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% (= $13/50 \times 100\%$) of the emails were in html format.

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

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

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

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

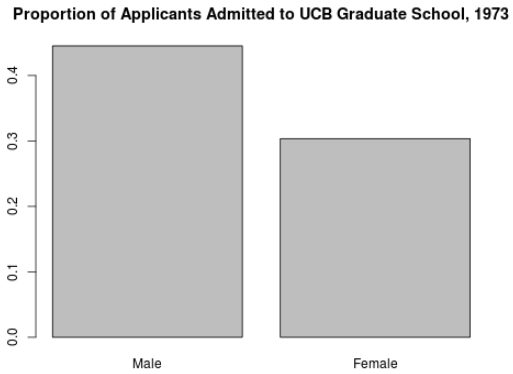
$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

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

2.2 Plotting Data

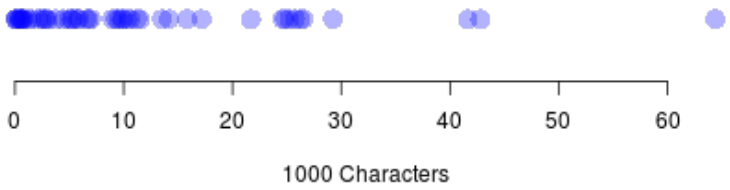
As with numeric summaries, the type of data determines the appropriate plot.

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.)

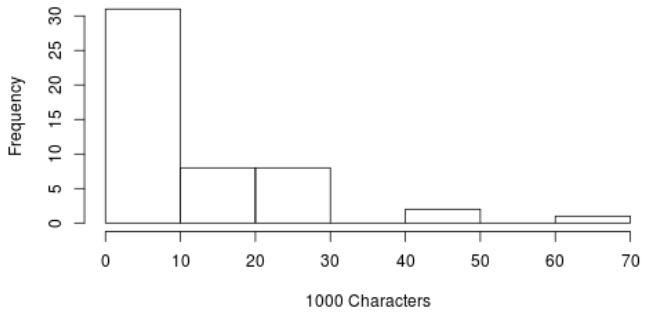


Quantitative variables are plotted with dot plots, histograms, density plots, and boxplots.

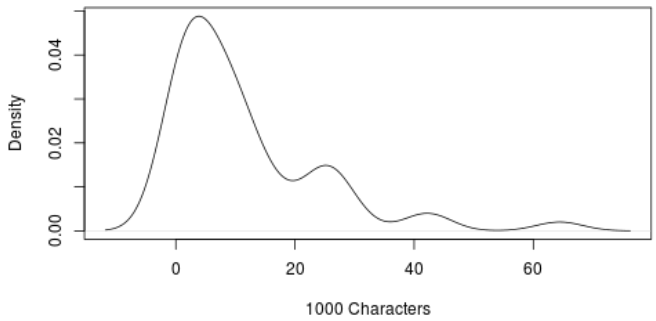
dot plots represent each point with a dot above the number line. This works well with small sample sizes. If the data are too close together to distinguish, we might stack them up to remove any overlap.



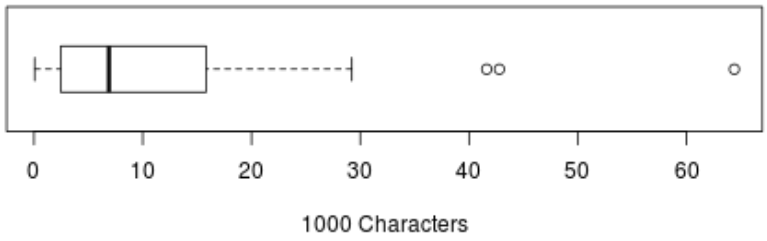
histograms divide the axis into "bins" and count the numbers of points falling into each bin. The height of each bin might show the count (frequency) of values in the bin or the proportion (relative frequency) for the bin. These plots work with moderate to large sized data sets. Choosing the best number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

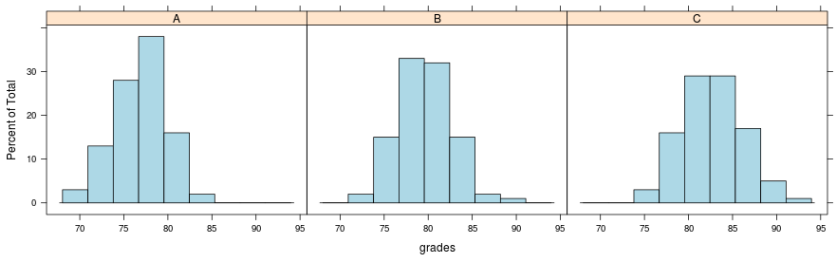
We also need to summarize categorical data, so we use proportions: the number in a certain category divided by the total number.

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

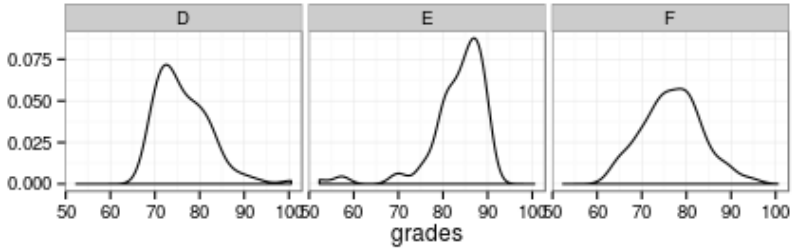
3.1 Comparing Distributions

Now we'll focus on quantitative data.

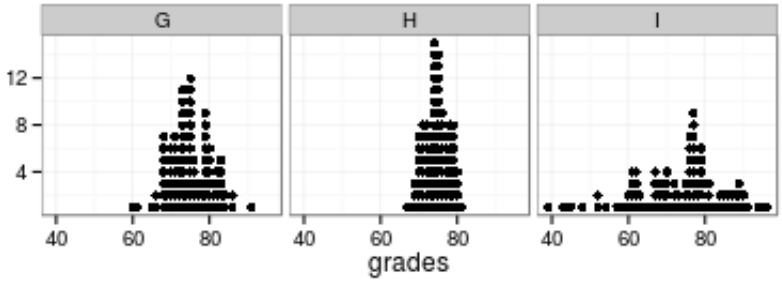
6. Suppose you are choosing which professors' class to enroll in. You have three choices, and have data on the grade distribution for each shown as histograms. Which class seems to have the best grade distribution? Explain.



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.



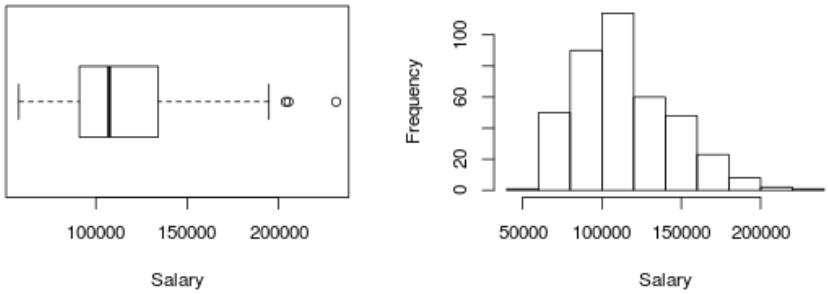
9. When comparing distributions there are several things to consider:

   - 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?

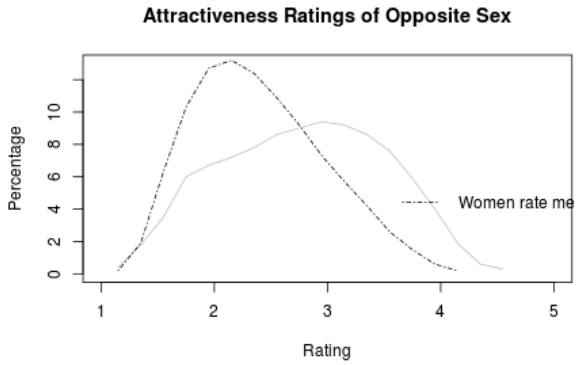
(b) Are any points flagged as outliers? If so, describe them.

(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 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 $\leftrightarrow$ 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.

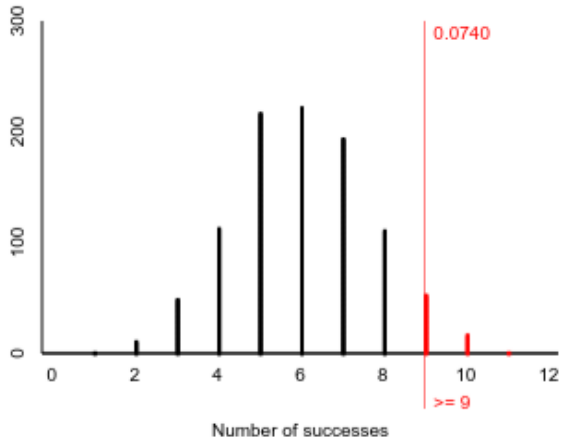


Figure 1: Simulation results obtained from the null model. The outcomes 9 and higher (74 out of 1000 trials) were as or more extreme as one group's number correct (of 12) and indicate the strength of evidence = 0.074.

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:

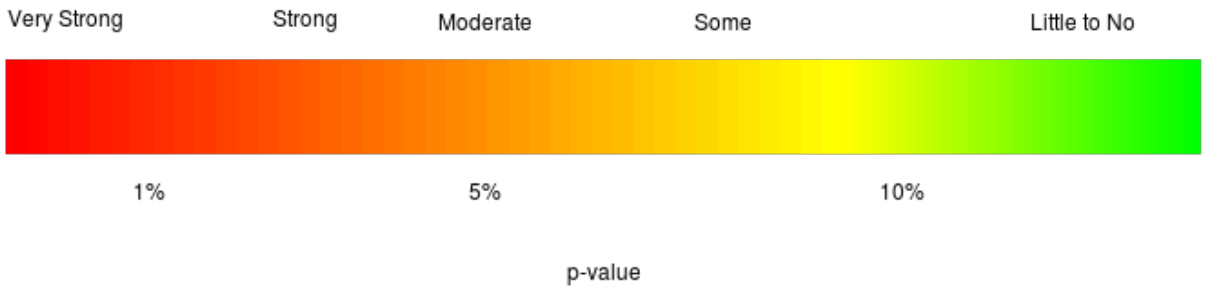


Figure 2: Numeric p–value and strength of evidence

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.

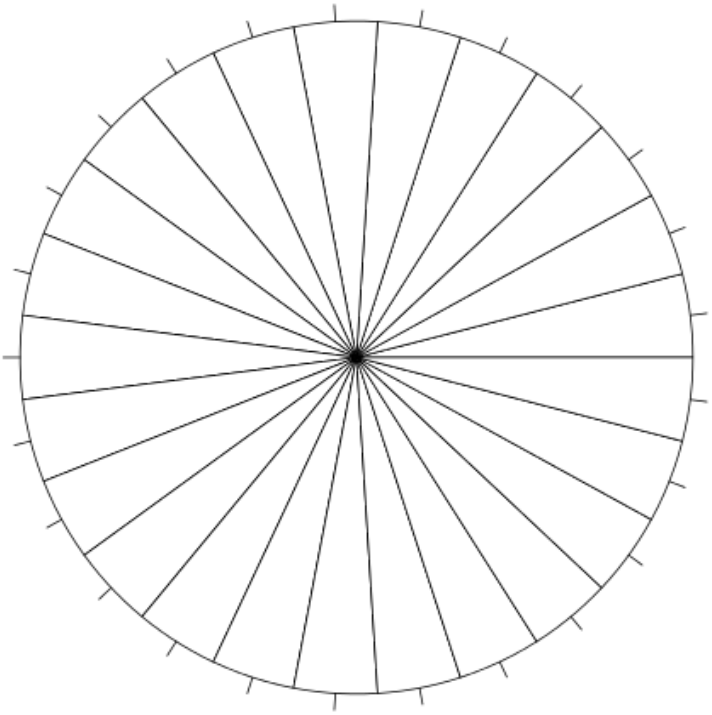
What is the "file drawer" effect?

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.

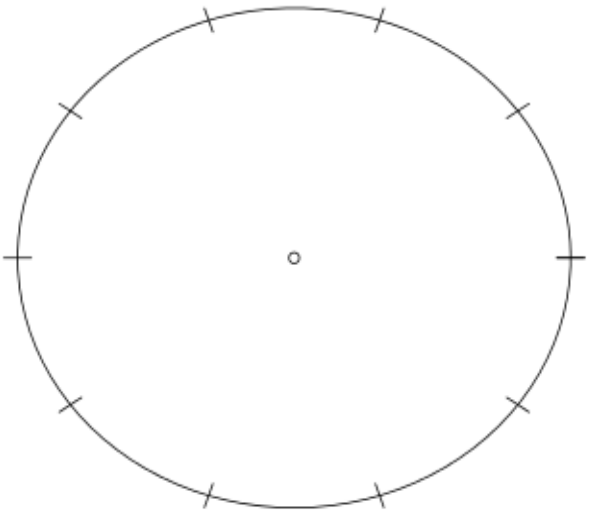
10. If Linzmayer is not just guessing and did have ESP, would you expect him to get a higher or lower proportion correct than the number from # 9? Use correct notation (an interval in parentheses) to denote this range of values.

    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 Test/Estimate under the One Categ menu. Enter the counts to show how many Linzmayer got right and got wrong. (These should add up to 25, but neither is 25.) Click "Use These Data" and record his proportion correct.

14. Now choose 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 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.

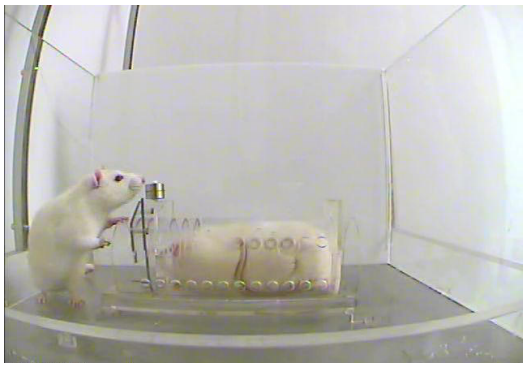
Watch this video:
`http://video.sciencemag.org/VideoLab/1310979895001/1/psychology`

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. This is a simple but powerful idea. By sampling **with replacement** we have the same conditions for every draw, and the probability of each outcome stays the same. Shuffle, draw a card, and record the outcome.

(c) Repeat until you have 30 outcomes chosen at random. What proportion of your rats were freed?

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

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?

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 it 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 80 and count: How many red points would you guess are in the left tail?
    How many reds in the right tail?
    How many blue points in the middle? Write the interval:

    (b) Click 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 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 numbers in each tail for, say, 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 One Categ in `http://shiny.math.montana.edu/jimrc/IntroStatShinyAp` and then select the Test page. Do we need to change the "Null value" for p?

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

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

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

(d) Now estimate the true population proportion.

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

ii. In order to generate simulated data,

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

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

C. Each couple leans right with what probability?

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

iii. Use the web applet to create 1000 or more resamples from the original data.
   A. Where is this distribution centered?

   B. What number describes the spread of the distribution?

iv. Compute a 99% confidence interval.

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

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