer is picking a card at random? Give a specific value and use correct notation to denote it.

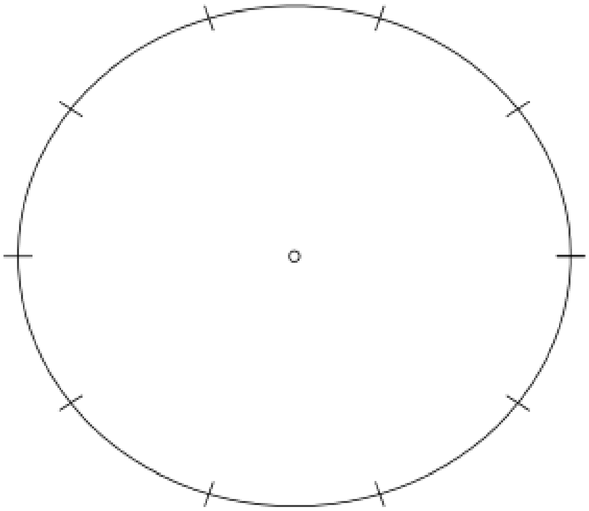
10. If Linzmayer is not just guessing and did have ESP, would you expect him to get a higher or lower proportion correct than the number from # 9? Use correct notation (an interval in parentheses) to denote this range of values.

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

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

    H_a:

12. Think of a "spinner" on a game board. How would you subdivide and color it so that each spin would be equivalent to Linzmayer randomly guessing one card and getting it right/wrong with the null hypothesis probability. (Hint: you do not need 5 segments.) Sketch your spinner on the circle below and shade the area where he gets it right just by chance. Put a paper clip on the paper with a pen to hold one end at 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 $\boxed{\text{Test/Estimate}}$ under the $\boxed{\text{One Categ}}$ menu. Enter the counts to show how many Linzmayer got right and got wrong. (These should add up to 25, but neither is 25.) Click "Use These Data" and record his proportion correct.

14. Now choose $\boxed{\text{Test}}$. and enter the value from 9 as the True Proportion. Run 5000 or more samples and sketch the plot below.

15. Check the summary statistics inside the plotting window. Does the mean agree with the null or alternative hypothesis? Explain why it should.

16. What proportion did Linzmayer get correct?
Type that value in to the box just after "than" below the plot. Select the direction (less, more extreme, or greater) based on the alternative hypothesis in 11. Click $\boxed{\text{Go}}$ and record the proportion of times this occurred.
Would you consider this an unlikely result?

17. Go back to figure 2 on page 39 to report the strength of evidence against H_0. Give the numeric value and a verbal description of the strength of evidence.

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

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

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

Take Home Messages

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

Assignment

- We strongly encourage you to get help in the Math Learning Center.
- Watch video # 6 before the next class.
- Read the next two pages.

10 Do rats feel for others?

Title: “Empathy and Pro-Social Behavior in Rats”

Authors: Inbal Ben-Ami Bartal, Jean Decety, Peggy Mason

Journal: *Science* **9** December 2011: Vol. 334 no. 6061 pp. 1427-1430

ABSTRACT

Whereas human pro-social behavior is often driven by empathic concern for another, it is unclear whether nonprimate mammals experience a similar motivational state. To test for empathically motivated pro-social behavior in rodents, we placed a free rat in an arena with a cagemate trapped in a restrainer. After several sessions, the free rat learned to intentionally and quickly open the restrainer and free the cagemate. Rats did not open empty or object-containing restrainers. They freed cagemates even when social contact was prevented. When liberating a cagemate was pitted against chocolate contained within a second restrainer, rats opened both restrainers and typically shared the chocolate. Thus, rats behave pro-socially in response to a conspecifics distress, providing strong evidence for biological roots of empathically motivated helping behavior.

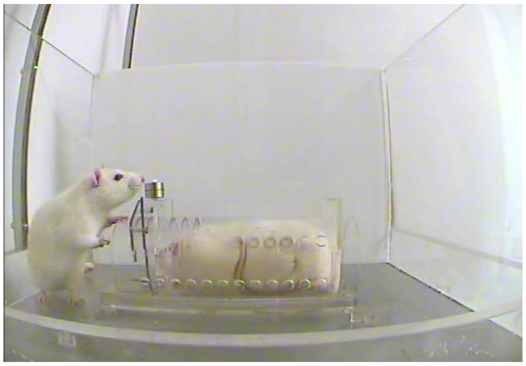
Watch this video:
`http://video.sciencemag.org/VideoLab/1310979895001/1/psychology`

Questions:

- What simple example of “emotional contagion” is mentioned?
- What was the free rat’s immediate reaction after first opening the cage door?
- What did both rats do when the caged rat was freed? (the first time).
- How did the free rat’s reaction change as it got used to the setup?
- Did the free rat open cages that contained:
  - chocolates
  - a toy rat
  - nothing
- What does Peggy Mason conclude is “in our brain”?

11 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.

Step 2. Design a study and collect data.

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

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

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

Step 3. Explore and summarize the data.

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

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

7. If the experiment were repeated with another 30 pairs of rats, do you think you would get exactly 23 who opened the cage again? Explain.

Step 4. Draw inferences beyond the data.

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

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

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

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

8. Your group will get 30 cards on which you will write (or check that the previous class properly wrote) 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, shuffle and draw a new card. 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. Record your second 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 – whereas we can repeat the resampling process many times.

After collecting many resampled statistics, we'll use the **percentile method** to compute a confidence interval.

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 Test or Estimate under the One Categ. header in the web app at `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`. Enter the rat data to look like:

Freed	23
Not	7

Then choose Estimate. What proportion of the rats were freed in your first resample? (Click the blue dot to see the resample.)

10. Now resample several 1000 times and copy the picture you get here.

Where is the distribution centered? (We are looking at the distribution of resampled $\widehat{p}$'s. The values in the corner of the plot are the mean and standard deviation of all blue dot values.)

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

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

12. You should have several thousand blue dots and the plot should have stabilized so that adding another 1000 doesn't change the shape much. Below the plot we have options for confidence limits for our interval estimate.

    (a) Click $\boxed{80}$ and count: What proportion are red points in the left tail?
    What proportion are red points in the right tail?
    What proportion are red points in the middle? Write the interval:

    (b) Click $\boxed{90}$ and estimate: What proportion are red points in the left tail?
    What proportion are reds in the right tail?
    What proportion are blue points in the middle? Write the interval:

    (c) Click $\boxed{95}$ and count: What proportion are red points in the left tail?
    What proportion are reds in the right tail?
    What proportion are blue points in the middle? Write the interval:

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

    (e) Play with the "Confidence Limit" buttons more to explain: How are the endpoints of the interval estimate related to the colors of the points in the plot?

    (f) Predict: what will happen to the interval endpoints of a 90 % interval, if we go from 5000 to 10000 resamples?

    (g) Try it and see: were you right?

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

Take Home Message

Several very BIG ideas:

- We only get one sample, but we can create many "resamples" using sampling with replacement (also called bootstrapping).
  Because we are estimating (not assuming a null value), we must sample **with replacement** to make each point come from the same distribution.
- 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.
- Any questions?

Assignment

- Watch videos # 7 and 8 before the next class.
- Fill in the simulation confidence interval box in column 1 of the Review Table.
- Read the next reading.

12 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.

Read the explanation from the Pew Research Center of how they conducted the poll, `http://www.pewinternet.org/2015/07/01/appendix-a-about-the-general-public-survey-2/`. The margin of error they give is for what confidence level?

How large is the margin of error for Republican/lean Republican?

For Democrat/lean Democrat?

12.1 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 $\sin(\theta)$.)
Open the web app: `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`.

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:

   (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).

   (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.

   (d) Compare the two intervals. Is one wider? Is there a shift?

13 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.

Simulation

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

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

The first slider on this page allows us to set the sample size – like the number of 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: (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.

1. 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)?

2. How does the center of the green (or red) interval relate to the $\widehat{p}$ you've clicked?

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?

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

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

6. Change the confidence level to 95%. Does the upper plot change? Does the lower plot? Describe any changes.

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

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

Data Analysis

9. Now back to the Pew study you read about for today. 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 as the first count in the Test or Estimate option under the One Categ menu and the balance of those 737 in the bottom box. Relabel the categories, then click Use These Data 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?

      ii. Click Estimate and run several 1000 samples. What is the SE?

iii. Find the “margin of error” for a 95% Confidence interval and create the interval.

iv. Are the endpoints close to those we get from the web app?

(b) Repeat for the Democrats:

i. Numbers of “successes” and “failures”.

ii. Margin of error and 95% CI related to it.

iii. Percentile interval and comparison.

(c) Explain what we mean by “confidence” in these intervals we created.

(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.

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 we have to tolerate some uncertainty.
- Any questions? How would you summarize this lesson?

Assignment

- Read the next two pages.

14 MIT – the Male Idiot Theory - Reading

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?"

Questions

1. What population is represented by these winners of the Darwin Awards?

[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.

2. Rephrase the researchers' question in your own words.

3. What parameter would answer that question?

4. What statistic gives us information about the parameter? What is its value? (Use correct notation.)

5. Would the question be better answered with a confidence interval or a hypothesis test? Why?

15 MIT – the Male Idiot Theory - Activity

1. What is the parameter of interest?

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

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

   (a) State null and alternative hypotheses in symbols and words.
   H_0 :

   H_a :

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

   (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 2 relative to the distribution you created? Explain in words where it falls relative to the plotted points.

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

4. Instead of considering a test of the true population proportion, we will switch gears and now estimate it.

   (a) What is our "point" estimate of the true proportion of idiots who are male (the sample statistic)?

   (b) In order to generate simulated data,

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

      ii. Explain how you would mark 318 cards and use them to simulate the gender of one individual, and then another.

      iii. What probability of being male is used?

      iv. After resampling 318 individuals, what number do you compute?

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

      i. Where is this distribution centered?

      ii. What is the spread of the distribution of resampled proportions?

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

(e) Explain what the word "confidence" means for this confidence interval.

5. Interpret this confidence interval.

6. Compare results from the hypothesis test and the interval estimate. If the null hypothesis is true, what value should be included in the 95% CI? Explain. Do the two methods agree to some extent?

Take Home Message:

- You just did two inferences on the same parameter. First, we tested the null hypothesis that half the world's idiots are male.
  You should have reported very strong evidence against that null hypothesis (less than 1/1000). We can feel quite confident that the true proportion of males in this exclusive group is more than one half.
- Secondly, we computed a 95% confidence interval for the true proportion of idiots who are male and you interpreted the interval. In 4e 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.
- Questions? Make your own summary of the lesson.

Assignment

- Review for the exam.
- Read pages the first two pages of Unit 2 before the next class after the exam.

16 Unit 1 Review

Vocabulary Define each term:

- sample
- population
- statistic
- parameter
- types of variables
- measures of center
- measures of spread
- sampling bias
- p
- $\widehat{p}$
- Null hypothesis
- Alternative hypothesis
- Strength of evidence
- Confidence interval
  Interpretation in context
  Meaning of “confidence”
- Margin of error
- Sampling with replacement

Simulation

1. If we repeat the "Helper – Hinderer" study and 10 of the 16 infants chose the helper (6 chose hinderer):

   (a) How would you assess the strength of evidence using the same simulation we already performed?

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

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

   (d) If we redid the study with 8 infants, and 7 chose the helper, is this stronger, weaker, or the same amount of evidence against the null hypothesis?

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

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

2. A German bio-psychologist, Onur Güntürkün, was curious whether the human tendency for right-sidedness (e.g., right-handed, right-footed, right-eyed), manifested itself in other situations as well. In trying to understand why human brains function asymmetrically, with each side controlling different abilities, he investigated whether kissing couples were more likely to lean their heads to the right than to the left. He and his researchers observed 124 couples (estimated ages 13 to 70 years, not holding any other objects like luggage that might influence their behavior) in public places such as airports, train stations, beaches, and parks in the United States, Germany, and Turkey, of which 80 leaned their heads to the right when kissing.

   (a) What parameter is of interest?

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

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

      i. State null and alternative hypotheses in symbols and words.
      H_0 :

      H_a :

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

iii. Use the $\boxed{\text{One Categ}}$ – $\boxed{\text{Test}}$ 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.

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

(d) Now estimate the true population proportion.

i. What is our "point" estimate of the true proportion of couples who lean right?

ii. In order to generate simulated data,

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

B. Explain how you would mark 124 cards and use them to simulate the lean of one couple, and then another.

C. Each couple leans right with what probability?

D. After resampling 124 individuals, what number would you compute?

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

A. Where is this distribution centered?

B. What number describes the spread of the distribution?

iv. Compute a 99% confidence interval.

v. Explain what the word "confidence" means for this situation.

(e) Compare results from the hypothesis test and the interval estimate. If the null hypothesis is true, what value should be included in the 99% CI? Explain. Do the two methods agree to some extent?

Unit 2

17 Experiments and Observational Studies

17.1 More Types of Variables

Sometimes a change in one variable causes another variable to change. For example,

- hours spent studying might affect a person's grade in statistics
- weight of a car might affect the gallons of fuel needed to go 100 miles
- major in college might affect "employment in field" six months after graduation. It might also affect starting salary.

In these cases, we have an **explanatory** variable which has an effect on a **response** variable. In other cases, variables are simply **associated**. For example:

- weight and length of Rainbow trout
- population of a city and the amount of taxes collected
- number of beds and number of patient deaths per year in hospitals
- cultivated acres and annual agricultural income in counties

Another variable might be causing changes in both (for example the age of the trout would affect both weight and length), or the connection might be more complex. Throughout this course we will be looking for associations and trying to determine if one variable is **causing** changes in the other.
Caution: We will often find that variables are associated (or not), but **Association does not imply causation!** You may have heard people say: "Correlation does not imply causation." They are trying to say the same thing, but in statistics, correlation has a very specific meaning. It is the strength (and direction) of a linear relationship between two quantitative variables and should not be used when one variable is (or both are) categorical. For now just say "association" instead. We will get to correlation in a few weeks.
For the following scenarios, circle any explanatory variable (if there is one) and draw an arrow to the response variable (if there is one).

1. Textbooks: cost, type of binding, number of pages

2. Chickens: breed, diet, weight gain

3. Test Scores: Exam 1, Exam 2, Exam 3

17.2 Types of Studies

Because we do often want to say that changing one variable causes changes in another, we need to distinguish carefully between two types of studies. One will allow us to conclude that there is a causal link; others will not.

Experiments are studies in which treatments are assigned to the units of the study. Units (or cases) can be people – called subjects – or things like animals, plants, plots of ground. A treatment variable has **levels** like different drugs or dosages or times or oven temperatures, so it is a categorical variable.
Randomized Experiments are those in which treatments are **randomly** assigned. These are very special – we'll keep an eye out for them.

Observational Studies are studies in which no variables are set, but each is simply observed, or measured, on the units (or cases). The variables could be categorical or quantitative and the units, as with experiments, can be subjects (people), animals, schools, stores, or any other entity that we can measure.

The next activity will discuss the differences between experiments and observational studies. You'll need a bit more information about how experiments are set up.

17.3 Four Principles of Experimental Design

Control. We want the different treatments used to be as similar as possible. If one group is getting a pill, then we give a "control" group a **placebo** pill which is intended to have no physical effect. The placebo may have a very real psychological effect because often subjects who believe that they are getting a drug do actually improve. To avoid these complexities, it is also important that subjects be "**blind**" to treatment – that they not be told which treatment they were assigned, and, in cases where measurements are subjective, that the raters also not know who received which treatment (a double–blind study is when neither subject nor rater knows which treatment was assigned).
In an agricultural experiment which involves spraying a fertilizer or an herbicide, the "control" plots would also be sprayed, but just with water.

Randomize. When doctors first started applying treatments, they thought that they should choose which treatment was given to each patient. We'll soon see that such choices lead to biased results, and that randomization is a better strategy because it makes treatment and control groups most similar in the long run. Statisticians recommend that we randomize whenever possible, for example, we like to randomize the order in which we make measurements.

Replication. Larger sample sizes give more accurate results, so we do want to make studies as large as we can afford them to be. Additionally, a principle of science is that whole studies should be replicated. With people, one study most often uses a very narrow slice of the human population, so it's a very good idea to run it again in a different country. The Skeptic reading from page 37 mentioned "meta analysis" which combines results from multiple ESP experiments to broaden the inferences we can make about ESP.

Blocking Agronomists comparing yields of different wheat varieties (treatments), have to worry about the fact that in any field there are patches which are more (or less) productive than others. Instead of randomly assigning varieties to plots across a large field, they split the field into "blocks" which contain more uniform plots, and randomly assign varieties within each block.

In a comparison of a new surgical technique with an old one, doctors might split patients into three different risk levels and assign treatments randomly within each group (block). Then each treatment group has roughly the same number of high, medium, and low risk patients, and we will get a comparison of the treatments which is not "confounded" with risk level. Studies are often blocked by gender or age, as well.

Two variables are "confounded" if their effects cannot be separated. For example, the first times we offered this curriculum, students taking a Tuesday–Thursday class all had the new curriculum in TEAL rooms, and the students taking MWF classes all had the old curriculum in non-TEAL rooms. We saw a large improvement in average test scores, but could not separate the effects of "day of week", curriculum, and room type.

Important Points:

- Know the difference between an explanatory variable and a response.

- What is the difference between an experiment and an observational study?

- How could MSU run a randomized experiment to compare two different types of STAT 216 curriculum? How would we set that up?

- Why do clinical trials expect doctors to use a double-blind protocol?

- What are two ways in which replication is used?

- Suppose we want to compare two weight reduction plans. How would blocking on gender change the set up of the experiment?

Does Music Help Us Study?

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

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

   A. With music

   B. In silence

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

   (a) what volume?

   (b) with lyrics? or instrumental?

   (c) what general category do you prefer?

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

   (a) The survey will ask for details on the music type each student prefers for studying, but we will also need a way to measure how effective their studying is. How could we measure a **response** to use for comparison – to see how much people are learning while studying?

   (b) In discussing the response, you probably found difficulties which make it hard to compare people. What differences in students make it hard to get a clear comparison between different music types? List at least three variables that we should consider. For each: is it categorical or quantitative? Focus in on one categorical and one quantitative variable.

3. In a 2014 study of the effect of music type on studying treatments were assigned. Perham, N. and Currie, H (2014). Does listening to preferred music improve reading comprehension performance? *Applied Cognitive Psychology* **28**:279–284.

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

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

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

4. Looking back at the studies in 2 and 3 above, which was an experiment?

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

Advantages of Randomized Experiments

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

5. 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?

6. 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.

7. 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: `https://jimrc.shinyapps.io/Sp-IntRoStats` 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.

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

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

8. As we said above, we need to think about repeating the shuffling process over and over. Create at least $\boxed{5000}$ repeats and sketch the right-hand plot (above). Describe the center, spread, and shape of this distribution.

   center

   shape

   spread

9. 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?

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

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

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

13. **Very Important Question:** In general, how similar are group mean IQs when we randomly assign people into two groups?

14. Another lurking variable would be the fact that some people have taken a short course as an SAT prep and others haven't. If the course does what it claims, then it could be the reason for one group to score higher than the other. We will look at the proportions who have taken an SAT prep course in the treatment and control groups.

    (a) Is "took SAT prep course" a categorical or quantitative variable?

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

    (c) Again go to the web app and click Lurking Demo, but this time under One Categ. header. Change A to Prep and B to No prep and enter the numbers observed in each group. Then click Use These Data. Run 5000 shuffles and record the mean and SE of the differences $\widehat{p}_1 - \widehat{p}_2$.

    (d) You will have a few shuffles that give -1 or 1. How could that happen?

    (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. You can split the 28 evenly between Control and Treatment in the first row, and split the 72 evenly in the bottom row. We should them have 50 Controls and 50 in treatment. Click Use These Data. What proportions of "Prep" in treatment and what differences do you get for the first two randomizations?

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

(g) Compare with other groups. Do the pictures look the same? What are its:

center?

shape?

spread?

15. When we randomly assign people to two groups:

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

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

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

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

Take Home Messages

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

Assignment

- Watch videos # 1 (experiments and Observational Studies) and # 2 (Scope of Inference) under Unit 2 before the next class.
- Read the next two pages before your next class.

18 The Sampling Distribution

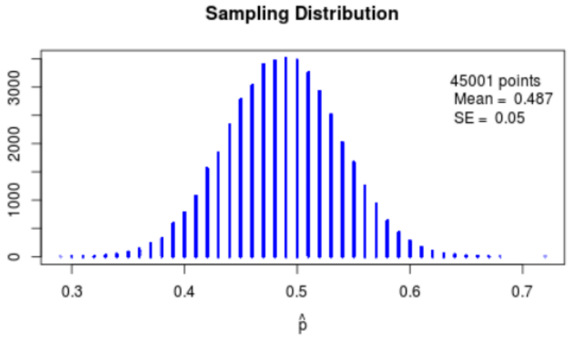
Statistical inference is based on simple ideas of random treatment assignment, random selection, and random sampling. **RANDOM** means that the outcome we will get cannot be known, but the distribution of possible outcomes can be known.

18.1 Sampling Distribution for $\widehat{p}$

Consider selecting a random sample of 100 people with season passes to a local ski run and asking if they snow board more than they ski. Our sample will produce a sample proportion – a number which we cannot know ahead of time. We will only select one sample, and will only get to see one sample proportion, but we can think about the process of random selection and consider all the sample proportions we might have obtained. This is a powerful way to think abstractly about the random selection process:

- We observe one sample.
- What else might we have observed?

The **sampling distribution** is a description of all possible outcomes and the probabilities of obtaining each outcome. If, for example, actually 48.7% of season pass holders board, then we could use a spinner to simulate the sampling distribution and would get a picture like this:

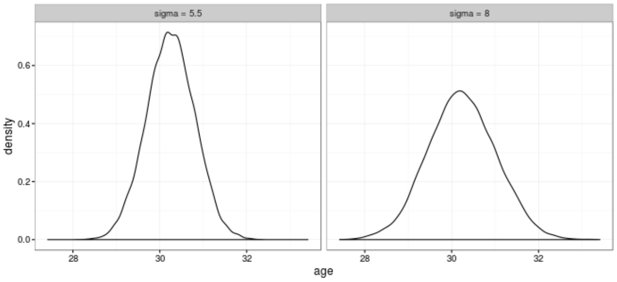


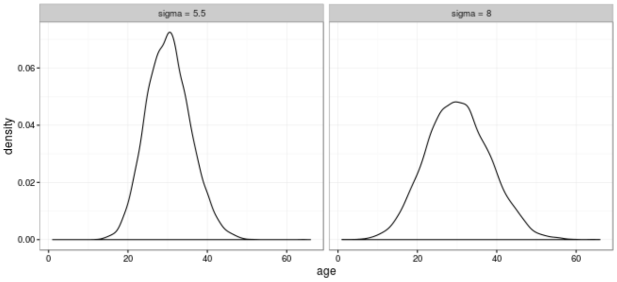
It is centered at the true value, 0.487 as the proportion of boarders, and has spread indicated by $SE(\widehat{p}) = 0.05$. It would be very useful to know the sampling distribution so that we could find the center part of the distribution for a confidence interval.

Sampling distribution for $\widehat{p}$ depends on two things: the true parameter p (unknown), and the sample size, n. As sample size increases, spread gets smaller.

18.2 Sampling Distribution for $\overline{x}$

Now suppose we are interested in the average age of season pass holders in a sample of size 100. The sampling distribution for the sample mean age, $\overline{x}$, again depends on the unknown parameter which is now μ, the true mean age in the population, and sample size, n, but it also depends on the spread in the original distribution, σ. The left hand pair of plots below is for individual season pass holders created (not real data) under two different assumed values for spread, either $\sigma = 5.5$ or $\sigma = 8$.





The right hand pair of plots shows the distributions of **MEAN ages of 100 season pass holders.**

A common confusion is to think that means will have the same distribution as the individuals. The two distributions will have the same centers, but when we average to get means, we pull in the extreme points. The distribution on the right has more younger and older ages, but still, when averaged over 100 people (top plot) we rarely see a sample average as low as 28.

Our point is that even though the distributions shown have the same means μ, and the same $n = 100$ the different values for σ change the sampling distributions for $\overline{x}$, the sample mean.

Important Points

- What does it mean to say that an outcome is "random"?

- Why does a plot of a sampling distribution show many points, not just one? After all, there is just one sample, right?

- Sampling distributions of $\widehat{p}$ and of $\overline{x}$ both depend on:

- Additionally, the sampling distribution for $\overline{x}$ depends on:

- Why is the sampling distribution for $\overline{x}$ less spread out than the distribution of the original data?

19 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? 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 allowed us to test a pre-specified value, not estimate an unknown parameter. We have a point estimate: a sample mean, $\overline{x}$, but we don't know how variable it is because we don't know σ, the true standard deviation of the data points.

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 parameters μ and σ. How can we estimate μ?

Solution:
Use the "Resampling" or Bootstrap distribution as a substitute for the unknown sampling distribution.

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 One Quant .

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. We will pretend that this is the entire population in order to see how well our methods work.

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

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

(c) Slow it down to where you can answer these questions: For one resample, which of the original eight values got used more than once? which not at all?

(d) Get 8 cards from your instructor and write each of the 8 values in your sample on a card. Create your own bootstrap resample to mimic what the computer does. Which of these methods works? (Circle one.)

   i. Select one card at random, leave it out, and select another card. Continue until you use all the cards.

   ii. Select one card at random and write down its value. Replace it, reshuffle, and select another. Continue until you've written down eight values.

(e) What statistic are we interested in (from the sample)? Compute it for the resample.

(f) Click 100 in the "Many Resamples" choices.

   i. Explain what values are being plotted.

   ii. A common quiz/exam question is "What does one dot represent?". Explain where the values came from and what statistic was computed to make one dot.

(g) Click 500 in the "Many Resamples" choices. Write down the interval estimate. Count (approximately) how many circles are outside the red lines at the left and at the right.

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

(h) Click 1000, 5000,and 10000 in turn. Write down three CI's for each. Compare the CI's. Are some groups off-center compared to others? More variable?

(i) Go back to 500 resamples. What happens to length of intervals when we change confidence levels? Hint: choose a different confidence level with the buttons, then click 500 again to obtain the interval.

`from 95% to 99% -- intervals` ______________________

`from 95% to 90% -- intervals` ______________________

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

(a) Look back at the 90% interval you wrote down. Did it contain the true value? Write "Hit" or "Miss".

(b) We'll now pretend that we can grab new samples and we will build two 90% CI's from each as a check of consistency. For each row of the table, click Get New Sample once, then click 1000 to get a 90% CI for μ. Record whether your first interval covers 34.5 (Hits) or not (Misses). Click 1000 again, and write "missed" or "hit" in the third column.

Click New Sample	1000 Hit or Missed?	1000 Hit or Missed?	Same?
1			
2			
3			
4			
5			

Click New Sample	1000 Hit or Missed?	1000 Hit or Missed?	Same?
6			
7			
8			
9			
10			

In each line above put a check in the last column if the 2 intervals agreed (both hit or both missed). Does coverage depend more on the sample or on the particular resample?

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)
   $\overline{x} =$

   (b) a 90% CI for μ is (show work)

   (c) a 95% CI for μ is (show work)

   (d) a 99% CI for μ is (show work)

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

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

   Will smaller sample size change the width?

   Will larger sample size shift the center?

   Will larger sample size change the width?

   Try it and record what happens to center and spread. (Yes, it is important to write it down. It will show up on the exam.) Using just one sample may not give you a good comparison. Try several samples at each sample size.

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.
- Questions? What is not clear?

Assignment

- View the video on Bootstrap - # 3 under Unit 2.
- Fill in the top three boxes of column 2 in the Review Table. Skip testing and fill in the bootstrap confidence interval.
- Read the next two pages before your next class.

20 Comparative Studies

With the textbook cost we combined all MSU students together and did not try to compare parameter values across groups. In many situation, however, the point of a study is to compare two or more groups. Here are some such example studies for you to consider.

1. Depression is a serious problem which affects millions of people each year. Suppose that you are asked to design a survey to compare answers of men and women to this question: If you were feeling depressed, would you visit MSU Counseling Services?

    (a) How would you select male and female MSU students to interview?

    (b) What are the variables you would collect in each interview? Are they categorical or quantitative? Is one explanatory? Is one a response?

    (c) The statistical inference tools you learned in Unit 1 do not quite apply to this situation. Why not?

    (d) Would this be an experiment or an observational study?

    (e) What would be your scope of inference?

2. In a clinical trial 183 patients with chronic asthma were randomly assigned to either placebo (n = 92) or budesonide (n = 91). After 12 weeks of treatment, doctors measured their lung function (Forced Expiration Volume in 1 second, FEV_1) in cc's.

    (a) How do you think these patients were selected to be in the study? Is there a larger population they were drawn from?

(b) What are the variables mentioned? Are they categorical or quantitative? Is one explanatory? Is one a response?

(c) Would it be appropriate to "block" the patients before randomly assigning treatments?

(d) What parameters would you compare between treatment and control groups?

(e) Was this an experiment or an observational study?

(f) What would be the scope of inference?

3. Key Points: What are two main differences between studies 1 and 2, and how do they affect the "Scope of inference" for each study?

4. In the last 10 years, the proportion of children who are allergic to peanuts has doubled in Western countries. However, the allergy is not very common in some other countries where peanut protein is an important part of peoples' diets.
The LEAP randomized trial, reported by Du Toit, et.al in the *New England Journal of Medicine* in February 2015 identified over 500 children ages 4 to 10 months who showed some sensitivity to peanut protein. They randomly assigned them to two groups:

   - Peanut avoiders: parents were told to not give their kids any food which contained peanuts, and
   - Peanut eaters: parents were given a snack containing peanut protein and told to feed it to their child several times per week (target dose was at least 6g of peanut protein per week).

   At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no).

   (a) What variables were measured? Are they categorical or quantitative? Is one explanatory? Is one a response?

   (b) Was this an experiment? was it randomized? were subjects blinded to the treatment?

21 Peanut Allergies

In the reading for today you learned about the LEAP study of peanut allergies.

The researchers want to answer this question:

Does feeding children peanut protein prevent peanut allergies?

Discuss the following questions in your groups:

1. Is there a treatment condition in this study? (If so, what?)

2. What is the response variable in this study?

3. Are the variables above quantitative or categorical?

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

4. Organize the results into a 2 by 2 table

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

5. Of the 245 subjects assigned to the eat peanuts group, what proportion developed allergies? 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.

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

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

8. What proportion of all 500 subjects developed allergies? This is called a marginal proportion because it just uses totals (the margins of the table, not the numbers in the middle of the table). 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”.

9. Write a few sentences summarizing the results in the sample. This summary should include a summary of what the data suggest about: (1) the overall risk of becoming allergic to peanuts in these subjects; (2) the differences between the two treatment groups; and (3) whether or not the data appear to support the claim that peanut eating is effective.

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

- Does the higher allergy rate in the control group provide convincing evidence that the peanut eating is effective?
- Is it possible that there is no difference between the two treatments and that the difference observed could have arisen just from the random nature of putting the 500 subjects into groups (i.e., the luck of the draw)?
- Is it reasonable to believe the random assignment alone could have led to this large of a difference?
- Just by chance did they happen to assign more of the subjects who were going to developed allergies into the peanut treatment group than the control group?

To examine these questions, think about what you would generally 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. Under that assumption (the null), you would expect, on average, about 20 of those subjects to end up in each group.

Next we will **Write out the hypotheses**

Reminder: 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.

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:

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

We will use a **permutation** test to compute our strength of evidence. "Permutation" just means that we are mixing up, or permuting, the group assignments. In physical terms, it's shuffling the cards (40 "allergy" cards and 460 "no allergy" cards) and redealing them into two groups (treatment and control). 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.

11. Go to the web page: `https://jimrc.shinyapps.io/Sp-IntRoStats` and select Enter Data under Two Categ. Enter the numbers and labels from the table in 4. The proportions should agree with those above, but let's check:

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

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

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

12. Click Test under Two Categ, generate 1000 shuffles and sketch the plot below.

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

14. Report the approximate p–value (i.e., strength of evidence) based on the observed result. (Reminder: we did this in the helper – hinderer study on Activity 6.)

    Go back to Enter Data and click Use Data to clear the plot. Generate another 5000 random shuffles. How much does the strength of evidence change?

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

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

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

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

19. Write up the pertinent results from the analysis. When reporting the results of a permutation test the following must be included:

    - The type of test used in the analysis (including the number of trials [shuffles]);

    - The null model assumed in the test;

    - The observed result based on the data;

    - The p–value and strength of evidence of the test and your conclusion; and

    - The appropriate scope of inference based on the p–value and the study design. Include:

        - How were the subjects selected? If they are a random sample from some population, then our inference goes back to the population.

  - Were treatments assigned? If treatments were assigned at random, then we can state a causal conclusion.

Take Home Messages

- We are conducting a permutation test which simply mixes up the labels. Because of random treatment assignment, this is also a randomization test.
- We tested to see if two proportions were equal. This is much like what we did in Unit 1 with a single proportion, except that the null hypothesis states that the two population proportions are equal (instead of one proportion coming from "blind guessing").
- Question 19 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?

Assignment

- Fill in the top five boxes in column 3 of the Review Table.
- Read the next two pages before your next class.

22 Statistical Significance

Disclaimer: Most statisticians prefer to just report p-value as the “strength of evidence” and let readers evaluate the evidence against H_0 for themselves. Furthermore, if evidence from a study could be summarized with a confidence interval, then that is a good way to report results, too. However, you will see results from many studies summarized as “statistically significant” or as “not statistically significant”, so we need to talk about the meaning of those phrases.

Statistical significance means

1. that researchers decided to use a cutoff level for their study, typically 0.10 or 0.05 or 0.01, and
2. the p-value for the study was smaller (stronger evidence) than the chosen cutoff level.
3. researchers rejected the null hypothesis at the given α level.

Significance level, α is the chosen cutoff. The three values listed above are most commonly used, but there is no reason (other than an ingrained habit) that we could not use other levels. How does one choose a significance level?

- If very strong evidence is required (lives are at stake, or the null hypothesis has very strong support) then a small α like 0.01 would be used.
- If we are just exploring a new area or the null hypothesis is not something people believe in, then weak evidence would be all that is needed to reject it, and we could use $\alpha = 0.10$,

Note that people will have different opinions on the above. That’s why we would rather let a reader decide how strong the evidence needs to be. In reports of results, we still want you to report the p-value.

Problems with fixed α significance testing

- Suppose we decide to use $\alpha = 0.01$ and the p-value is 0.0103. We then “Fail to reject” at our set α level. But someone else might repeat the study, get very similar results, and a p-value of 0.0098 which is less than α, so they reject H_0. If we just reported the p-value and let readers look at the strength of evidence, we would be able to say that the studies pretty much agree. With fixed level testing, we make quite different conclusions from very similar results. That is a disturbing **inconsistency**.
- P-values are strongly related to **sample size**. Whether we “reject” or “fail to reject” depends as much on the sample size as it does on the true state of nature.
  For example, the LEAP study of peanut allergies in the last activity used sample sizes of

245 and 255 infants and we had a p-value < 0.0001. What if they had used one tenth the sample size as here:

	Peanuts	Avoiders	total
Allergic	0	4	4
Not Allergic	24	22	46
Total	24	26	50

The proportions are almost the same, the difference in proportions is -0.15 but the p-value is 0.066.
Lessons: The researchers were smart to use a large sample size. However, if there was only a little difference in the two groups and they used a huge sample size, they would reject $p_1 = p_2$ when the difference in proportions allergic is very small. Which leads to our third point:

- Obtaining "Statistical significance" does not tell us anything about "**practical importance**". In common usage, the word significance means something like "importance". For example, we talk about "significant events" in history. Consider these examples:
  - YouTube carefully examines how people navigate their web site. Suppose they test two web page designs (assigned at random) on large groups of randomly selected viewers and find that there is a "significant" difference in mean time spent on their site (between the two designs) of 0.56 seconds (α was set to 0.05). Is that an **important** difference?
  - Many businesses use a call center to answer questions from their customers. They have to decide how many staff to have on hand to answer questions during business hours. If they have too few people, wait times get long, and customers hang up without getting to a support team member. Suppose they have to decide between having 6 or 8 people and they do a test to measure the proportion of individuals who hang up before getting help. With 6 people the hangup proportion is 0.34 and with 8 people it is 0.33. Because they gathered data on several thousand customers, the p-value for the test is very small, 0.002 which is "statistically significant" at the $\alpha = 0.05$ level. Is a difference of 1% of practical importance?

Important Points

- Remember: smaller p-values provide stronger evidence against the null. When we set a **significance level**, we reject H_0 for p-values **smaller** than α.
- Still report p-values.
- Statistical significance $\neq$ practical importance.
- When we do use a fixed α significance level, we will say either that we "reject H_0" (because p-value $\leq \alpha$) or that we "Fail to reject H_0" (when p-value $> \alpha$). We never accept the null hypothesis, but we can reject it.

23 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

   H_a

2. Go to the web apps page: `https://jimrc.shinyapps.io/Sp-IntRoStats` and select Two Categ, Test or Estimate. Type our numbers into their cells so that Success is "over"–estimate, 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.

[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.

(b) After checking the data, select $\boxed{\text{Test}}$, and generate 1000 shuffles. Describe what the computer does for a single "shuffle".

(c) Why is the plot centered where it is centered?

(d) Enter the observed difference in proportions, select the correct direction for the alternative, and write down the p–value and strength of evidence.

(e) Which is the more plausible explanation:

- These results occurred just by chance (We just got unlucky and saw something weird) or
- Men and Women who don't know what their weight is really do differ in their self-perception of being over versus under weight.

(f) 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)

State your conclusion about H_0 in the context of this situation. Be as specific as possible about the true proportions overestimating their weight.

3. Now we will estimate the difference. You might want to look back at 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 $\boxed{\text{Estimate}}$.

(a) What is the point estimate for the difference in population proportions? Use proper notation.

(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?

(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?

For each man being resampled?

How would you resample from the women using 70 index cards? Explain what to write on each card and how to randomly draw a card for each woman.

(d) When you're sure you know what it is doing, click $\boxed{\text{1000}}$ several times. Note that it adds more rather than getting rid of the old ones. Record center and SE of this distribution (upper right corner of plot).

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

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

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

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

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

4. Summarize the hypothesis test results in a report as in question 19 of the Activity 12. 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 12, 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.
- We have just used confidence intervals to estimate the difference between two proportions. Recall from Activity 7: 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, $\widehat{p}_1 - \widehat{p}_2$.
- Questions? What do you see as the key points of the lesson?

Assignment

- D2Box 5 is due Feb 25.
- **D2Quiz 6** is due Feb 29. Fill it in online.
- Watch video 4 under Unit 2 Examples.
- Fill in the Simulation Confidence Interval box in column 3 of the Review Table.
- Read the next two pages before your next class.

24 Energy Drinks - Reading

In the next activity, we will use data from this study:

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

Researchers used 27 volunteer college students from a private college in the US. All were women because the college has 3 times as many women as men, and the researchers thought it would be too hard (take too long) to get enough of both sexes. They advertised for volunteers, obtained consent to drinking beverages with alcohol and or caffiene, and were told that the study involved measuring creativity effects.
The abstract says:

> In a double-blind, placebo-controlled design, 27 non-caffeine-deprived female participants were randomly assigned to consume a caffeinated energy drink alone (RED), one containing alcohol (RED+A), or a non-alcoholic, non-caffeinated control beverage. Pre- and post-test assessments were conducted using alternate forms of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS).

RBANS comes in two forms (A and B) which have been shown to give the same scores to people in terms of cognitive status. One form of the test was administered several days before the subjects were given the treatment, and another was given 30 minutes after they had drunk the assigned beverage. The response we will look at is the change in RBANS scores (post score minus pre score). Initially we'll compare only the control and RED+A groups.

The RBANS test is called a "battery" because it is intended to measure many aspects of human cognition including: Immediate memory, Visuospatial/constructional, Language, Attention, and Delayed memory. The scores we will examine are combined (Total scale) across all those dimensions.

Two of the drinks were commercial products: RED = Green Monster (caffienated) and RED+A = Orange Sparks (6% alcohol and caffiene) while the control was colored Diet 7-Up.

Questions:

1. What does "double-blind" mean for this study?

2. What are the variables of interest? What type are they?

3. Was this an experiment or an observational study?

4. Were subjects randomly chosen from some larger population of people?

5. If we find strong evidence of an association between the drinks and the "change in RBANS" response, what will the scope of inference be? (Clearly explain and include your reasoning.)

25 Energy Drinks

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

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

1. Why are energy drinks popular?

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

3. How do energy drinks interact with alcohol?

4. An experiment tried to compare the effects of energy drinks with and without alcohol on human subjects. Pharmacology is the study of how drugs affect the body, and "psychopharmacology" studies effects of drugs on the nervous system. An article in *Human Psychopharmacology* in 2009 reported on an experiment intended to tease out some of

the effects and to compare an energy drink without alcohol to one with alcohol and to a non-energy drink. The research question is:

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

Higher RBANs scores indicate better memory skills.

Go to the site:
`https://jimrc.shinyapps.io/Sp-IntRoStats` Select Test or Estimate under One of Each. Select PreLoaded Data as the data entry method, and select REDAvsCntrl and Use These Data to load today's data.

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

Center

Spread (SD and IQR)

Shape

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

5. Here are the data in the way the experiment was run.

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

Use the web app to find the means of these groups and subtract to find the difference. (Control mean minus REDA mean.)

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?

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

7. How would you use cards to reassign these 18 women into new treatment and control groups, and then obtain means and a new difference in means?

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)

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?

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

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

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

   (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?

10. There are other reasons that one person might show more change in RBANS than another person. Write down at least one. (Again, don't get stuck here.)

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

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

13. We will also estimate the difference in the true mean RBANS change (REDA versus control) using the web app. Select the Estimate option under One of Each. Click 10 and mouse over the extreme dots in the plot. Notice how the lower boxplots, the lower summary table and the difference in means change. The app is creating a bootstrap resample from the 9 REDA values and a bootstrap resample from the 9 Control values, finding the mean of each, and subtracting to get a difference in means which shows up in the plot. Click again until you are positive you know what it is doing. For your last resample, write the two bootstrap means and their difference.

14. Obtain a 95% bootstrap percentile interval based on several 1000 resamples and write it here.

15. Write your interpretation of this interval.

16. Build a 95% confidence interval using “estimate $\pm t^*$SE” where the estimate is the observed difference in means, $t^* = 2.12$, and using the SE the plot. Does it contain zero?

17. 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.
- When we report results about a difference in means, we **must** specify the direction of subtraction.
- Questions? Summarize the lesson.

Assignment

- Watch videos 5 and 6 under Unit 2 Examples.
- Fill in the top six boxes in column 4 (one of each type) of the Review Table.
- Read the next page before your next class.

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.

26 Hypothesis Test for a Single Mean

Watch the video: Hypothesis Test for a Single Mean found:
??

Questions:

- Why might people care about average snow depth in the mountains around Bozeman and whether or not it has changed recently?

- Over the 30 years before 2011, what was average snow depth at Arch Falls?

- Over the past 5 years, what has average snow depth at Arch Falls been?

- What variable is of interest? Is it quantitative or categorical? What statistic will we use to summarize it?

- What are the null and alternative hypotheses?

- Step 2 – to create the null distribution – uses a technique you've never seen before, and it's a bit weird. We'll work on it in the next class activity, so if you don't get it from the video, that's OK.

- Aside from getting the simulated distribution, everything else should seem familiar. Why do we want the distribution to have the center it does?

- How do we determine the p-value in this case? What is it?

- What do we conclude?

27 Arsenic in Toenails

Symptoms of low–level arsenic 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). They want to know the mean arsenic concentration for New Hampshire residents drinking from private wells.

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

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

Step 1. State the research question.

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

Step 2. Design a study and collect data.

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

[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?

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

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

6. Calculate the observed statistic and use correct notation to denote it (check your answer with another group!).

7. Could your answer to 6 have happened if the arsenic concentrations in New Hampshire residents are not hazardous?

8. Do you think it is likely to have observed a mean like the one you got in 6 if the arsenic concentrations in New Hampshire residents are not hazardous?

Step 4. Draw inferences beyond the data.
We'll start with the first research question asked because we have done confidence intervals for a single mean back in Activity 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.

10. Explain how you can use a deck of 19 cards to create the bootstrap distribution. (Go back to Activity 11 if you don't remember.)

11. Use the One Quant option in the web applet `https://jimrc.shinyapps.io/Sp-IntRoStats` to use the pre-loaded data (arsenic) and then generate a bootstrap distribution with 5000 or more bootstrap statistics. Draw the plot below and record the summary statistics.

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

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

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

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

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

16. Interpret the 90% confidence interval from 15.

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

There are two possibilities for why the sample average was 0.272. List them here and label them as the null and alternative hypotheses also write the null and alternative in notation.
H_0 :

H_a :

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

We can simulate the behavior of arsenic concentrations in New Hampshire ground water if we assume the null hypothesis which gives a specific value for the mean. The two key ideas when creating the reference distribution are:

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

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

17. 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.)

18. How far is the sample mean from this null value?

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

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

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

20. Now we want to resample the shifted values. To speed up the process, we use Test option under One Quant at `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps`. You should have already loaded the data, but if not, go back to Input Data and select the preloaded Arsenic data.

- 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 left labeled Original Sample. Does the mean match your answer to 6? If not, consult with your instructor!

21. What is the statistic from the first resample? (Click the blue dot to see.)

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

23. Generate 5000 or more randomization samples. Copy the summary statistics and the plot of the randomization distribution below

24. Where is the distribution centered? Why does that make sense?

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

25. 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.

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

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

27. How strong is the evidence against the null hypothesis? Look back to the guidelines for assessing strength of evidence using p-values given on page 39.

Step 5: Formulate conclusions.

28. Based on this analysis, what is your conclusion about the residents in New Hampshire who own a private well based on this study?

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

Take Home Messages

- We first reviewed building a CI for a single mean.
- You need to know when to discuss means versus proportions. If the response has two categories, then we deal with proportions. If the response is quantitative, then we estimate and test means.
- The new twist today was to do a simulation for testing $H_0 : \mu = \mu_o$ that the mean is some particular value. We had to modify the data to make H_0 true, shifting it from its center at $\overline{x}$ 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).
- Questions? What is your summary of this lesson?

Assignment

- Fill in the simulation confidence interval box in column 2 of the Review Table.
- Watch the “Types of Errors” video (# 4 under Unit 2) and Example 8 under Unit 2.
- Read the next two pages before your next class.

28 Decisions and Justice

Read this article about the justice system and the errors that are possible when trying a suspect. `http://www.intuitor.com/statistics/T1T2Errors.html`

Important points

- Assumption:
  - In the justice system, what assumption is made about a defendant's guilt (or innocence)?
  - The web page points out the similarities between the justice system and statistical hypothesis testing. What is the usual assumption in hypothesis testing (which parallels the justice system assumption above)?
- Rejecting the assumption:
  - In the justice system, what information causes a jury to reject the assumption they start with? What is the standard for them to decide to reject it?
  - In hypothesis testing, what information is used to reject the assumption? How do we set the acceptable error rate?
- Conclusions:
  - In the justice system, if the jury decides it can reject the assumption, they find that the defendant is:
  - In hypothesis testing that is equivalent to:
  - If the jury decides to **not** reject the original assumption, they find the defendant:
  - In hypothesis testing that is equivalent to:

Their quality control example uses the assumption that a batch of some product is “not defective”. Someone would test the batch to see if it has any problems, in which case the whole batch would be rejected.

We have not yet used the normal distribution which they show. You can instead think of the distributions of dots from simulations you have seen in our web app. The idea of p-value is the same – that a statistic further out in the tail of the distribution gives a smaller p-value and is stronger evidence against H_0.

- Stronger evidence. Near the bottom of the web page they mention two ways we can reduce the chance of both type I and type II error. Here they are for the justice example. Fill in the same ideas for the hypothesis testing situation.
  - More witnesses
  - Higher quality testimony.

You might like their applet which illustrates how we can reduce the chances of error, but it requires java which is not supported on Mac devices.

29 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. An FAA weather forecaster has to decide if it is too dangerous (due to storms or potential icing conditions) to allow planes to take off. What is the cost of grounding a flight? What could happen if a flight should have been kept on the ground, but wasn't?

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?
What are the trade-offs?

29.1 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. Instead of focusing on the negative (error), we more often talk about the **power** of a test to detect an effect which is really present. Power = $1 - \beta$ is the probability of rejecting H_0 when it is false (that's a good thing, so we want power to be high).

Justice System and Errors

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 could report a big discovery when in fact, nothing is going on. (How embarrassing!) This is considered to be 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 I error rate is controlled by setting α, the significance level, and only rejecting H_0 when the p-value is less than α.

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

- 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, or 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.
- 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, 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.

The Power Demo web app lets us try different values to see what size power we get. Go to `http://shiny.math.montana.edu/jimrc/IntroStatShinyApps` and click Power Demo under One Quant.

Try different sample sizes.

5. Set SD to 2, Alternative Mean to 1.5, and Significance Level to 0.01. Find the power and the effect size for each sample size:

n	power	effect size
4		
8		
16		
32		
48		

6. What happens to power as you increase sample size? Explain why.

Try different standard deviations.

7. Set sample size to 16, Alternative Mean to 1.5, and Significance Level to 0.01. Find the power and the effect size for each Standard Deviation (SD). This SD is the spread of individual data points within our sample.

SD	power	effect size
0.4		
1		
2		
3		
4		

8. What happens to power as you increase standard deviation? Explain why.

Try different alternative means.

9. Set sample size to 16, SD to 2, and Significance Level to 0.01. Find the power and the effect size for each Alternative Mean.

Alternative Mean	power	effect size
0.4		
0.8		
1.2		
1.6		
2.4		

10. What happens to power as you increase the alternative mean? Explain why.

11. Look at the effect sizes in the last two tables. How do SD and Alternative Mean work together to determine power?

Try different significance levels.

12. Set sample size to 16, SD to 3, and Alternative mean to 1. Find the power and the effect size for each significance level (α).

α	power	effect size
0.01		
0.03		
0.05		
0.07		
0.10		

13. In which direction does power change when we increase the significance level?

Planning a new study

14. 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?

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

15. 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 a decrease in mean score (when people are sleep deprived) of 6 with 80% power. How large a sample is needed?

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

16. 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?

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

Take Home Message

- Errors happen. Use of statistics does not prevent all errors, but it does limit them to a level we can tolerate. We have labels for two types of error.
- The talk about probability of error is based on the sampling distribution 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.
- Questions? Summarize the lesson.

Assignment

- Read the next two pages before your next class.
- Watch the "Correlation and Regression" video (# 5 under Unit 2).

30 Correlation and Regression - Reading

Watch the “Correlation and Regression” video at ??

Important Points

- What type of plot is used to show the relationship between two quantitative variables?
- What does correlation measure?
- What are possible values for correlation?
- What symbol is used for population correlation? for sample correlation?
- What values for correlation indicate strong (weak) linear association?
- What symbols are used for the true parameters: intercept and slope in a regression model?
- What symbols are used for the sample statistics: intercept and slope of the least squares line?

- What is a residual? Which residuals are negative? positive?

- We found the estimated least squares line to be $\widehat{Temp} = 37.7 + 0.23 \times Chirps$
  - Interpret the slope.
  - What is the predicted Temperature if we hear 100 chirps per minute?
  - What is the residual if $y = 58$ and $x = 100$?

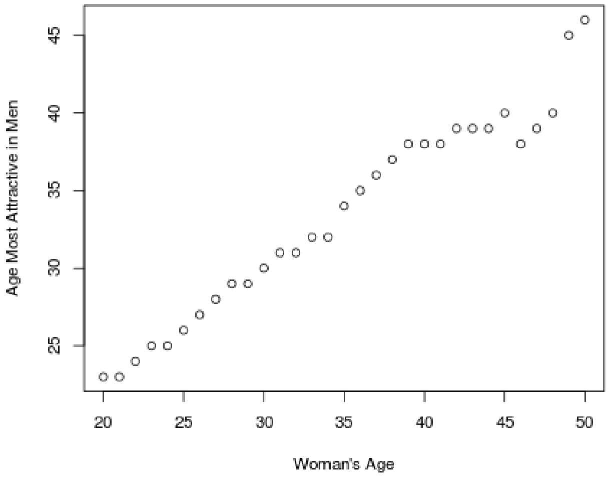
- Is it possible to have a negative slope and a positive correlation between two quantitative variables? Explain.

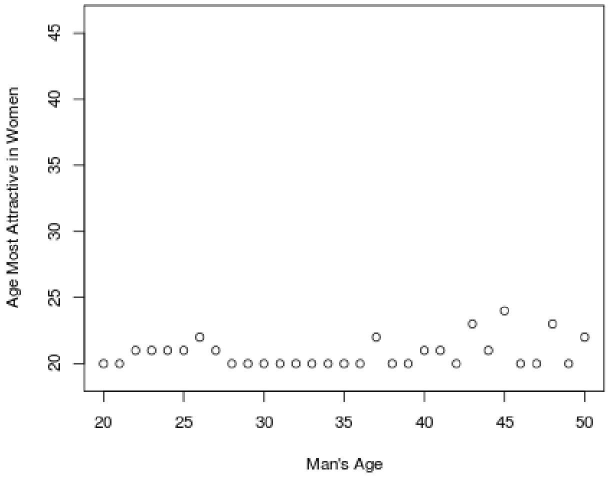
31 Attraction and Age

Or: Who Looks Good to You?

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

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





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

1. Suppose you're talking to a friend over the telephone, and you need to explain that the same two variables have a different relationship for women than for men. How would you describe the two plots?

2. What statistical summaries differ in the two plots?

3. As a review, in Algebra class you would have learned an equation for a linear relationship between x and y. What letters did you use for slope and intercept? What does "slope" mean? What does "intercept" mean?

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

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

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

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

4. Estimate the slopes for the two plots above.
   Hint: For each plot take a nice range of ages on the x axis, say from 20 to 40, and guess how much the line changes from the lower x value to the upper one. The change in y over change in x gives the slope estimate.

5. Interpret your estimated slopes.

31.1 Slope

6. In Algebra, a line is a collection of points satisfying an equation. In Statistics we start with data and have to find a good line to represent the relationship between two variables. When there is a lot of scatter in the points, many lines look reasonable. Go to `http://www.rossmanchance.com/applets/RegShuffle.htm` to see data on people's foot length versus height.
   (a) Is this a linear relationship?
   (b) Positive or Negative?
   (c) Strong, Moderate, or Weak?
   (d) Guess the correlation, then check with the button below the plot.

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

7. 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:

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

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

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

   (a) Was your slope too large? too small? about right?

   (b) What height does this line give for a person whose foot length is 0?

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

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

   (a) What are the coordinates of the added point?

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

   (c) Is correlation resistant to outliers? Is slope? Explain.

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

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

   (a) Which residual is largest? Find the (x, y) pair in the data table associated with that point.

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

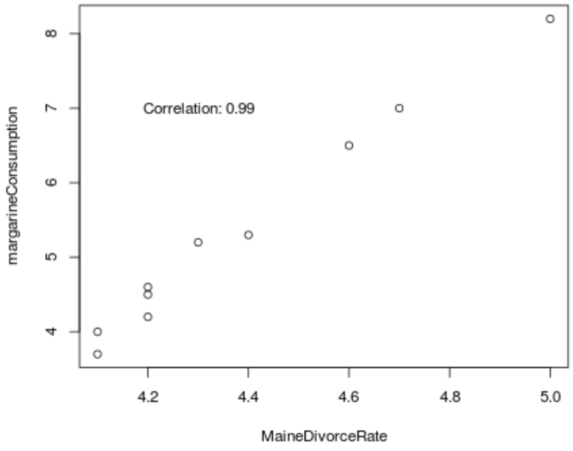
   (c) Now click Show Squared Residuals. These are important because we are using the "Least Squares" line. It picks slope and intercept to minimize the sum of all the squared residuals. Write down SSE (sum of squared errors).

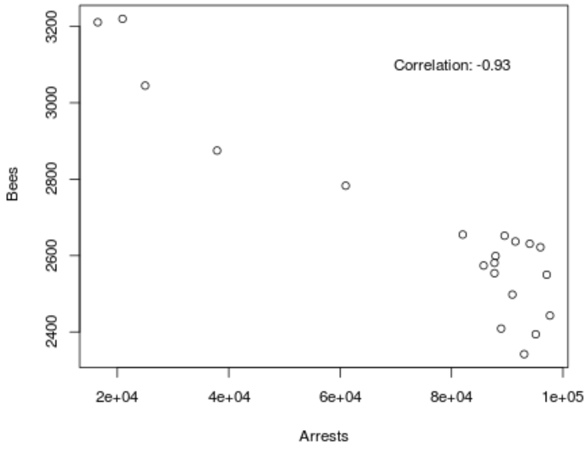
   Any other line will have larger SSE.

31.2 Correlation

11. **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).



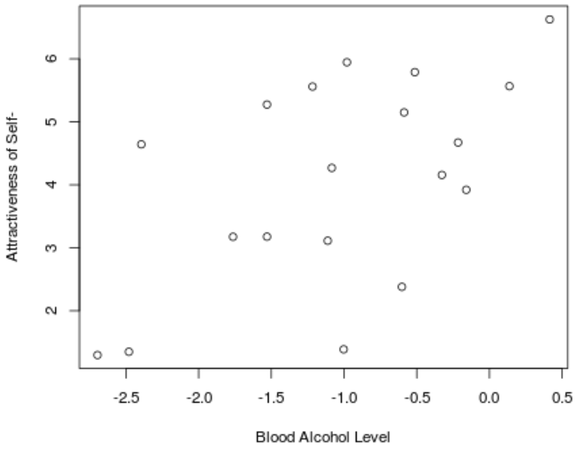


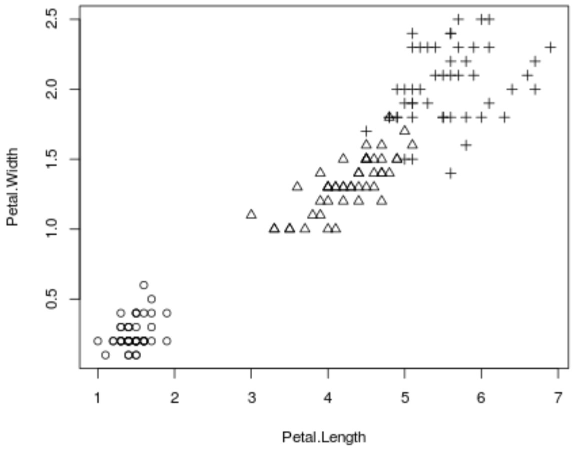
The point here is that if you search through lots of variables, you can find pairs that increase in the same way, or oppositely.

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

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

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





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

14. Look back at the Age-Attraction plots from OKcupid. Guess what those correlations are for women and for men.

15. 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?

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.

Assignment

- Fill in the simulation confidence interval box in column 2 of the Review Table.
- Read the next two pages before your next class.

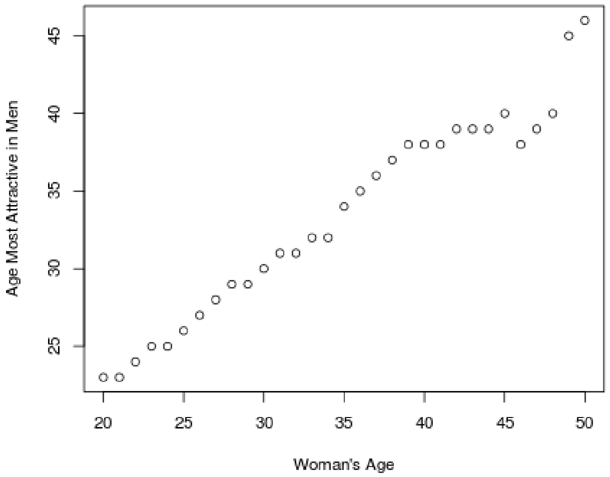
32 Testing Slope

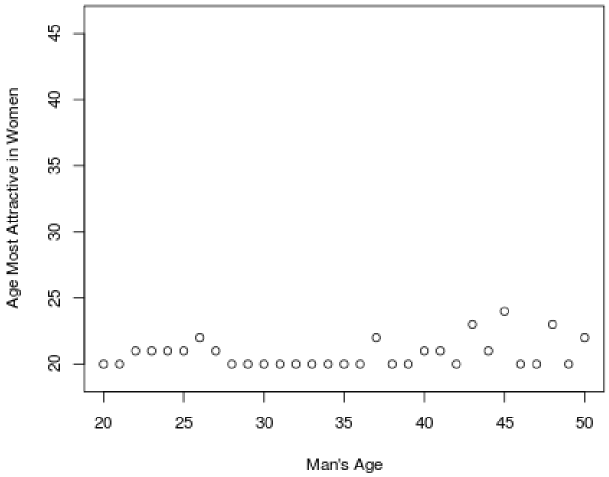
Watch this video: ??

- What is the usual null hypothesis for a test of slope? Be sure to use the right parameter.
- What is the alternative hypothesis for the example of taxi tips?
- What is the equation of the least squares line?
- If you are heading to NYC, what percentage would you expect to give a cabbie as a tip? (These data are for fares paid with credit cards. If we look at fares paid with cash, lots of people gave no tip. Apparently people doing 'business lunches' tend to use cards more, and they can include the tip as part of the expense, so they are more generous. Answer the question as if you are one of them.)
- What is the usual null hypothesis for a test of correlation? Be sure to use the right parameter.
- Explain the null hypothesis in your own words. What does it say about the relationship between the two variables?
- What is the alternative hypothesis for the example of taxi tips?

33 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. Use the least squares line to estimate what women aged 36.5 would say is the most attractive age of men?

2. Use the least squares line to estimate what men aged 49.5 would say is the most attractive age of women?

3. Discuss this alternative with your group. Perhaps the age of the men really doesn’t matter, and we’d be better off estimating their preference by using the mean “most attractive age for women” which is $\overline{y} = 20.77$ for all men, just ignoring the men’s age. Does that seem like a reasonable way to describe the second plot: “men of all ages find 20.8 years to be the most attractive in women”? Write down your group’s conclusion.

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

4. Go to the website `https://jimrc.shinyapps.io/Sp-IntRoStats` 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: $\widehat{mostAttrWomen} = 20.77 + 0\times\ mansAge$

   (a) Where have you seen the intercept estimate before (data input page)?

   (b) If you plug in any age for men, say 18 or 54, what result will you get from this line?

   (c) What does that say about the relationship between x and y?

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

5. If we want to test the null hypothesis "no linear relationship between men's age and the age of women they find most attractive", what is the null value of the true slope? Use β_1 as the true slope and fill in the hypotheses. Use a "right-tailed" alternative because we think ages should go up together.
   H_o :

   H_a

6. We need to do a simulation of possible slope values when H_0 is true. If there really is no connection between the two variables, x, and y, what kind of a "shuffle" could we do?

7. 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.

8. 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?

9. Take at least 1000 shuffles and compute the p-value. Explain which shuffles you are counting based on the plot provided and the original line.

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

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

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

    (a) State the hypotheses.
    H_o :

    H_a

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

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

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

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

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

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

14. Was some treatment randomly assigned?

15. What is the scope of inference?

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

Take Home Messages:

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

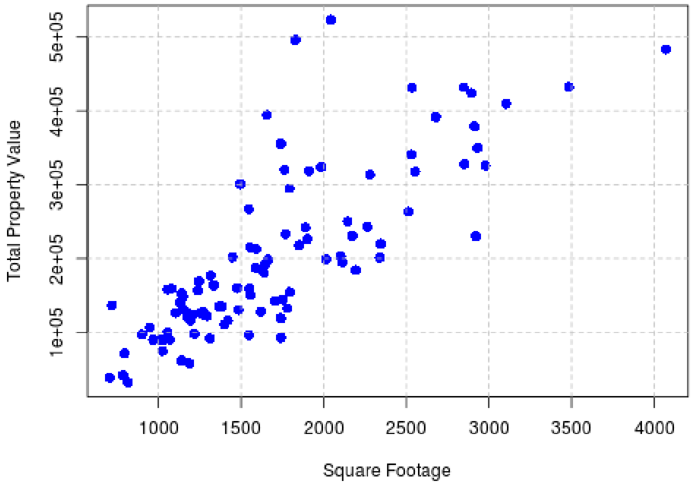
Assignment

- Fill in the top five boxes in the rightmost column of the Review Table.
- Instead of a reading for the next class, your assignment is to find an example of a simple linear regression model which is applicable to an area that interests you. Be ready to present it to your group: the variables, the estimated coefficients, and the meaning of the slope estimate. Do they provide any evidence that the slope is not zero?

?? suggested sites??

34 Regression Examples

1. One hundred homes[5] were randomly sampled from a database of all homes in Wake County, NC. We have the square footage of each home and an estimated (2008) value of the property. The plot shows 98 homes with value less than \$1 million.



2. Describe the relationship (linear or nonlinear? positive or negative? strong or weak?) and give a guess at the correlation between area and value.

3. The equation of the least squares line is:

$$\widehat{value} = -30748 + 137 \times sqft$$

   (a) Compute the fitted values for homes of size 1000 and 4000 sqft.

   (b) Mark the points on the plot and connect them to get the line of best fit. Note: `1.0+e05` means 1 times 10^5, or \$100,000.

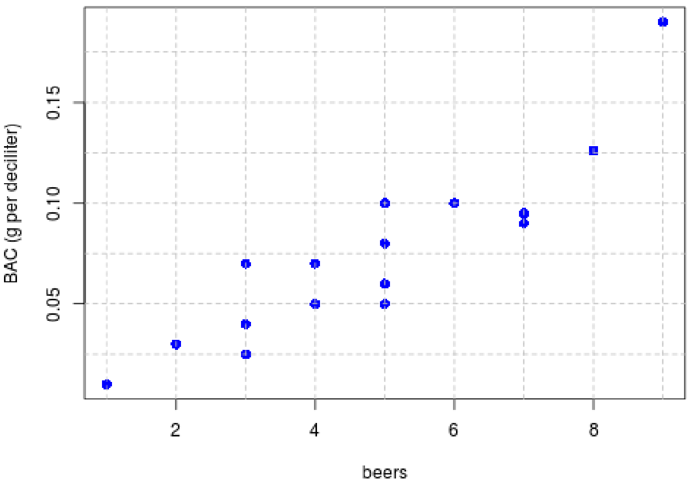
   (c) The observed value for a 3483 sqft home is 432516. Compute predicted price and find the residual.

[5] Dr. Roger Woodard, NCSU

(d) Two homes in particular do not seem to fit the linear relationship very well. Circle them and theorize: why might they have a value other than what the model predicts?

(e) Interpret the slope estimate.

4. Sixteen student volunteers[6] at Ohio State University were each randomly assigned a number of cans of beer to drink. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. First we'll look at a plot:



(a) Describe the relationship (linear or nonlinear? positive or negative? strong or weak?) and give a guess at the correlation between beers and BAC.

(b) The equation of the least squares line is:

$$\widehat{BAC} = -0.013 + 0.018 \times beers$$

i. Compute the fitted values for 2 and 9 beers.

[6] Diez, D.M., Barr, C.D., and Çetinkaya-Rundel, M. (2014). *Introductory Statistics with Randomization and Simulation* Exercise 5.28

ii. Mark the points on the plot and connect them to get the line of best fit.

iii. The observed BAC for the person who drank nine beers was 0.190. Compute the residual for the 9 beers person.

iv. The eighth beer drinker had observed BAC of 0.126. Compute that residual as well and explain why one residual is negative and the other positive.

(c) Interpret the slope estimate.

(d) Interpret the intercept estimate. Does its sign make sense?

Assignment

- Review for the exam

35 Unit 2 Wrapup

Vocabulary

- Random Sampling
- Principles of Designing experiments
- Response and Explanatory variables
- Random Assignment (why do we do it?)
- 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?
- Statistical significance is not the same as importance or practical significance.
- Interpretation of Confidence Interval
- Correlation, Slope. When are they useful? How are they interpreted?
- Type I Error probability of this error is limited by α, the significance level.
- Type II Error. Power is the probability we (correctly) reject the null hypothesis when an alternative is true. Power $= 1 - \beta$.
- What settings affect power of a study?

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

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

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

Extensions

2. Peanut Allergy Study

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

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

More Examples

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

3. Teen Hearing Study

   Headlines in August of 2010 trumpeted the alarming news that nearly 1 in 5 U.S. teens suffers from some degree of hearing loss, a much larger percentage than in 1988.[7]. The findings were based on large-scale surveys done with randomly selected American teenagers from across the United States: 2928 teens in 1988-1994 and 1771 teens in 2005-2006. The researchers found that 14.9% of the teens in the first sample (1988-1994) had some hearing loss, compared to 19.5% of teens in the second (2005-2006) sample.

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

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

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

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

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

4. Mammography Study

A mammogram is an X-ray of the breast. Diagnostic mammograms are used to check for breast cancer after a lump or other sign or symptom of the disease has been found. In addition, routine screening is recommended for women between the ages of 50 and 74, but controversy exists regarding the benefits of beginning mammography screening at age 40. The reason for this controversy stems from the large number of false positives. Data consistent with mammography screening yields the following table:[8]

	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?

(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?

(c) Estimate that error rate using these data.

(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?

(e) Estimate that error rate using these data.

If a patient tests positive for breast cancer, the patient may experience extreme anxiety and may have a biopsy of breast tissue for additional testing. If patients exhibit the symptoms of the disease but tests negative for breast cancer, this may result in the patient being treated for a different condition. Untreated cancer can lead to the tumor continuing to grow or spread.

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

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

5. Blood Pressure Study

   In a 2001 study, volunteers with high blood pressure were randomly assigned to one of two groups. In the first group – the talking group – subjects were asked questions about their medical history in the minutes before their blood pressure was measured. In the second group – the counting group – subjects were asked to count aloud from 1 to 100 four times before their blood pressure was measured. The data presented here are the diastolic blood pressure (in mm Hg) for the two groups. The sample average diastolic blood pressure for the talking group was 107.25 mm Hg and for the counting group was 104.625 mm Hg.

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

   (a) Do the data in this study come from a randomized experiment or an observational study? Explain.

   (b) Calculate the difference in the means.

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

   (d) Use our web app to do the appropriate test to determine if a difference this large could reasonably occur just by chance. Comment on the strength of evidence against the null model.

6. Investigators at the UNC Dental School followed the growth of 11 girls from age 8 until age 14. Every two years they measured a particular distance in the mouth via xray (in mm) . Assume that they want to test “Is the rate of growth zero?”. The data are 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.

   (b) How fast is this measurement changing?

   (c) What hypotheses are we testing?
   H_o :

   H_a

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

   (e) Give the scope of inference.