Stat 216 Syllabus – TEAL Sections

Sections Meeting Tu and Th

Note: This section meets in a **TEAL**, Technology Enhanced Active Learning classroom, (Wilson 1-119, 1-133, or Gaines 143). The curriculum is designed for you to work in groups, although the group work does not directly impact your grade. Every component of your grade is corrected and graded individually. There are other sections of Stat 216 which cover the same material using a more traditional book + lecture methods, but all sections will use some group work.

People

- Your Instructor: (Write contact info here)

- Student Success Coordinator: Jade Schmidt
  email: roskam@math.montana.edu Office: Wilson 2-260 406-994-5357

- Course Supervisor: Dr. Robison-Cox
  email: jimrc@math.montana.edu Office: Wilson 2-241 406-994-5340

Course Materials

- Buy the Stat 216 - TEAL - Coursepack from the MSU Bookstore and put it in a three ring binder. Do not buy a textbook for this section.

- Download the free textbook *Intro Stat with Randomization and Simulation by Diez, Barr and Cetinkaya-Rundel from*
  http://www.openintro.org/stat/textbook.php?stat_book=isrs

- *We will also use various web sites and readings.*

- *"Quizorks" (our word for very important homework sets) will be downloaded from D2L, so be sure you can log in to the MSU D2L system:*
  https://ecat.montana.edu/. If you have problems, view the help on that page.

The syllabus continues after this page relating to our inclass activity on Day 1. Read the syllabus for Day 2.

Detecting Fraud

If you look at the first digits in a bunch of numbers, for instance, values on the 1040 tax form, the distribution of digits 1 through 9 is rather odd. There are more 1's than 2's, more 2's than 3's, etc, with 9 occuring least often. Accountants at the IRS can use their knowledge of "Benford's Law" to identify people who are making up numbers to fill in the form.

Your instructor claims to know something you don't know about sequences of H's and T's (heads and tails) from coin flips. As a group of three, we want you to make up a sequence of 45 H's and T's and write it down. This will be your "fraudulent" sequence. Write it here.

Also, we need you to create a sequence of 45 H's and T's by spinning or flipping a coin (or coins – you could each do 15 and combine them to get 45). This is your "true" sequence. Write it here.

The whole class will need to see your sequences either on the whiteboard or on the document camera in shorthand. We are going to guess which one is the true sequence, so flip a coin and give the first one the "A" label if you get heads (if tails, second is "A"), and label the other "B". To save time, write 3H instead of HHH, and stack each sequence vertically. Copy both sequences to the previous page. We will need them in a few days. Label them so you know which one was made up.

Your instructors will tell you how we'll see if they can "detect fraud".

While waiting for the other groups, discuss: "What does random look like?" and "How can we tell if a sequence is random?"

At the end of class or soon after class, write up the skeleton of our reasoning. What two explanations could there be for the instructor having a high score? What sort of outcome on the instructor's part would lead you to conclude that she/he knows something you don't? You can't judge the instructor's score without some background for comparison. What is the background? Do you think students know something, too, or are blindly guessing?

Syllabus

Learning Outcomes for STAT 216

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

CORE 2.0: This course fulfills the Quantitative Reasoning (Q) CORE 2.0 requirement because learning statistics allows us to disentangle whats 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. If further assistance is necessary, contact the Course Coordinator, Dr. Robison-Cox.

Course Description

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

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

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

Prerequisites

This course is intended for undergraduate students who have completed the equivalent of a 100-level math course with a grade of C- or better, but not previously studied statistics. If you have previously taken Stat 216 or an equivalent course elsewhere, we recommend that you transfer to a different section. However, if you have a fervent desire to stay in this class and learn under this different instructional approach, please speak with your instructor.

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 created by the course coordinator for most in-class activities and homeworks. The website for this applet is `www.math.montana.edu/~jimrc/randomization` . We will also be using a web applets in Units 2 and 3 from the websites `http://www.rossmanchance.com/applets/index.html` and `www.lock5stat.com/statkey`.
- **Technology Policy**: This course utilizes technology extensively. Bring a laptop or tablet to class (your phone might get you by for web apps, but you'll also need a word processing program). You will need at least one laptop within your group each day.

Math Learning Center in 1-112 Wilson Hall is a very important resource. During the semester students may get help on Stat 216 topics at these times:
9am – 5pm and 6pm – 9pm Monday through Thursday, 9 am – 2 pm Friday.
We strongly encourage you to check out what's offered there (preferably before an exam is looming).

Assessment

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

- **Quizorks: 30%** Twice per week assignments of problems that will help you learn the course material and software through reflection and practice and are essential preparation for the exam.

  Format: Your instructors will tell you if you submit these as electronic files uploaded to D2L or as hard copies. If electronic, it needs to be in a format we can read. Adobe pdf is our standard. Submissions we can't read will not count.
- **Exams: 30%** There are two in-class individual (on your own, not with your group) exams, which count as 15% each. You will be allowed to bring in one page of handwritten notes for in-class exams.
- **Final Exam: 25%**. This exam will be cumulative in content. Again, you will be allowed to bring in one page of handwritten notes for the final exam.

- **Attendance/Participation/Preparation: 15%** . Class participation is an important part of learning, especially in courses like this one that involve group cooperation.

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

  Preparation: The in-class activities and out-of-class assigned readings are the primary source of information for this course. Take them seriously, work through them with care, and they will be very valuable on exams. As a way to provide further emphasis to the activities and readings, most classes will begin with a Readiness MiniQuiz 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 (90–100 is an A, etc.) Cutoffs for plus and minuses will be determined later.

Some Department Policies:

- Do not attempt to turn in any assignment in the math office. They will not be accepted.
- Do not call or email the math office for information on grades.
- Do not attempt to pick up your final exam in the math office.
- Your final exam can be mailed to you if you provide your instructor with a self-addressed envelope and two stamps during the semester.

University Policies and Procedures

Behavioral Expectations

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

Collaboration

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

Plagiarism

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

Academic Misconduct

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

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

Academic Expectations

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

A. be prompt and regular in attending classes;

B. be well prepared for classes;

C. submit required assignments in a timely manner;

D. take exams when scheduled;

E. act in a respectful manner toward other students and the instructor and in a way that does not detract from the learning experience; and

F. make and keep appointments when necessary to meet with the instructor. In addition to the above items, students are expected to meet any additional course and behavioral standards as defined by the instructor.

Withdrawal Deadlines

September 15, 2014 is the last day to withdraw without a "W" grade. University policy is explicit that after this date, the advisor and instructor must approve requests to withdraw from a course with a grade of W.

Group Expectations

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

We do hear students worry that their grade will get lowered if some in the group don't put in their fare share of work. The grades we record are all based on individual assignments and exams. Groups do not all get the same grade. Your table will have more than one group, so if some of your group members do not participate in what we hope will be interesting discussions, you can tune in and learn from a neighboring group. We're all in this together, and want you to enjoy the benefits of collaboration without worrying about harmful effects of "slackers".

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 making inference from a sample to the population.

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

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

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

> Four college friends were so confident that the weekend before finals, they decided to go to a city several hours away to party with some friends. They had a great time. However, after all the partying, they slept all day Sunday and didn't make it back to school until early Monday morning.
>
> Rather than taking the final then, they decided to find their professor after the final and explain to him why they missed it.
>
> They explained that they had gone to the city for the weekend with the plan to come back and study but, unfortunately, they had a flat tire on the way back, didn't have a spare, and couldn't get help for a long time. As a result, they missed the final.
>
> The Professor thought it over and then agreed they could make up the final the following day. The four were elated and relieved.
>
> They studied that night and went in the next day at the time the professor had told them. He placed them in separate rooms and handed each of them a test booklet, and told them to begin.
>
> They looked at the first problem, worth 5 points. It was something simple about exploratory data analysis. ”Cool,” they thought at the same time, each one in his separate room. ”This is going to be easy.”
>
> Each finished the problem and then turned the page. On the second page was written:
>
> For 95 points: Which tire?

2. Explain your method of selection. How did you choose your ten words? *Answers will vary. Some will be more representative than others.*

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

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

STOP!

Give your sample means to your instructor.

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

   on target = unbiased.

   Then the mean of the distribution matches our true parameter.

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

Simple Random Sampling

6. Point your browser to

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

7. Click Clear under the data box to erase all the data.

8. Copy the word data from this page: http://www.math.montana.edu/~jimrc/classes/stat216/data/jokeData.txt. You can select the entire file with control-A, copy it to the clipboard with control-C (or use the right mouse button to copy) and paste it into the data box (use control-V or the mouse option). Click Use Data, then select variable length. You should then see Population size change to 241 and a histogram of length. This is our population of 241 words.

9. Click the box next to "Show Sampling Options" set Number of samples to 1 and Sample Size to 10. Click Draw Samples. A box then shows which 10 words were selected and two plots appear. The first shows the 10 sampled lengths, the second shows just their mean (or median if you select that, don't worry about the t-stat now). Write down the 10 word lengths in this sample. *AWV*

10. Record the average word length for the ten randomly sampled words. Remember, the sample average is an estimate of the true mean word length in the population. This value should appear at the top of the data plot and as a point in the right-hand plot as well. *AWV*

11. Click Draw samples again and record the next average. *AWV*

12. Change Number of samples to 1000 and click Draw Samples three times. How many sample averages are now represented by the histogram? Record the mean and standard deviation (SD) of all the sample means. (See top of the right–most plot. The SD is a measure of the average spread of the points away from the mean.) *Should be close to 4.257 for mean, 0.60 for st.dev*

13. If the sampling method is unbiased, the estimates of the population average should be centered around the population average word length of 4.257. Does this appear to be the case? Describe what you see. *Should see a fairly symmetric distribution about the center (mean close to 4.257). Mine goes 2.5 to 7*

14. Return to number 4 where you took your own sample and found its average. Put that sample mean into the empty box below the plot next to Count Samples Greater than and click Count. How many of the thousand random samples had a mean this large or larger? Is this a typical value for the distribution shown, or is it unusual? Explain. *AWVF. We hope most people had a high number like 6 which shows up in only 19 of my 3011 samples (.6% of the samples). This is quite unusual.*

15. **Class Samples**: Now your instructor will display the estimates from each person in the class. Sketch the plot of all of the sample estimates. Label the axes appropriately. *Hope to see some bias here. Discuss estimates close to 4.25.*

16. The actual population mean word length based on all 241 words is 4.257 letters. Where does this value fall in the above plot? Were most of the sample estimates around the population mean? Explain. *Expect them to say: No, we got fooled into picking the larger words.*

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

18. Based on your answer to question 17, are "by eye" sample estimates just as likely to be above the population average as to be below the population average? Explain. *No, they are biased to generally be larger.*

19. Compare the web applet plot from question 12 with the plot from 15. Which method is closer to being **unbiased**? Explain. *Random sampling should win the day here. It is unbiased.*

Examining the Sampling Bias and Variation

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

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

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

Even when an unbiased sampling method, such as simple random sampling, is used to select a sample, you dont expect the estimate from each individual sample drawn to match the population mean exactly. We do expect to see half the estimates above and half below the true population parameter.

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

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

20. Back in the web applet, change sample size from 10 to $\boxed{25}$ and click $\boxed{\text{Reset}}$. Draw 3000 random samples of 25 words, and write down the mean and standard deviation of the values. *AWV*

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

22. Record the mean and standard deviation of the sample averages. *AWV*

23. Does the sampling method still appear to be unbiased? Explain. *Yes, because the distribution is centered at the true mean.*

24. Compare and contrast the distribution of sample estimates for $n = 10$ and the distribution of sample estimates for $n = 25$. How are they the same? How are they different? *Same in that both are centered at 4.257. Different in that the st.dev is larger for* $n = 10$ *(it is 0.60) than for* $n = 25$ *(0.38).*

25. Compare the spreads of the plots in 12 and 21. You should see that in one plot all sample means are closer to the population mean than in the other. Which is it? Explain. *Sample size 25.*

26. Using the evidence from your simulations, answer the research question: does changing the sample size impact whether the sample estimates are unbiased? *Yes, as sample size gets bigger, the st.dev goes down.*

Population Size

Now we examine another question:

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

27. Increase the size of the population in our sampler by selecting the button on x4 instead of the default x1 to make four copies of the data instead of just one. What is the new population size?

    964

28. With sample size set to $\boxed{25}$, draw a few single samples to see if they look similar, then draw 3000 random samples and record the average (mean) of all the average word lengths. *AWV*

29. Sketch the plot of the sample estimates based on the 3000 samples drawn. Label the axis appropriately. *Hope to see center and spread have not changed much.*

30. Record the mean and standard deviation of the sample averages. *AWV*

31. Does the sampling method still appear to be unbiased? Explain. *Yes. It's centered at about 2.457.*

32. Compare and contrast the distribution of sample estimates for $n = 25$ now that you are sampling from a larger population to the distribution of sample estimates for $n = 25$ from before. How are they the same? How are they different? *Means of the two distributions are essentially the same, so also is the st.dev.*

33. Try it one more time with the x40 button checked. Again use 3000 samples of size 25. How similar are the mean and SD of the distributions? *Means of the two distributions are essentially the same, so also is the st.dev.*

34. Use the evidence collected from the simulation to answer the research question: does changing the size of the population impact whether the sample estimates are unbiased? *No. The sample mean for 25 observations has roughly the same mean and st.dev as it did with 241 in the population.*

35. When we actually collect data, we only get a single sample. In this exercise, we started with a known population and generated many samples. How did we use many samples to learn about properties of random sampling? *We sampled over and over to see how variable our statistics will be. We compared SE's for different sample sizes and population sizes.*

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

Summary

- 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).

Cell Phone Data

Do others use their cell phones as much as you? Or do they use them in different ways?

We want to informally investigate how people use their cell phones, so you were asked to bring them to class today.

What data should we examine? A phone has lots of info stored, so this is quite open ended. We also need some data to describe you. Some possibile quantities are listed below. As a class, we'll discuss which would be most interesting to examine.

- Your gender.
- Are you over 25 years old?
- Are you in–state? or paying out–of–state tuition?
- Do you pay the phone bill, or does someone else?
- Do you have a data plan with the phone?
- What company do you use for phone service?
- On a scale from 1 (very unhappy) to 10 (very satisfied), how satisfied are you with the service?

From your phone's memory or system log:

- How many calls did you make yesterday?
- How many calls did you receive yesterday?
- How many text messages did you send yesterday?
- How many text messages did you receive yesterday?
- Do you store pictures you've taken on this phone? If so, how many in the last month?
- Do you receive and store pictures from others? How many in the last month?

In choosing variables, think about which comparisons are interesting. What differences might their be between men's and women's cell phone usage?

Cell Phone Answers

1. Write down your answers to the questions selected.

   To record data for the whole class, we might make a table of data with a row for each of us and a column for each question, or "variable". Organize your group's data that way. Use an abbreviated column header to remind us which variable is in each column.

2. Separate the variables into categorical and quantitative.

3. As a group, define what makes a variable categorical or quantitative.

   According to your rule, is zip code categorical or quantitative? Explain why.

4. We need a plot to compare one qualitative aspect of phone usage for two or more groups of people. As a group discuss what type of picture or plot we can use. As a class, combine your data and copy the plot here.

5. Compare the distributions across categories. How do we differ, or are they about the same?

6. Think about the quality of the data.
   Did everyone count messages (phone or text) and pictures in the same way?
   Discuss within your group and write down an explanation of what was counted.

7. How would you plot data from two categorical variables, say for example: gender, and in/out of state? Show what we might get here:

8. An important aspect of statistics is to take information from a small sample and use it to figure out what is going on in a larger population. We use a **statistic** from the **sample** to estimate a **parameter** from the **population**. That works well if the sample is just like the population, or is **representative** of the population.

   (a) Is your group representative of this whole class?

   (b) Is this class representative of all Stat 216 students?

   (c) Is this class representative of all MSU students?

   (d) List some variables which might differ (on average) between this class and the general MSU population of students.

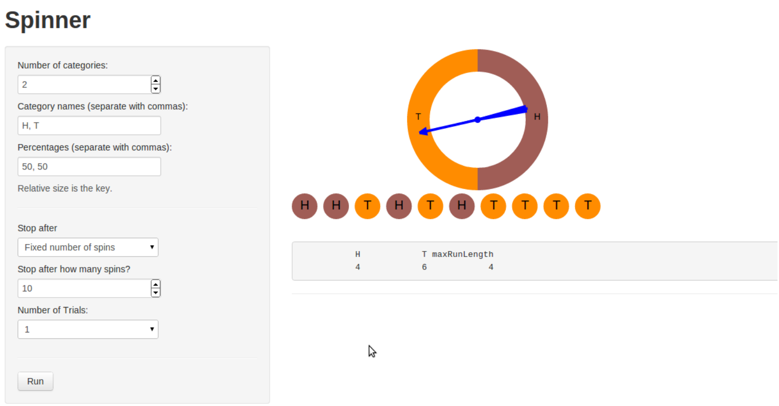
Take Home Message:

Collecting high quality data is tough work. We have to define questions carefully and must agree on how to measure variables. The type of variable determines how we work with and visualize it. (Plots are very important.) To compare distributions we look at center (mean or median) and spread. Distributions can also show lots of skewness, and we might have some extreme outliers to deal with. We saw on Day 2 that selecting a sample is also difficult, and our class might not be representative of all MSU students.

Patterns in Random Behavior

Modeling Coin Flips

On the first day of class, we spun or flipped a coin 45 times. We can use this web app to do the same job more quickly: `http://spark.rstudio.com/jimrc/spin/`
For example, here we asked it to spin either an "H" or a "T" with probability .50 for each. Note that just because there are two options, does not mean that they have to be equally likely. We could make H appear 75% of the time and T only 25%.



Spinner

Number of categories: 2

Category names (separate with commas): H, T

Percentages (separate with commas): 50, 50

Relative size is the key.

Stop after: Fixed number of spins

Stop after how many spins? 10

Number of Trials: 1

```
        H          T maxRunLength
        4          6            4
```

1. Change the inputs to use $\boxed{2}$ categories labeled $\boxed{\text{H,T}}$ with equal probability. Tell it to stop after $\boxed{45}$ spins and leave it on $\boxed{1}$ trial. What do you get as the maximum run length?

2. Can you see where the maximum run of H or T occurred? Was it H's or T's?

3. Now we want to see what happens when we do this more than once. Change $\boxed{\text{Number of Trials}}$ to 100 or 1000.

4. When you do this, the app then needs to know what summary of the data to save from each trial. Choices are:

   - Count in first category, (here H because H is first in input box 2)
   - Count in second category (here it's T), or
   - Maximum run length.
     In the figure above, we had a run of 3 tails, then a run of 4 heads, and finally 3 tails to give 10 spins. The maximum run length for this trial was four.
     The smallest maximum run length possible is one. Write a sequence of five heads and tails with max run length = 1. Show 3 sequences, one from each group member. *HTHTH or THTHT and repeat one again.* Write as many different sequences of five heads and/or tails as possible that have the largest possible maximum run length (five in this case). *TTTTT or HHHHH, that's it.*

In simulation experiments after a trial is carried out, a result is recorded. A result is simply a numerical summarization of the outcomes in a particular trial. We might want to look at the maximum run length in 45 coin flips.

5. Go back to the sequences you created on Day 1.
   What was the maximum run length for the "fraudulent" sequence? *AWV, hopefully smaller than that of the true sequence.*
6. What was the maximum run length for the "true" coin flip sequence? *AWV, hopefully larger.*
7. Sketch the plot from the web app (using 1000 trials) and show where these two values fall. Is one more in the center of the distribution than the other?
8. How do you think the instructors decided which sequences were fraudulent? Based on the maximum run lengths we actually see in coin flips, how would you now make your guesses? *AWV, Just picking the sequence with larger max run length will often work.*

Intuitions About Dice

Imagine that you are rolling regular a six-sided die 10 times. In those 10 rolls, consider the number of times that you would expect to see the outcome of three.

Imagine repeating this process 100 times.

9. Which outcome (zero 3's to ten 3's) would you expect to occur most often? *AWV – hopefully less than 5. Expected count is 10/6 = 1.67.*
10. What percentage of the time would you expect to see an outcome of five 3's? Explain. *Rarely, because 1 3 is slightly rare, so seeing lots of 3's is quite unusual. The exact binomial mass is about 0.013.*
11. Which outcome, two 3's or eight 3's, would you expect to see more often? Why? *It's hard to get one three in one roll, so I expect to get low numbers out of 10. 2 is more likely than 8. If they say "equally likely" they are missing the fact that "3" and "non-3" are not equally likely.*
12. What percentage of the time would you expect to get an outcome of all ten 3's? *one per billion? precisely it's 1.653-08*

Modeling Dice Rolls

Now you will use the "Mixer" web app `http://spark.rstudio.com/jimrc/mix/` to simulate rolling a die 10 times. We will use the data you generate in the simulation to check your initial intuitions about the number of threes that would occur in 10 rolls.

Setting Up the Model: At the above web site create a box with six balls numbered 1 through six in it. Drawing a ball numbered four is equivalent to rolling four on a die.

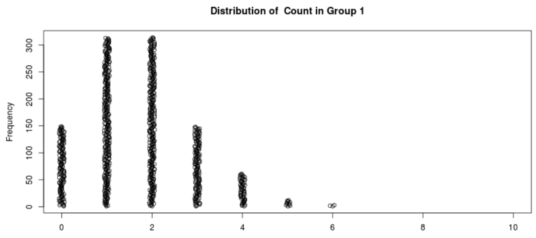
- Change number of categories to 6
- Change category names to 3,1,2,4,5,6
- Change numbers of each type to 1,1,1,1,1,1 to get fair dice.
- Leave Replace each draw? set to yes. If it were "No" we'd be taking away each outcome as it occurs.
- Change Stop after how many draws? to 10. (Leave other choices the same. If you happen to change the "Stop after" choice, switch it back to Fixed number of draws.)
- Click the Run button.

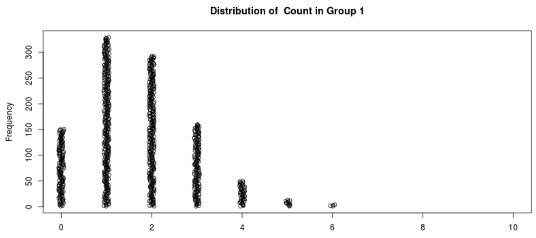
In this simulation, the trial consists of rolling the die 10 times. The result from each trial is the number of 3's that occur.

13. Record the number of 3's that occurred in the trial in a case table (the summary output table found at the bottom of the mixer applet).
Carry out nine more trials. Record the results from each trial into your case table.

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
3	2	0	0	2	1	2	1	1	2

14. Create a plot of 1000 results and sketch it here. Run it again, sketch and compare. How similar are they?





15. Which outcome (zero 3's to ten 3's) occurred most often? *one*

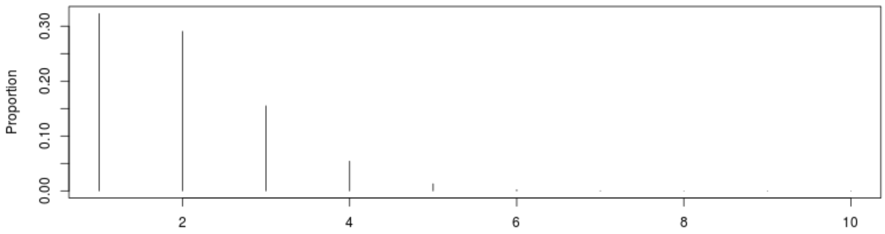
16. What percentage of the time did an outcome of five 3's occur? *0.13%*

17. Which outcome, two 3's or eight 3's, occurred more often? *two*

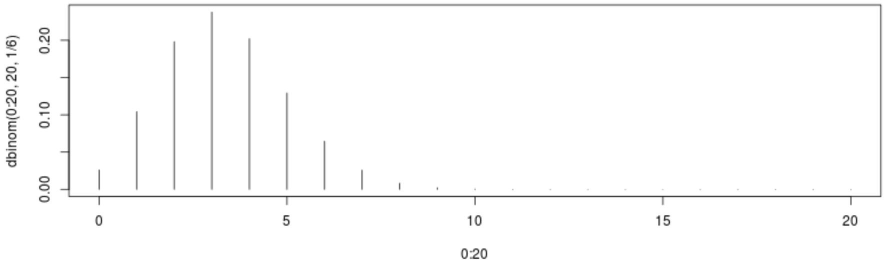
18. What percentage of the time did an outcome of seven 3's occur? *zero*

Extensions

19. Sketch the plot you would expect to see if you could simulate the results from 10,000 trials (rather than 1000) of the die rolling experiment. *AWV, here's the ideal:*



20. Sketch the plot you would expect to see if you could simulate the results from 10,000 trials of the dice rolling experiment, but instead of rolling the die 10 times, you rolled it 20 times. *AWV, here's the ideal:*



21. In the Quizork for today we asked you to view a video about Galton's Board. We don't show an animation of balls falling down, but we can use the spinner or the mixer to get a pattern very similar to the one the video ends with. Challenge: try different settings using 2 categories to see if you can get that nice symmetric distribution. Hints: you need a fairly large number of spins or draws, at least 20, and lots of trials, like 1000. Have it keep track of the count in group 1 or in group 2, not run length. Record your settings here.

22. If time allows: Next class we will simulate the genders of children born into a family. If gender is random and knowing the gender of one baby does not help predict the gender of another baby born to the same couple, would you use a spinner or a mixer to set up a simulation of births? How would you set it up?

One Child? – One Son? – One of Each?

According to Wikipedia,

"The one-child policy, officially the family planning policy, is the population control policy of the People's Republic of China. Many demographers consider the term "one-child" policy a misnomer, as the policy allows many exceptions: rural families can have a second child if the first child is a girl or is disabled, and ethnic minorities are exempt. Families in which neither parent has siblings are also allowed to have two children. Residents of the Special Administrative Regions of Hong Kong and Macau, and foreigners living in China are also exempt from the policy. In 2007, approximately 35.9% of China's population was subject to a one-child restriction. In November 2013, the Chinese government announced that it will further relax the policy by allowing families to have two children if one of the parents is an only child.

"This policy was introduced in 1979 to alleviate social, economic, and environmental problems in China. Demographers estimate that the policy averted 200 million births between 1979 and 2009. The policy is controversial both within and outside China because of the manner in which the policy has been implemented, and because of concerns about negative social consequences. The policy has been implicated in an increase in forced abortions, female infanticide, and underreporting of female births, and has been suggested as a possible cause behind China's sex imbalance. Nonetheless, a 2008 survey undertaken by the Pew Research Center reported that 76% of the Chinese population supports the policy. [1]"

Scholars have wondered how things would change if instead of a one-child policy, a country adopted a one-son policy. A "one son" policy would allow families to keep having children until they had a son. If a family's first child is a boy, they stop having children. If the first child is a daughter, they can try again. They can continue having children until they have a son and then they stop having children.

Discuss:

1. According to the 2000 Census, the average number of children per family in the United States is 1.86.[2] If the United States adopted (and enforced) this "one son" policy, how do you think the average number of children per family would compare to the current average? Explain your reasoning.

   AWV, They might guess that half the families will have 1 kid, 1/4th will have 2, 1/8 will have 3, ... so the average seems to approach 2

Modeling family size under the "one son" policy

[1] One-child policy. (2013, Dec 14). In Wikipedia, the free encyclopedia. Retrieved Dec 18, 2013, from http://en.wikipedia.org/wiki/One-child_policy

[2] Of families with children. Including couples without children brings it down to 0.90.

Your group will try to answer the research question by conducting a simulation study. Write a brief description of the model that you can use to generate data to answer this question.

2. Model: Describe the model you will use to generate outcomes (what are the potential outcomes; sampling with or without replacement; probabilities of the potential outcomes; etc.);

   Assume that each pregnancy is equally likely to produce a boy or a girl, and that birth genders are independent of each other. We will sample with replacement so that the probability stays at 50/50%.

3. Based on your description of the model, set up the model in the web app using a spinner or mixer. Describe the inputs you gave for the web app.

   Use the spinner with outcomes $\boxed{\textit{B,G}}$ *or* $\boxed{\textit{M,F}}$ *and percents set to* $\boxed{\textit{50,50}}$. *Or use the mixer with the same categories and equal counts, like* $\boxed{\textit{5,5}}$ *BUT then we must set "Replacement" to* $\boxed{\textit{yes}}$.

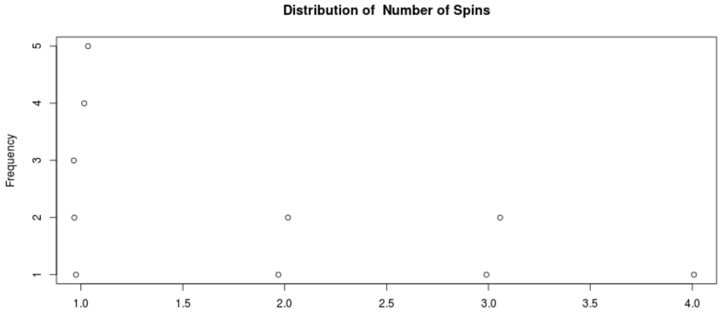
4. **Defining a Single Trial**

   Unlike previous simulations, a trial in this simulation is not defined by a set number of outcomes. Rather, a single trial will end when "a male is born." We need to use a stopping rule rather than a fixed number of outcomes. The spinner and mixer apps both have a "Stop after" choice which should be set to $\boxed{\text{One spin in 1st category}}$ or $\boxed{\text{One draw in 1st category}}$

   (a) Carry out a single trial of the simulation. How many kids did it take to get one boy?
   AWV, I got 1

   (b) Carry out 10 trials of the simulation.

   i. Sketch the results.



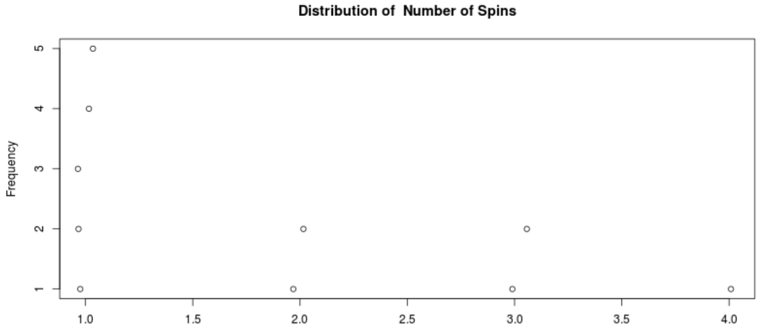
   ii. What is the average number of children per family? Show the results table and write down directions for computing the average.

```
           1  2  3  4
 Counts    5  2  2  1
```

   $(5 \times 1 + 2 \times 2 + 3 \times 3 + 1 \times 4)/10 = 19/10 = 1.9$

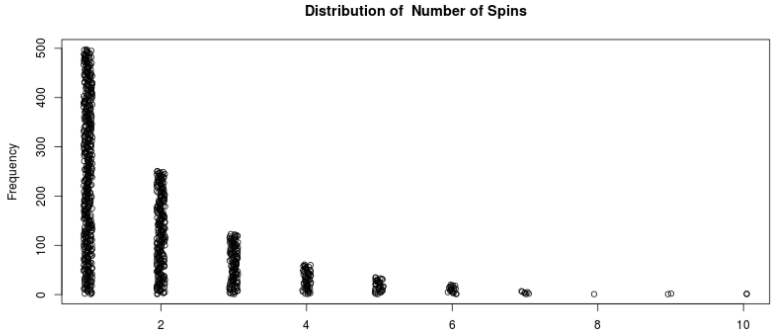
   iii. Where does that average number of children per family appear in the output?
   Under "Mean" in the summary table

(c) Carry out 100 trials of the simulation and sketch the results.



Average number of kids: *1.84*

(d) Carry out 1000 trials of the simulation and sketch the results.



Average number of kids: *2.03*

One of Each

What if a country adopted a "one of each" policy, allowing couples to keep having children until they have both a boy and a girl?

5. Discuss: What would happen to the US average number of kids per family (1.86 in 2000) under this policy? How would that number change? Each group member should answer. Include your reasoning.

   AWV, but something less than 2 makes no sense.

   Conduct a simulation study.

6. Model: Describe the model you will use to generate outcomes (what are the potential outcomes; sampling with or without replacement; probabilities of the potential outcomes; etc.);

   Assume that each pregnancy is equally likely to produce a boy or a girl, and that birth genders are independent of each other. We will sample with replacement so that the probability stays at 50/50%.

7. Based on your description of the model, set up the model in the web app using a spinner or mixer. Describe the inputs you gave for the web app.

   Use the spinner with outcomes B,G or M,F and percents set to 50,50. Or use the mixer with the same categories and equal counts, like 5,5 BUT then we must set "Replacement" to yes.

8. **Defining a Single Trial**

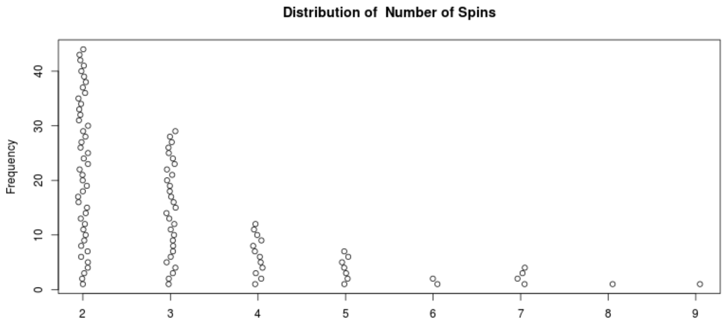
   Unlike previous simulations, a trial in this simulation is not defined by a set number of outcomes. Rather, a single trial will end when the family has at least one boy and at least one girl. We need to use a stopping rule rather than a fixed number of outcomes. The spinner and mixer apps both have a "Stop after" choice which should be set to One of Each.

   (a) Carry out a single trial of the simulation. How many kids did it take to get one of each?

   AWV, I got 9!

   (b) Carry out 100 trials of the simulation.

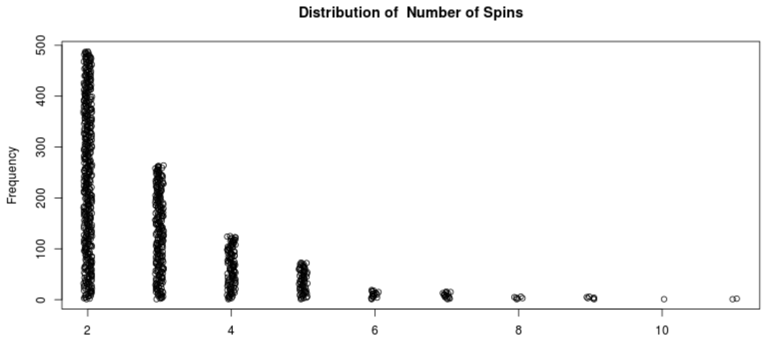
   i. Sketch the results.



ii. What is the average number of children per family? Show the results table the mean and standard deviation.

```
               2   3   4  5  6  7
 Counts  48  20  17  8  6  1
Mean:  3.07   Std.  Dev:  1.297
```

(c) Carry out 1000 trials of the simulation and sketch the results.



Average number of kids: Std Dev.: *Mean: 2.99, Std Dev: 1.40*

9. **Evaluate the Results and Answer the Research Question** Based on the results of your simulation, provide an estimate of the average number of children per family under the one-of-each rule. Explain where the answer came from.

   Hopefully based on 1000 trials because it contains more information, and is less easily thrown off by one odd trial.

10. What real life factors did you not consider when developing your model? How do you think these factors would affect your results if you included them in the simulation?

    Not everyone wants to have a boy, and not all families can keep having kids indefinitely til they get one.

11. In reality, slightly more boys (50.3%) are born than girls. However, females live longer and in the population of all US residents there are about 45% males. Run the "one son" model again with 45% boys. Find the numerical average number of children per family (which number of trials did you use?). How does this number for the new "one son" model compare with your answer to Question 9?

    We should do 1000 trials to get more accuracy. Then I get an average family size of 2.24. It should be bigger because its slightly harder to get a boy (which allows us to quit trying).

12. Compare the two model means (45% / 55% in gender versus 50% / 50%). Explain the relationship.

    We increased chances of girl by 5% and the family size increased by .24, which is 12% of the value 2 where we started.

13. Discuss the "other real world processes you consider random" which you thought of on the last question of Quizork 4. Could you simulate them with a mixer or spinner?

    AWV

Helper – Hinderer

We all recognize the difference between naughty and nice, right? What about children less than a year old? Do they recognize the difference and show a preference for nice over naughty? In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying for the foundation for social interaction.[3] In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were alternately shown two scenarios for the climbers third attempt: one where the climber was pushed to the top of the hill by another character ("helper") or one where the climber was pushed back down the hill by another character ("hinderer"). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The researchers found that the 14 of the 16 infants chose the helper over the hinderer.

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

Discuss:

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

   14/16

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

   14/16 is a high proportion. If there really is nothing going on, the infants should chose the toy randomly, making the proportion 8/16. So the vast difference between what actually happened and what would be expected if the infant did not notice the behavior suggests that infants do in fact notice and react to social behaviors of others.

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

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

   Yes because anything is possible in a random sequence, but it is not very likely.

4. Would the observed result (14 of 16 choosing the helper) be very surprising if infants had no real preference, or somewhat surprising, or not so surprising? How strong do you believe the evidence is against the chance model?

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

Pretty strong.

Modeling the Helper or Hinderer Problem

The key is to determine the likelihood of the observed result (14 of 16 choosing the helper) under the assumption that infants have no real preference, that is, the model based on random chance. To find out this likelihood, you will model the process of 16 hypothetical infants making their selections using random chance. Then, you can count how many of these "infants" choose the helper toy. After you have run many trials of the simulation, you can examine the variation in the results from this process and determine the probability of our result (or a more extreme result) under the chance model.

The model in this situation, and in many other situations, reflects the assumption of no preference, or nothing affecting their selection other than random chance. Because of this, statisticians refer to the model as the null model (nothing's going on except randomness).

5. Write a brief description of a simulation study that can be performed to model the assumption of no preference (i.e., infants select a scenario at random). Be sure that you explicitly identify each of the following in your description:

   (a) A **model** used to generate outcomes (what are the potential outcomes; sampling with or without replacement; probabilities of the potential outcomes; etc.)

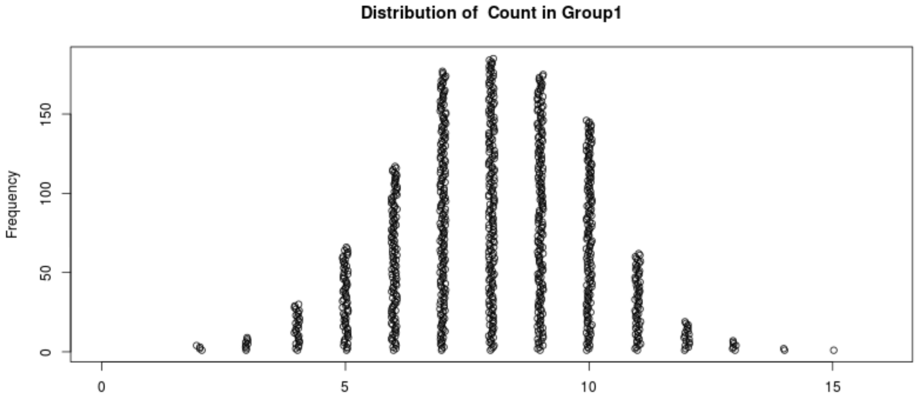
   (b) **Result** collected from each trial of the simulation
   Outcomes: Helper, Hinderer (equally likely, with replacement) and probability of helper or hinderer is 1/2
   Trial: 16 infants chose one outcome (16 spins or draws)
   Result: Number of helpers chosen.

6. **Carry Out the Simulation**
   Set it up in Spinner or Mixer.
   Run a single trial of the simulation. How many choose "Helper"?

   12 in my first, 8 in my second

Evaluate the Results and Answer the Research Question

7. Sketch a plot of 1000 trial results.



8. Locate where the observed 14 of 16 falls in your plot. Is the result from the original experiment likely or unlikely under the null model? Explain.

   Unlikely since one simulation in 1000 picked 14 or more helpers out of 16.

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

   No, because it is unlikely that the 1 sample in the experiment is the 1 in 1000 samples from the null model. The observed result provides evidence that the children do seem to favor the toy that helps.

10. What do these results suggest about the null model? Explain.

    This suggests that the model is not valid. In other words, the observed result suggests infants do favor one toy over another.

Strength of Evidence

The observed result gets compared to the distribution from the simulation to gauge the evidence against the null model. 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 help others understand how likely or unlikely the observed result is under the null model, we want to report the "strength of evidence".

The strength of evidence is quantified by answering the question: "What proportion of the simulated results indicate at least as much evidence as the observed result against the null model?" For example, consider the results from the "Detecting Fraud" data on day one if we had 12 pairs to pick the true sequence from. Instead of using the class choices as a model for "just guessing" we could use spinner or mixer to set up a model for the number right out of 12 in the "just guessing" case. Suppose the instructor got 10 of 12 right. The numbered outcomes on the plot are those as extreme or more extreme as the 10 of 12. The chance is only 16/1000 of getting a result this extreme when the null model is true. We can think of 0.016 as the strength of evidence against the null model for a result of 10 correct picks. It is the probability of obtaining results as extreme or more extreme when the null model is true.

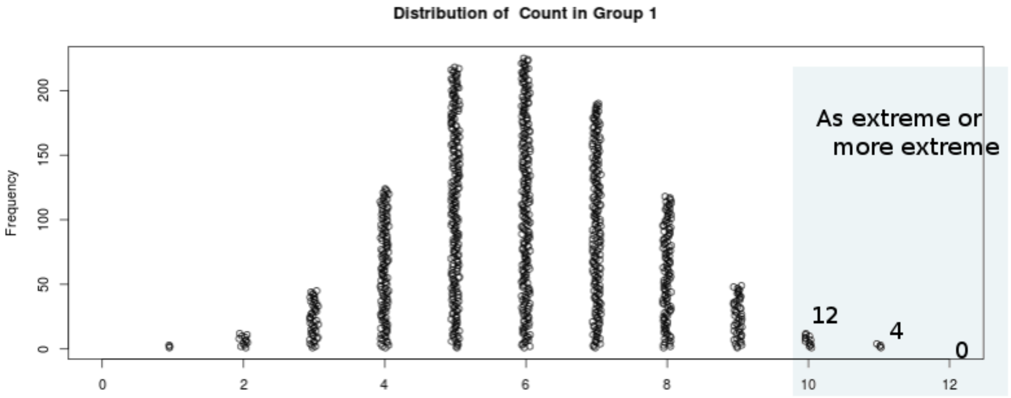


Figure 1: Simulation results obtained from the null model. The numbered outcomes (out of 1000 trials) were as or more extreme as the instructor who got 10 right of 12 and indicate the strength of evidence of 16/1000.

The strength of evidence against the null model based on the observed result of 10 correct is 0.016. This suggests that ten correct is quite unlikely under the null model – it is in the most extreme two percent of the simulated results.

11. Quantify the strength of evidence for the observed result of 14 out of 16 infants choosing the helper toy.

    1/1000 on my simulation

12. Based on your analysis, how strong is the evidence against the null model?

    Very strong

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

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

14. Put the following steps into their proper order:

    (a) gather data *2*

    (b) formulate a hypothesis *1*

    (c) report strength of evidence *5*

    (d) simulate a distribution *3*

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

Unit 1 Wrapup

Vocabulary Define each term:

- Model *What we simulate. Identifying the model means noting what the outcomes are, the probability associated with each outcome, whether you are sampling with or without replacement, what constitutes a trial and when a trial ends, as well as the result or statistic you are collecting from each trial.*
- Trial *An event of interest. Know when a trial ends and what you will be collecting from each trial.*
- Observed result *The result from an experiment or sample. We will use this to provide evidence (strong or weak) against the null model. We compare it to the null model to see if the data is consistent with (weak evidence against) or contradicts (strong evidence against) the null model.*
- "Blind Guessing" Model *Null model in which outcomes are equally likely.*
- Null model *The model we simulate. Null model typically will be random or no effect. We simulate what would happen if the null model were correct (true) in order to get the null distribution. We compare the observed result to the null distribution to see if the observed result is a likely value. If it is unlikely (typically defined as less than a 5of happening under the null model), we say we have strong evidence against the null, reject the null hypothesis in favor of the alternative hypothesis, and call the observed result statistically significant.*
- Strength of evidence *The probability (proportion of simulations) of results as or more extreme as the observed result.*
- Patterns in random behavior *Galton's board showed that random left–right deviations could lead to a bell–shaped curve. Our simulations have also given patterns, but they could be skewed. Not all are bell-shaped.*

Simulation

1. Suppose that you receive 10 text messages per day and that half of your messages come from family members. Describe how you would set up a simulation to model the number of texts from family each day using

   (a) 10 coins *Heads = family, Tails = other. Flip a coin 10 times to simulate 1 trial (1 day) and collect the number of heads. Repeat that many times (collect the result from many trials) and plot the number of heads for each trial.*

   (b) Spinner

   Label two categories "family", "other" and give probabilities 50,50. Set 10 spins and 1000 repeats. Store count in category 1

(c) Mixer: *label two categories "family", "other" and give counts 5,5 with replacement. Set 10 draws and 1000 repeats. Store count in category 1*

2. Simulate the "Detecting fraud" activity with 15 pairs of sequences to choose the "true" one from. Use the web app to simulate 1000 trials under the "blind guessing" model. How many correct choices (of 15) would someone have to get to provide convincing evidence that they are able to detect fraud better than the model suggests? Explain your reasoning – not with a mathematical formula, but just by eye–balling the distribution. Certainly picking all 15 pairs correctly would be strong evidence, but that's so rare, you probably didn't see it happen in 1000 trials. What other numbers of matches would you find to be strong evidence? Why?

   I would say getting 11 or more out of 15 would provide convincing evidence against the blind guessing model. This is because the probability of getting 11 or more out of 15 is 59/1000 = .059 = 5.9% which is close to the traditionally used 5% cut off value (significance level).

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

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

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

   (b) What is the strength of evidence against the null model provided by this new data?

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

   (c) What conclusion would you draw about the null model?

   There is little to no evidence against the null model in this case so I would not reject it and can conclude the null model could be correct. Infants may chose a toy at random.

   (d) What conclusion would you draw about the null model if 13 of the 16 infants chose helper?

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

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

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

   Change the Stop After? Fixed number of spins or draws to 8 instead of 16. The only thing that has changed in the model is when a trial ends (here after 8 children pick instead of after 16).

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

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

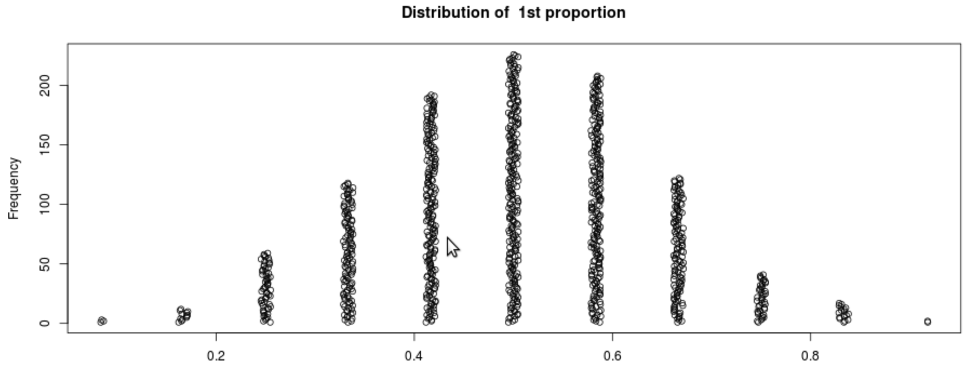
4. Researchers asked the question, "Do more than half of kissing couples lean their heads to the right?", and they collected data on 12 couples, of whom 8 leaned to the right.

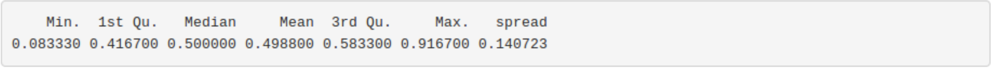
   (a) How would you setup a simulation to obtain strength of evidence?

   Spinner: Number of Categories = 2
   Category names = Left,Right
   Percentages = 50, 50
   Stop After = Fixed Number of Spins
   Stop after how many spins? = 12,
   Number of Trials = 1000.
   Store what result? = Count in group 2

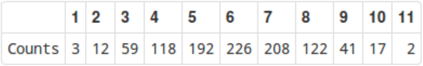
   Mixer: Number of Categories = 2
   Category names = Left, Right
   Number of Each = 1, 1
   Replace each draw? = Yes
   Stop After = Fixed Number of Spins
   Stop after how many spins? = 12
   Number of Trials = 1000
   Store what result? = Count in group 2

   (b) Use the figure below (based on 1000 trials) to answer the question. Be sure to give the strength of evidence and your conclusion.





```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.   spread
0.083330 0.416700 0.500000 0.498800 0.583300 0.916700 0.140723
```



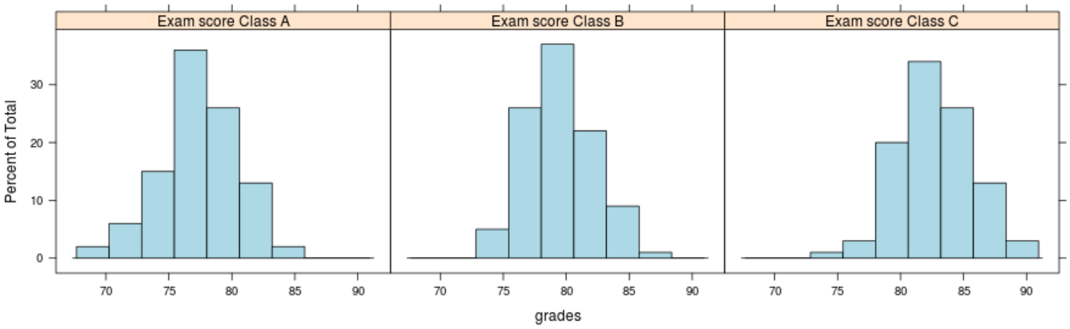
	1	2	3	4	5	6	7	8	9	10	11
Counts	3	12	59	118	192	226	208	122	41	17	2

With a strength of evidence of 185/1000 = .185 = 18.5% (because 185 of the 1000 trials gave 8 or more couples leaning to the right), there is little to no evidence against the null hypothesis which states that couples lean to the right or left randomly (they *are equally likely to lean to the right or left). We do not have enough evidence to conclude couples do tend to lean to the right when they kiss. The null model might be true, as the data do not argue against it.*

Unit 2

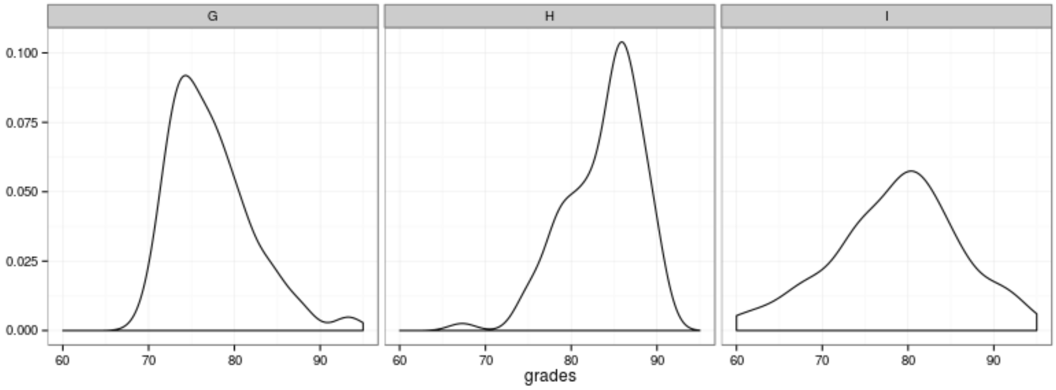
Comparing Distributions

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



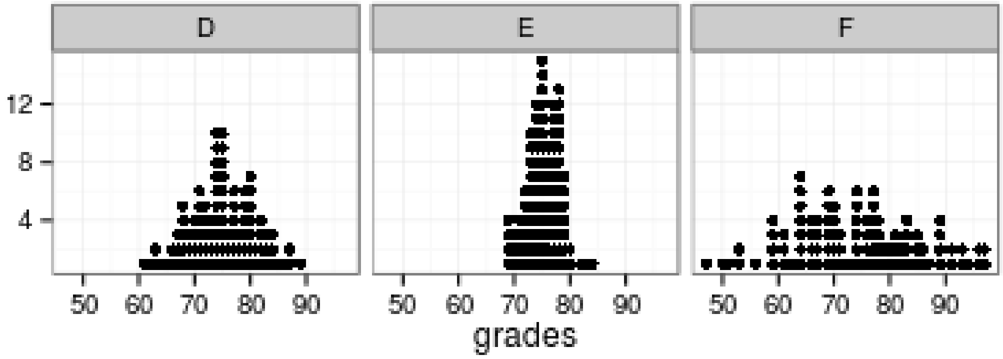
Class C has the highest center, so most will vote for it.

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



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

3. And here's a third set. Which do you prefer? Explain the differences.



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

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

   (a) Comparing location or center (measured by mean or median) tells us which class did best "on average".

   (b) Comparing spread (interquartile range or standard deviation) tells us which class is generally closest to its mean.

   (c) Comparing skew (could be left or right) to symmetric. 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.
   1 _____ *center* _____ 2 _____ *skewness* _____ 3 _____ *spread* _____

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

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

6. The cost of a college education is frequently discussed as a national problem. One factor we might consider is the amount of money students can earn over summer to help pay for college expenses. In particular, ask yourselves: do students who work summer jobs in Montana (let's also include northern Wyoming and eastern Idaho) make more or less per hour than students who work further away? Discuss with your group and write down your own opinion and whether or not the group all agrees.

   AWV

7. We are going to poll this class to see how much you all made (hourly) this summer. Discuss with your group and write down two or three types of people we do not want to collect data from.

   AWV. Older people with established trade or occupation, those who just vacationed, ...

8. Go to the web site given on D2L and answer the questions there.

Comparing Ebay prices for a new versus a used video game

9. While we're waiting for everyone to enter their data, we'll look at a tool which makes it easy to compare groups.

   Go to the site:
   http://www.rossmanchance.com/applets/Dotplot.html. Click Stacked (Group Value) and, under the data box, Clear.

We'll look at data from Ebay auctions of the video game MarioKart for Wii in October 2009. The variables of insterest are condition (new or used) and total price. The data are available from D2L or at `http://www.math.montana.edu/~jimrc/classes/stat216/data/marioKart.csv`

Copy the data (cntrl-A, cntrl-C or use right mouse options) and paste it (cntrl-V) into the data box. Click Use Data and see what you get for a plot. Sketch the vague shape of it here.

(a) You will see some unusual points. Discuss: "Are these outliers?" and write the group conclusion here.

points from lines 20 and 65 are quite high, 326 and 118 versus tyical price in the 80 dollar range.

(b) Looking at the descriptions of items sold, we see that the game in line 20 came with a game console, and the game in line 65 was part of a package of 10 video games. Discuss: should we remove these from the data? (Other descriptions were similar to "Wii MarioKart and Steering Wheel"). Explain your conclusion here:

Omit because these are not comparable items.

(c) Regardless of your conclusion, we'd like you to look at the data without lines 20 and 65 (look for prices 326.51 and 181.5), so delete them from the data box and reclick Use Data. Check that the two large prices are now gone. View dotplot, boxplot, and histogram.

i. With a lot of points, dotplots are not very useful. Write down two reasons the dotplots don't work well with these data. Be specific about some way in which they are hard to interpret or compare.
Points run together and create a blur across the height of 1 dot, so we can't see how many there are.

ii. Click histogram and move the slider for bin width. Find a number of bins which gives a "not too bumpy" view of the data without lumping it into to few bins to be useful. (There is not one right answer. You could give a range of values.)
I like 10 to 15 bins.

iii. Click boxplot and compare the medians of the two groups. How much more valuable is new than used? Are the spreads about equal, or is one group more spread out than the other? How are you judging spread?
Median for used is about $45, and for new about $55. Widths of the boxes is similar, about 9 to 10 $, and standard deviations are both near 7.3

iv. Are the distributions symmetric or skewed?
I'd say slightly right skewed.

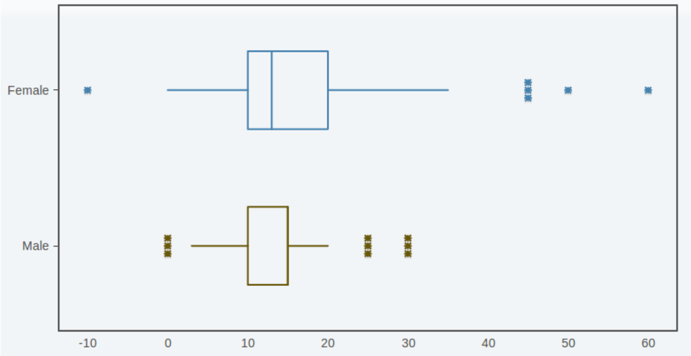
Comparing Wages

10. Click Clear and delete all the video data.
Go to the Google Doc spreadsheet (see link on D2L) to see class responses from the web data form.

Shift-drag over columns A and B and push CONTROL-C for copy.

Paste the data into the data box. Look at the dotplot, histogram, and boxplot.

(a) Sketch the boxplots here.



(b) Write down the means and medians for near Bozeman and far away. Is this an important difference to a college student considering where to work next summer?

Medians:
Means:

(c) Write down the standard deviations and IQR = Q3 − Q1 for both groups. Is it obvious that spread differs?

IQR:

(d) Describe the shapes of the distributions. Are they similar?

(e) Are there outliers in either group? If so, can we find out why some are extreme? Write down an example of a high wage you think does not belong with these data, and another of a high wage which could be part of the wages of interest.

11. Column C of the spreadsheet contains gender. Copy columns B and C into the data box and compare the two distributions. Is there much difference?

12. Comment on what you've found. Do women get paid less per hour than men?

Take Home Message:

We make comparisons every day. When comparing distributions of data, we often care about mean or median, but we may also care about the spread of the distribution. When data are highly skewed, a few outliers make a big difference. In the rest of this course we will be comparing centers of distributions. This is not appropriate unless the distributions are of similar shape and spread. However, it is the most commonly used comparison. If you understand the principles used to make statistical comparisons, you'll be able to easily modify them for other types of comparisons such as comparing slopes when fitting regression lines (take Stat 217 to learn more).

Energy Drinks

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

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

1. Why are energy drinks popular?

   AWV

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

   AWV

3. How do energy drinks interact with alcohol?

   AWV

4. An experiment tried to compare the effects of energy drinks with and without alcohol on human subjects:
   Pharmacology is the study of how drugs affect the body. An article in *Human Psycopharmacology* in 2009 reported on an experiment intended to tease out some of the effects and to compare an energy drink without alcohol to one with alcohol and to a non-energy drink. The research question is:

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

   Higher RBANs scores indicate better memory skills. Refer back to today's reading.

   Go to the site:
   http://www.rossmanchance.com/applets/Dotplot.html. Click Stacked (Group Value) and, under the data box, Clear. Paste in the energy drink data from D2L.

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

Examine the data using the three plotting options: dotplot, boxplot, and histogram. Describe any differences you see between Red+A and Control groups.

Center
Mean of RED+A is clearly the lowest (-8.3), Control the highest (1.2)
Spread
Control has the smallest spread (3.9 sd), followed by RED (6.4) and largest spread is RED+A (10.0)
Shape
RED+A seems most symmetric (boxplot), the other two both have outliers. From the histograms, the Control group may be left skewed and RED seems symmetric also.

GROUP TASK

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

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

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

Consider this important question:

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

6. Discuss this within your group and write down your answer. Don't say that it has anything to do with the drink they were given because we are assuming the drinks are all having the same effect.

   Random variation!

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

   Do these new means agree well with the original ones? If not, which ones changed by approximately how much? Are they similar?

   They should be similar but not exactly the same.

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

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

9. Go to the applet: http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2 Again, click Clear, and paste in data for RED+A and control.

   (a) Do the means in the summary table match what we had earlier?

   They should!

   (b) Click Show Shuffle Options, keep Number of Shuffles set to 1 and change Data to Plot . Watch what happens when you click Shuffle Responses. What are the means for control and RED+A in this reshuffled version? The difference?

   AWV

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

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

   (d) Set Number of Shuffles to 1000 and click Shuffle Responses three times. Where is the plot centered? Why is it centered there?

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

10. Below the plot change $\boxed{\text{Greater Than} \geq}$ to $\boxed{\text{Less Than} \leq}$ and enter the observed difference in means from the original data (it appears to the left of the window center). What proportion of the samples are this extreme?

    0.008 in my sample.

11. There are other reasons that one person might show more change in RBANS than another person. Think of three or more and write them down.

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

12. Variables like those you just listed are called "lurking variables" when we don't measure them and adjust for their effects. When we randomly assign treatments, how should the groups compare on any lurking variable?

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

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

    Most will say probably no because of the lurking variables or sample size (?) but you can since we have random assignment. This is a good place to start discussing scope of inference.

Important Ideas:

1. If there is no treatment effect, then differences in distribution are just due to the random assignment of treatments. This corresponds to a "null hypothesis" of no difference between treatment groups.

2. By randomly applying treatments, we are creating groups that should be very similar because differences between groups (age, reaction to alcohol, memory) are evened out by the random group allocation. If we see a difference between groups, then we doubt the null hypothesis that treatments don't matter. Any difference between groups is caused by the treatment applied. Random assignment is a very powerful tool. When reading a study, it's one of the key points to look for.

Reference

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

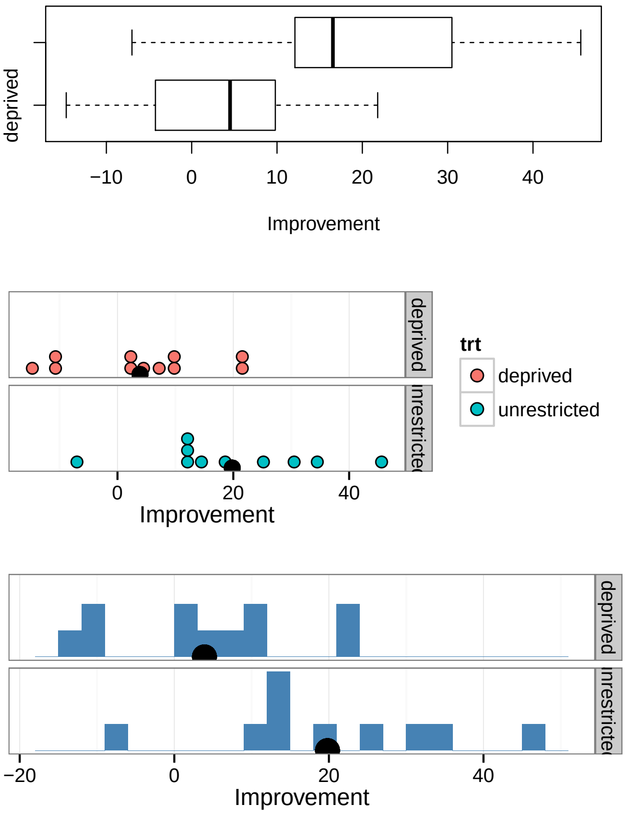
Sleep Deprivation Study

Researchers have established that sleep deprivation has a harmful effect on visual learning. But do these effects linger for several days, or can a person "make up" for sleep deprivation by getting a full night's sleep in subsequent nights?

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

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

```
trt, improvmnt
unrestricted, -7
unrestricted, 11.6
unrestricted, 12.1
unrestricted, 12.6
unrestricted, 14.5
unrestricted, 18.6
unrestricted, 25.2
unrestricted, 30.5
unrestricted, 34.5
unrestricted, 45.6
deprived, -14.7
deprived, -10.7
deprived, -10.7
deprived, 2.2
deprived, 2.4
deprived, 4.5
deprived, 7.2
deprived, 9.6
deprived, 10
deprived, 21.3
deprived, 21.8
```



Subjects' performance on the test was recorded as the minimum time (in milliseconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen. The sorted data and plots presented here are the improvements in those reporting times between the pre–test and post–test (a negative value indicates a decrease in performance). Black dots in the bottom two plots show position of the group means (3.90, 19.82).

Discuss and record your groups opinions:

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

   Yes because the mean is higher and only 1 data value from the unrestricted sleep did not improve (improvement score was negative) and only 2 people in the sleep deprived group were above the average of the unrestricted sleep group.

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

   Yes, $19.8 - 3.9 = 15.9$. *This difference does appear to be large.*

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

   Of course, as every outcome is possible (though not equally likely) under random chance.

In this study, the random chance is introduced not through the sampling process (like in Unit 1), but rather in the random assignment to groups. It is possible that sleep deprivation has a harmful effect, but we want to consider the other possibility: that the treatment has no effect, but researchers happened to assign the subjects with high improvement in their test scores into the unrestricted sleep group.

As in the simulation study in the Energy Drinks activity, consider what you would likely see if there really is no difference in test score improvement between the two conditions. (This is the null hypothesis!) In that case, the assigned treatment is just a meaningless label. The subjects would improve the same amount regardless of which group they had been assigned to because the effect on test scores would be identical for both groups.

Since there are many possible ways to randomly assign 20 subjects into 2 groups, it is possible that the random assignment that came up was just unlucky and happened to assign the subjects who were going to have more improvement in their test scores into the unrestricted sleep group. A different random assignment might have spread them out so there were no harmful effects, or it could have reversed the groupings and made the effect of sleep deprivation positive. The good news is that we have tools to evaluate how unusual this group assignment is **under the "no effect" model**.

The key statistical question is:

When treatments have no effect on test improvement, how unusual is it to see a result this extreme or more so due simply to the process of random assignment?

MODELING THE SLEEP DEPRIVATION STUDY

We will now conduct a randomization test to find out how likely or unlikely the observed result (change in mean of 15.9) is under the null model that sleep deprivation has no effect. If it has no effect, then the difference in means happened just because people with larger scores got randomly assigned to one group and low scores to the other.

Because (under H_0) treatments are having no effect other than sticking meaningless labels on our units, we can re–randomize to look at different random assignments that could have occurred. The difference in the mean improvement scores is then computed under each re–randomization. Repeat this process of re–randomizing the data and computing the differences many times. The **distribution** of these differences shows what we can expect to see under the **null hypothesis**. We use the **null** model of no treatment effects to evaluate how unusual our observed data are. It is important to realize we can get lots of trials from the null model, but we only have one observed data set.

Go to the applet web page: `http://www.rossmanchance.com/applets/index.html` and select $\boxed{\text{two means}}$ in the second column under Statistical Inference. Click $\boxed{\text{Clear}}$, and insert the data provided.
Click $\boxed{\text{Show Shuffle Options}}$ with Number of shuffles set to $\boxed{1}$ and get a new trial. Write the means and their difference.

Modeling a Set of Fixed Responses Under the Null Model

Under the null hypothesis of no difference between the two conditions, these response values are fixed - they will always be the same for the subjects. The labels will change.

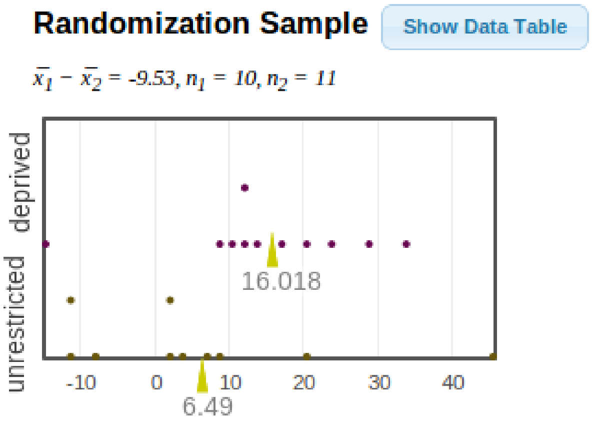
4. Can you predict the means generated in each trial?

   No, because the responses are randomly assigned to treatments.

5. Is the webapp generating outcomes at random? Explain.

   It is generating group labels at random, which makes the means change for each group, but it is not generating scores at random.

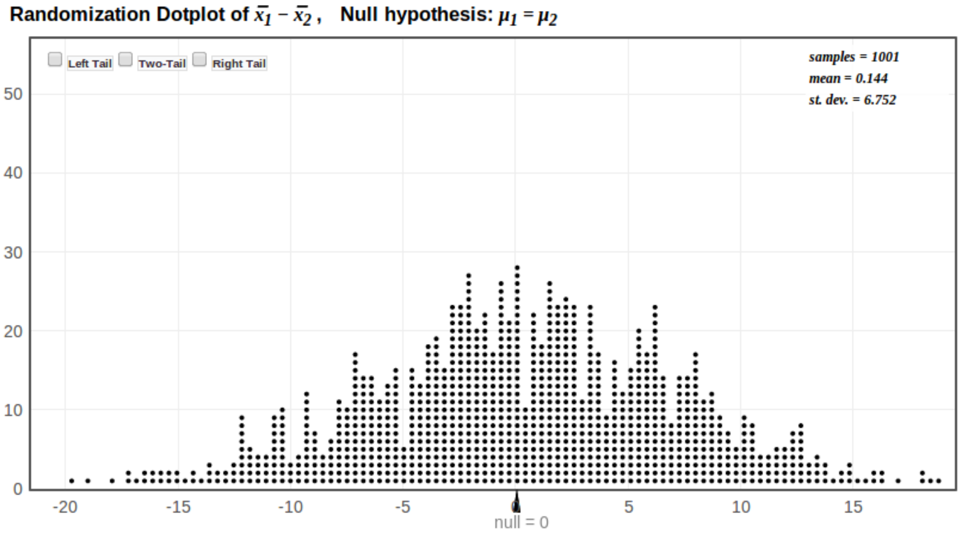
6. In this simulation, the trial represents what might have occurred for another random assignment of subjects to conditions. The relabeled data are plotted. Sketch the dot plot below.



For 1 sample: my means were 16.08 for deprived and 6.49 for unresricted.

Evaluate the Results

7. Run 3000 shuffles and sketch the plot of the the differences in means here.



8. What are the cases in the plot? (Hint: Ask yourself what each individual dot represents. Note what happens in the shuffle plot when you click a different part of the histogram.)

   Each dot represents the difference in mean improvement scores between the unrestricted and sleep deprived groups (unrestricted sleep deprived) for one re–randomized trial.

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

   Plot is centered at 0 because if sleep deprivation has no effect on improvement scores, then the difference in scores between unrestricted and sleep deprived should be 0.

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

    Unlikely as it is in the very tail of the distribution.

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

    One–tail (can a person make up for sleep so our alternative is that the improvement scores are higher (right–tail) for the unrestricted group). P–value of 0.0080 in my simulation.

Strength of evidence is referred to as a **p–value.**

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

    From this study, it appears that no, people cannot make up for sleep. The mean improvement score for those in the unrestricted group was significantly higher than the mean improvement score for those in the sleep deprived group (p–value of 0.0080). This is strong evidence against the null model and we have evidence to conclude that sleep deprivation is harmful.

    The purpose of the simulation was to see what differences in mean improvement scores would be likely if sleep deprivation really had no effect on improvement scores. Because our observed difference in mean improvement scores of 15.9 (unrestricted deprived) is unlikely to occur if sleep deprivation really has no effect, we have evidence to support that sleep deprivation is harmful.

13. The plots on the webapp site resemble the plots we got from the mixer and spinner web apps. For example, with the helper–hinderer we made a "strength of evidence" estimate from the tail of the plot in the web app. Explain how both the webapp and the spinner app use random outcomes, but in a different way.

    Spinner and mixer are randomly generating outcomes (essentially it is randomly generating responses) whereas the StatKey app is keeping the responses fixed and randomly generating what treatment group the responses came from.

Running Shoes

Manufacturers of running shoes recommend that people with low arches be fitted with a motion control shoe, that people with medium arch height use a stability shoe, and that those with high arches wear a cushioned shoe. They claim that taking arch size into account when choosing shoes will decrease foot and leg pain and reduce injuries.

How can we determine whether the manufacturers claim about shoe fitting is correct?

Discuss:

1. Suppose a friend who runs a lot and has a low arch tells you she has switched to a motion control shoe and now has much less pain in her feet when running. Is this strong evidence that these shoes really do decrease pain? Explain.

   No, it is only 1 person. This is not enough data.

2. Now suppose that you take a random sample of runners by randomly selecting them from the population. You identify whether or not they wear a shoe fit as recommended to their arch, and then compare the two groups levels of pain while running. If, on average, the group which wears shoes fitted to their arch height has a much lower level of pain, is this compelling evidence that the manufacturers' claim is correct? Explain.

   No because the people who chose to use recommended fits could be different from people who buy any shoe (better athletes or more accustomed to running so they experience less pain).

3. What factors influence the amount of foot and leg pain people experience when running? List at least three.

   How often they run, age, previous injury, etc

Both descriptions in questions 1 and 2 lack compelling evidence to support the claim. The results from question 1 are based on a person's story and are called **anecdotal evidence**. Anecdotal evidence can sound compelling, but it is just an unsubstantiated story and is of little value in scientific research. The practice of statistics involves designing studies and collecting data so people do not have to rely on anecdotal evidence.
The problem with the question 2 is that you do not know whether or not the two groups might differ in more ways than simply choice of shoes. For example, subjects who choose to wear the recommended shoes could be more athletic to begin with than those who opt to wear the ordinary running shoes.

When investigating whether or not one variable causes an effect on another, researchers seek to exert control by creating a comparison group and then assigning subjects to the explanatory variable groups.

An **experiment** is a study in which the experimenter actively imposes the **treatment** condition on the subjects. Ideally, the groups of subjects are identical in all respects other than the treatment, so the researcher can then see the variables direct effects on the response variable.

A 2010 study published in the *American Journal of Sports Medicine* investigated the claim with groups of 840 men and 571 women who were undergoing training in the Marine Corps. Suppose we are designing this experiment.

The control group will get a standard stability shoe. The treatment group will get shoes recommended for their arch size.

4. How would you handle the differences between men and women?

   Make sure the women and men are split evenly between the groups since one group might have a higher threshold for pain (block on gender). The researchers analyzed men and women separately.

5. Consider just the men. How might you assign subjects to control and treatment groups in an effort to balance out potentially confounding variables?

   Assign them randomly!

People have tried various schemes to make sure that two groups are as alike as possible. **Random assignment** is the "gold standard" method of assigning subjects to treatment conditions in an experiment. It means that each subject has the same chance of being assigned to the treatment group. You will explore the properties and benefits of random assignment in this activity using just 12 people. Another word used for random assignment is **randomization**.

6. Describe in detail (so another student could replicate the process) how you might implement the process of randomly assigning 12 subjects to 2 treatments.

   Flip a coin for each person: H goes into standard, T into recommended (as 1 example).

Suppose that your 12 subjects are listed in the following table. We know their weight in pounds and whether or not they exceed the Marine Corps' weight–to–height standard.

Name	Weight	OverWt	Name	Weight	OverWt	Name	Weight	OverWt
Andy	186	no	Kyle	182	no	Patrick	226	yes
Ben	214	yes	Mark	196	yes	Peter	158	no
Brad	195	yes	Matt	204	no	Russ	161	no
Jorge	201	no	Michael	199	no	Shawn	187	no

7. Write each subjects name on an index card. Shuffle the cards and randomly deal out 6 for each group. Record the names weights and if they were over weight for their height here:

Treatment Group		
Name	Weight	OverWt

Control Group		
Name	Weight	OverWt

Answers will vary.

8. Calculate the proportion over "weight guidelines" in each group. Also subtract these two proportions (treatment proportion minus the control proportion).
Treatment group proportion overweight:

Control group proportion overweight:

Difference in proportions (treatment – control):

9. Calculate and report the average weight in each group. Also subtract these two averages (taking the treatment groups average minus the control groups average).
Treatment group average weight:

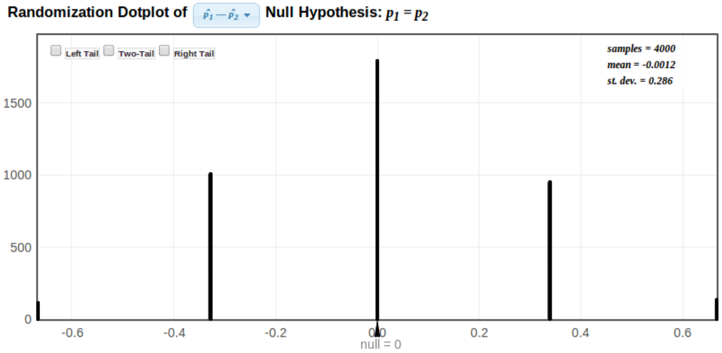
Control group average weight:

Difference in averages (treatment – control):

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

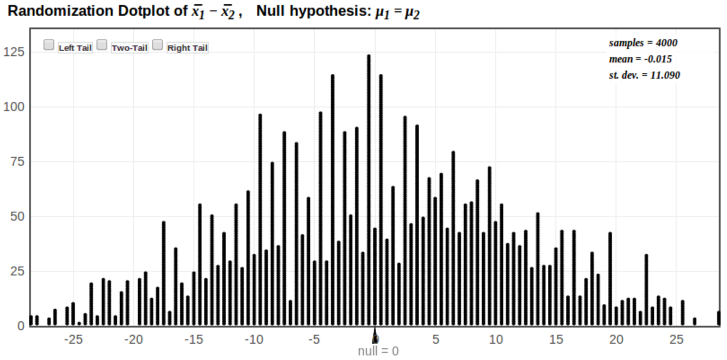
    Usually we get similar group, but not always when sample sizes are small.

11. Go to the Rossman-Chance site http://www.rossmanchance.com/applets and click Dolphin Study Ap to see what happens under repeated re–randomizations. Change the numbers of successes (over weight guideline) to 2 for both groups and the numbers of failures to 4. Click Show Shuffle Options and watch what the cards do as you click Shuffle several times. When you've got it figured out, select 1000 shuffles and sketch the plot of the differences in proportions of overweight soldiers.



12. Next use the Rossman-Chance site Randomization test for quantitative response to look at the weights. Click Clear to remove their data and insert ours. Click Generate 1000 samples four times and sketch the plot of the differences in average weights.

```
group, weight
trt,  186
cntrl, 214
cntrl, 195
trt,  201
trt, 182
cntrl, 196
trt, 204
trt, 199
cntrl, 226
trt, 158
cntrl, 161
cntrl, 187
```



13. Where are the plots in #11 and #12 centered? Explain why this indicates that randomization is effective.

    Centered at 0. When we randomly assign people to treatment, in the long run half the time each person goes to treatment and half the time to control. Their weight (or overweight status) then is subtracted half the time and added half the time, so the plots are symmetric about 0. In the long run, it all evens out.

14. Does randomization always balance out the weights exactly between the two treatment groups in each randomization? Explain.

    Not exactly in every single trial but on average the groups are equivalent.

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

    Yes because the distribution of the difference between the two groups is centered on 0.

16. The abstract mentions several other variables which were accounted for in the analysis. What are they?

    fitness, smoking, prior physical activity (link in the end of this activity)

17. Adjusting for variables which might affect our outcome is important, but we always fear that some important variable was not recorded. Other variables that were not measured are called **confounding variables**.

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

    Yes because the other two variables evened out in the long run so that would likely happen with any variable we measured on the participants.

18. What do the plots suggest about the effect of randomization on unrecorded or unseen variables?

    That it will even them all out in the long run.

19. How unusual it is to obtain a difference in mean weights that's more than 20 pounds? Fill in 20 next to the Greater than box and record the proportion above 20 in the upper tail. Also select Less Than and change the number to –20. What proportion of shuffles have differences in means more than 20 (in both directions)? What proportion are inside those bounds?

    0.076 outside, 0.924 inside

20. What values for difference in mean weights do you feel would be highly unusual? Explain your choice.

    AWV. Perhaps 22 or 23 since the probability of more extreme is less than 0.05.

21. Now suppose you conduct this random assignment and find that the treatment group has substantially less pain than the control group, on average. Would you be comfortable concluding that the treatment caused the decrease? Give an argument that no confounding variable was responsible.

Since randomization evens out the effect of all variables other than the treatment applies, any difference in the response must have been caused by the treatment.

22. Results from the study showed that the strength of evidence was .15 for men and .083 for women. If we use a cutoff value of 0.10, is this significant evidence against the null hypothesis?

    It is for women ($.083 < .10$*), but not for men (*$.15 > .10$*).*

Experimenters try to assign subjects to groups so that lurking and potentially confounding variables tend to balance out between the two groups. This activity demonstrates that randomization generally achieves its goal of creating treatment groups that are similar in all respects except for the treatment imposed. Furthermore, we can see from the plots how unusual it is to get a large difference. If the randomly assigned groups turn out to differ substantially on the response variable, you can draw a cause–and–effect conclusion between the explanatory and response variables.
As you have just seen, random assignment does not always create completely identical groups. If the difference in the values of the response variable between two groups is so large that such an extreme difference would rarely occur by random assignment alone, then the difference between the groups is said to be statistically significant. When a randomized experiment produces a statistically significant difference between the groups, then it's reasonable to conclude that the explanatory variable caused the difference in the response.

Dolphin Therapy

People find it exhiliarating to swim with dolphins, but can this adrenaline rush be therapeutic? To find out, researchers recruited 30 subjects with mild to moderate depression between ages 18 to 65. The subjects were taken off drugs and psychotherapy for four weeks, then flown to Honduras for a swimming and snorkeling vacation. Half (randomly assigned) swam in the vicinity of bottlenosed dolphins, the other half were not near the dolphins. After two weeks, all were tested and assigned a "depression score". The same test had been given at the start of the vacation, and the researchers are interested in whether depression score was lower (some improvement) or not (no improvement).

Does swimming with dolphins reduce depression?

Discuss the Following Questions

1. What is the treatment condition in this study?

   Swimming with dolphins or not

2. What is the response variable in this study?

   Improved or not

   Results: of the 15 subjects getting "dolphin therapy", 10 showed improvement, whereas in the control group, only 3 of 15 improved.

3. Organize the results into a 2 by 2 table

	dolphin swimmers (A)	control group (B)	total
Improved	10	3	13
Not improved	5	12	17
Total	15	15	30

4. Of the 15 subjects assigned to the dolphin therapy condition, what proportion improved? We will label this $\widehat{p}_A$ because we think of p as the true proportion who would improve if all depressed peole swam with dolphins, and we stick the "hat" on top to show that this is an estimate (or a statistic) computed from the observed sample. Finally, the "A" subscript is just to distinguish the groups.

   10/15 = 0.667

5. Of the 15 subjects assigned to the control condition, what proportion improved? We'll call this $\widehat{p}_B$, using a "B" for this second group.

   3/15 = 0.20

6. Find the difference between the proportion of subjects assigned to dolphin therapy condition that improved and the proportion of subjects assigned to the control condition that improved. $\widehat{p}_A - \widehat{p}_B =$

   $0.667 - 0.20 = 0.467$

7. What proportion of all 30 subjects improved? This is called a marginal distribution because it just uses totals. If the treatment has no effect, then this will be a good estimate of the true marginal probability that any depressed person who goes on a swimming vacation to Honduras will improve, so label it $\widehat{p}_m$ where m means marginal.

   13/30 = 0.433

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

   The data suggests that overall approximately 43.3% of participants improved, but the difference between the proportion who improved in the two treatment groups (swam with dolphins control) was 46.7%. This appears to be a large difference which supports the idea that swimming with dolphins may increase improvement.

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

- Does the higher improvement rate in the dolphin group provide convincing evidence that the dolphin therapy is effective?
- Is it possible that there is no difference between the two treatments and that the difference observed could have arisen just from the random nature of putting the 30 subjects into groups (i.e., the luck of the draw)?
- Is it reasonable to believe the random assignment alone could have led to this large of a difference?
- Just by chance did they happen to “assign” more of the subjects who were going to improve into the dolphin therapy group than the control group?

One way to examine these qustions is to consider what you would likely see if 13 of the 30 people were going to improve (the number of subjects who improved in our sample) regardless of whether they swam with dolphins or not. If that is the case, you would have expected, on average, about 6 or 7 of those subjects to end up in each group (the null model suggests this).

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

MODELING THE DOLPHIN THERAPY STUDY

You will answer this question by using the Rossman–Chance web applet to conduct a randomization test which lets us see the results one can get just dueto variation in random assignment. We'll operate under the null model assumption that the control and dolphin therapy conditions are equally effective (or ineffective) at improving depression.

Go to the Rossman–Chance web page: http://www.rossmanchance.com/applets and select Dolphin Study applet in the second column. This time we don't have to clear the data, because they are using the same data we are studying.

10. The web page shows "Success" and "Failure". Relate those terms to the labels in our table above. *Success means improved, failure means "did not improve".*

11. What proportion of subjects assigned to the dolphin therapy condition are successes? 0.667 What proportion of subjects assigned to the control therapy condition are successes? 0.20 What is the difference in proportions between these two groups?

    Should see the same answers in the webapp as you got for 5, 6, and 7.

12. Click Show Shuffle Options and leave the display choice set to Cards. Explain why there are 10 blue and 5 green cards in Group A and 3 Blue, 12 Green in Group B.

    Blue is representing Success, green Failure.

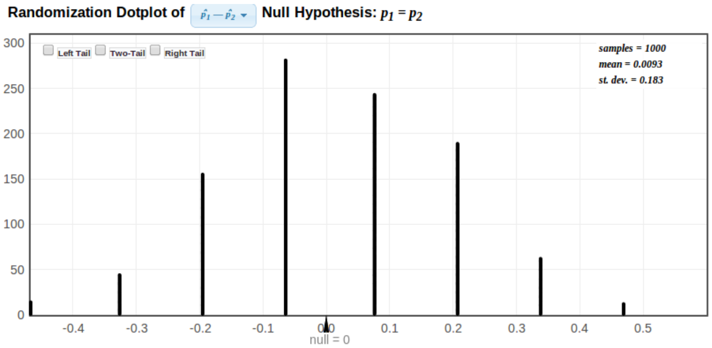
13. Predict what will happen when you click Shuffle. How many groups? How many of each color of card? How will it be different from this view?

    We'll again get 2 groups of 15, but with some new blue/green split in each (still 13 blues and 17 greens)

14. Do one shuffle and report what you see. Did it do what you thought it would? What proportions Blue do you have? What difference in proportions Blue?

    AWV

15. Generate at least 1000 trials and sketch the plot below.



16. What are the cases in the plot? (What does each individual dot represents? If you click on a bar, what changes in the display?)

Each dot represents a single re–randomized trial in with the 13 improved participants were randomly assigned to one of the two treatments. The location of the dot gives the difference in the proportion of improved in each group (dolphin therapy – control).

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

    It is centered at 0 because it represents the null model and the null hypothesis is that the treatment has no effect on improvement in which case we should see the same proportion of improved participants in both groups (or the difference in proportions should be 0).

18. The null hypothesis can be phrased: $H_0 : \ p_A = p_B$ or $p_{treat} = p_{control}$. Is the researcher's question looking for an increase, decrease, or change in either direction? Fill in the blank with <, >, or ≠ for the alternative hypothesis:

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

    This tells you whether to look in the left tail (<), or right tail (>), or both (≠) for results "this extreme or more extreme".
    Report the approximate p–value (i.e., strength of evidence) based on the observed result. (Reminder: we did this in the sleep deprivation study on Day 11.)

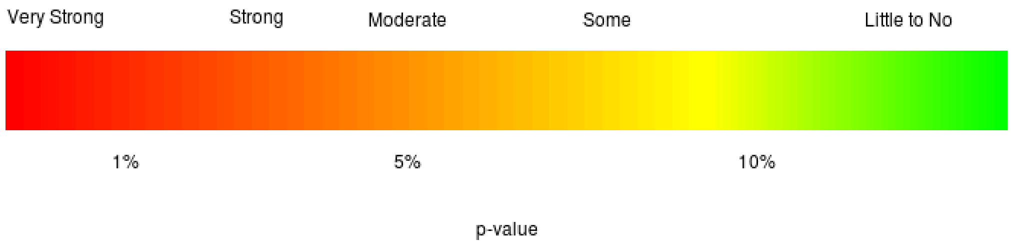
    Less than 0.001

    Collect another 1000 trials to see how much stength of evidence changes.

    I got 1 trial with a result as or more extreme than the observed, so the p–value was $1/2000 = 0.0005$.

The p–value gauges the *strength of evidence* against the null hypothesis. A small p–value says that the data are very unusual under our null assumption. How small is small? That depends on the application area. Common rules of thumb are:

- p–value < 0.01 is very strong evidence against the null.
- p–value < 0.05 is strong evidence against the null.
- p–value < 0.10 is some evidence against the null.
- p–value > 0.10 provides little or no evidence against the null.



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

    Very strong

20. Based on the p–value, provide an answer to the research question. *With a p–value of 0.0005, there is very strong evidence to reject the null hypothesis. We can conclude that in this sample, swimming with dolphins was therapeutic for patients suffering from depression as evidenced by a higher proportion of improved participants in the treatment group over the control group.*

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

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

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

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

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

    - The type of test used in the analysis (including the number of trials [shuffles]);
    - The null model assumed in the test;
    - The observed result based on the data;
    - The p–value for the test; and
    - The appropriate scope of inference based on the p–value and the study design. Include:
        - How were the subjects selected? If they are a random sample from some population, then our inference goes back to the population.
        - Were treatments assigned? If treatments were assigned at random, then we can state a causal conclusion.

A randomization test for a difference in proportions with 2000 trials was used to test the null hypothesis that dolphin therapy has no effect on whether or not a patient with depression will improve. In 30 participants split evenly and randomly between the treatment and control groups, 10 people in the treatment group improved whereas 3 people in the control group improved. This gave an observed difference in proportion of improved people between the treatment and control group is 0.467 (dolphin therapy – control). This resulted in a p–value of 0.0005, which constitutes very strong evidence against the null hypothesis. Since participants were randomly assigned to groups but do not represent a random selection of all patients with depression, we can conclude that in this sample, dolphin therapy caused a higher proportion of participants to improve as compared to a control group.

What is Normal?

Before computers became cheap and easily accessible, we could not use simulation techniques to find p-values, but had to use known distributions and tables of probabilities. Most stat intro courses still rely on z and t tables to do hypothesis tests. This activity relates the tools we're using to the older technology tools.

You have learned to use randomization web applets, and you can actually use them to analyze real data in the future. However, if you want to learn more powerful techniques – regression for example - you really should take more statistics and learn software better suited to higher powered methods of analysis. We are happy to visit with you about possibilities for more statistics (Stat 217 is a great second course).

As you read articles which discuss statistical results, you might see the word "simulation" or "randomization test", but will certainly see reference to tests built on the normal distribution – the familiar "bell–shaped" curve. This activity will show you how the randomization or permutation test on proportions is similar to the z-test based on normality.

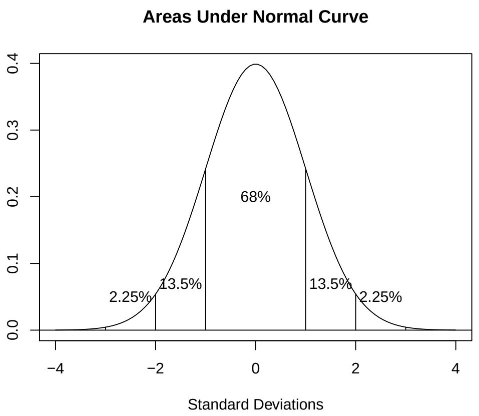
Normal Distributions

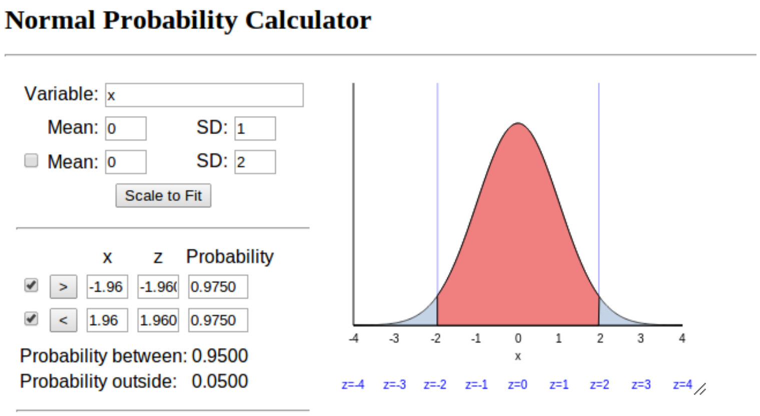
The classic "bell-shaped" curve we call the "normal" distribution has two parameters: the center μ and the standard deviation, σ. Changing the center slides the curve to the left or right on the number line. Increasing σ flattens and broadens the curve, while decreasing σ makes it tighter and higher. We mentioned, back on Day 10, that standard deviation is a measure of spread. Later we'll see how to compute the sample standard deviation, labeled s, which is commonly used to estimate σ, the true (unknown) population standard deviation.

Empirical Rule (meaning a rule which works well in practice):
All normal distributions have the same basic shape with area under the curve of 1. And all follow these rules:

- 68% of the probability (area under the curve) lies within one standard deviation of the mean.
- 95% of the probability (area under the curve) lies within two standard deviations of the mean.
- 99.7% of the probability (area under the curve) lies within three standard deviations of the mean. (We rarely see any data more than 3 SD's from the mean).

In other sections of Stat 216, students have to memorize the empirical rule. Rossman and Chance have another app called the Normal Probability Calculator which gives us any normal probabilities.





Normal Probability Calculator

Variable: x

Mean: 0 SD: 1

□ Mean: 0 SD: 2

Scale to Fit

		x	z	Probability
☑	>	-1.96	-1.960	0.9750
☑	<	1.96	1.960	0.9750

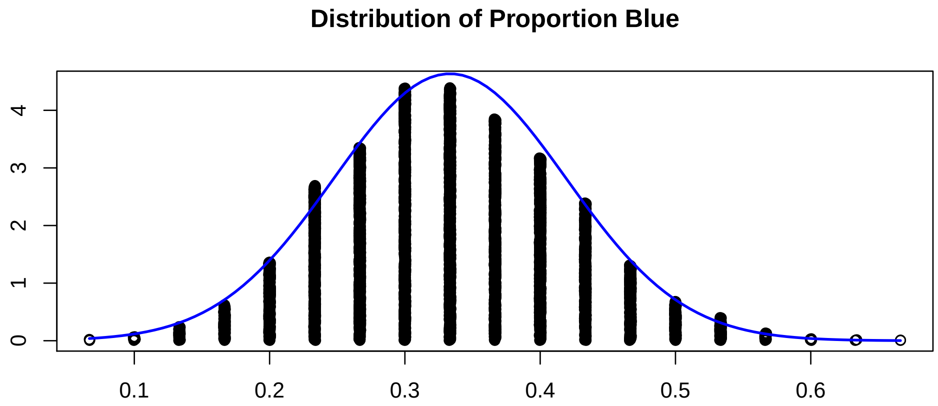
Probability between: 0.9500
Probability outside: 0.0500

The plot on the right from the web app shows us that the “Empirical Rule” is not exact. To get an area of exactly 0.95, we should go 1.96 σ’s from the mean in each direction. In practice, using 2 instead of 1.96 is “close enough” for most applications.

How do we use the Normal Distribution?

Important facts:
Statistics vary from sample to sample, and the pattern is predictable.

For sample proportions, the pattern is often close to that of a normal distribution. For example, here’s a distribution of 5000 repeated trials where we keep track of the number of blue balls in a sample of size 30 randomly sampled with replacement from a box containing 40 blue and 80 gold balls.



On top of the simulated distribution we’ve added the normal density curve with mean $\mu = \frac{1}{3}$ and standard deviation $\sqrt{\dfrac{\frac{1}{3} \times \frac{2}{3}}{30}}$, which is the standard deviation of the sample proportion. The theoretical curve is a good match to the simulation distribution, and we can compute probabilities from areas under the curve (instead of counting points and dividing by 5000) to find p-values.

Cautions: The simulation test we have used with sample size n requires:

- A population of size at least $10n$.
- A representative sample.
- Independent trials. (One person's response should not influence anothers.)

The normal approximation also needs those assumptions to be met. In addition it needs:

- Large enough sample size to expect at least 10 successes ($np_0 \geq 10$) and at least 10 failures ($n(1-p_0) \geq 10$) where p_0 is the value for p used in the null hypothesis.

Facebook and Job Applications

In 2009, researchers at CareerBuilder.com surveyed 2667 hiring managers and asked if they viewed applicants social media sites when considering them for a job.[4] Of the 2667 respondents, 1200 said they did, which seemed a bit of a change from 2008 when 43% of hiring managers looked at the social sites. We will test the hypothesis:

$$H_0: \ p = 0.43 \text{ versus } H_A: \ p \neq 0.43$$

using the a permutation test and the z-test.

1. Do the assumptions hold? Let's assume they took a simple random sample from a list of 50,000 managers.

   - Do these managers come from a population of size at least $10n$?

     Yes.
   - Is it a representative sample?

     I doubt it, because they had to volunteer to respond.
   - Are answers independent?

     Yes.
   - Is sample size large enough so that:

     Is $np_0 > 10$?

     Is $n(1-p_0) > 10$?

     Yes. $p_0 = .43$ *so* $np_0 = 2667 \times .43 = 1147 > 10$ *and* $n(1-p_0) = 2667 \times .57 = 1520 > 10$

[4] http://www.careerbuilder.com/

Discuss the assumptions and record your thoughts here.

2. Run the hypothesis test in the web applet called $\boxed{\text{One Proportion Inference}}$.

   - Where it says "Probability of Heads" put in the value for the null hypothesis. (Now click anywhere and it changes the label to "Probability of Success, π").
   - Change sample size to the number of managers surveyed.
   - Leave number of samples on $\boxed{1}$.
   - Change "Number of Successes" to "Proportion of successes".

   (a) What proportion of managers in 2009 view applicants social media sites? $1200/2667 = 0.45$

   (b) Click $\boxed{\text{Draw Samples}}$ to see one sample. Change the number and get several thousand samples. *I did 5000* Click "Summary Statistics". Where is the null distribution centered? Explain why.

   0.43, the null value for p

   (c) Copy down the "*SD.*" from your plot (we need it later).

   0.0094

   (d) How unusual is it to get a value for $\widehat{p}$ as big as .45 or bigger?

   happens only 0.016 of the simulation runs

   (e) How unusual is it to get a value as small as .41 or smaller?

   .017

   (f) Why do we need to compute both? What is the p-value?

   add the 2 together because we have a two–sided alternative. p–value is 0.033

   (g) At significance level $\alpha = 0.10$, what do you conclude about the null hypothesis?

   Reject the null.

   (h) Has there been a change in the proportion viewing social sites? State your decision.

   There is moderately strong evidence to suggest that the true proportion of managers who check social media pages of applicants is not 0.43, but is, in fact, larger than it was the previous year.

3. Now we'll do the same test using a z-test.

   (a) Where is the center of the distribution under the null hypothesis?

   0.43

   (b) In 2c the permutation applet gave us the standard deviation of the sampling distribution based on the points you generated. Another way to get that is to use a formula: $\sqrt{\frac{p_0(1-p_0)}{n}}$ or, in this case: $\sqrt{\frac{.43(.57)}{2667}} = 0.00959$

   How well does this agree with your answer in 2c?

   pretty close

   Note: We have two ways to describe the same measure of spread. Any statistic has a "Standard Error" (SE) which is the spread (that is, the Standard Deviation) of the sampling distribution of that statistic. In the traditional STAT 216 sections, we emphasize the formulas for SE. With simulations, it's easier to read SD from the plot of the sampling distribution. Either method should give similar results.

(c) Next we "standardize" by subtracting the mean and dividing by standard deviation of the statistic (SE).
$z = \frac{\widehat{p}-p_0}{SD} = \frac{0.450-0.43}{\sqrt{\frac{.43(.57)}{2667}}} = \frac{.02}{.00959} = 2.087$
Note: you do need to know how to use your calculator to compute the numerator and denominator of the fraction, and then divide to get the answer. One approach is to start with the denominator (bottom) and store it in memory. Then compute the numerator (top) and divide by the number in memory.

(d) Now open the normal probability app: http://www.rossmanchance.com/applets/NormCalc.html and check the square to the left of the $\boxed{<}$ under the line x z Probability. Put the z–statistic you just computed in under z. Check the second line as well and put in the negative of your z and switch $\boxed{<}$ to $\boxed{>}$. Now the probability outside will be your p–value.

.037

(e) How similar is this p-value to the one in 2f?

very similar: .038 versus .033

Is your conclusion the same? or has it changed?

No change. There is moderately strong evidence to conclude that the proportion of managers who check out applicants Facebook pages has gone up since 2008.

Confession: In the CareerBuilder survey from 2008, actually only 12% of managers looked at social media sites, then it jumped way up. We changed it to be closer to the 2009 proportion so that the p-values were not so small we couldn't compare them under the two methods.

4. **Dolphin Therapy** revisited.

In our last class, we tested a hypothesis about swimming with dolphins. We'll redo that analysis using a Z-test. Go back to your last activity and copy in these values:

Description	Value
Number in Dolphin group	*15*
Number improved in Dolphin group	*10*
Number in Control group	*15*
Number improved in Control group	*3*
Proportion to improve in group 1, $\widehat{p}_A$	*.667*
Proportion to improve in group 2, $\widehat{p}_B$	*.20*
Difference in proportions, $\widehat{p}_A - \widehat{p}_B$	*.467*
Overall proportion to improve, $\widehat{p}_m$	*.433*

(a) Rerun the Randomization test doing 3000 to 5000 trials, and report your p-value here.

less than 0.001

Leave the web app window open.

(b) In question 18 on Day 13, we phrased the hypotheses in terms of the true parameters, p_A and p_B. Copy the hypotheses here. (Use p_A and p_B with no hats, because a hypothesis is a statement about the true parameters, not about the numbers we observed in the samples.)

$H_0 : p_A = p_B$

$H_A : p_A > p_B$

Explain in your own words what the null hypothesis means.

Answers vary. Something like "No difference between the two groups in proportion improving."

(c) If the null hypothesis is true, which estimate from the table above gives the best overall estimate of the proportion who improved? Explain why.

$\widehat{p}_m = .433$ *Because there is really just one big group, and we should combine across the treatments to get a better estimate of* p.

(d) If the null hypothesis is true, what would you expect to see for the difference of means, $\widehat{p}_A - \widehat{p}_B$?

about 0

(e) To use the Z-test, we need the observed difference, $\widehat{p}_A - \widehat{p}_B$, the value we expect for it under the null, and a standard deviation of the sampling distribution (called the standard error of the estimate). There is a formula for standard error, but to keep our lives simple, go back to the web app window and copy it from that plot.

st. dev = 0.184

In traditional Stat216 we would use this formula:

$$SE(\widehat{p}_A - \widehat{p}_B) = \sqrt{\widehat{p}_m(1 - \widehat{p}_m)(\frac{1}{n_A} + \frac{1}{n_B})}$$

where $\widehat{p}_m$ is the marginal proportion of successes when we combine both treatment groups into one.
Build the test statistic:

$$z = \frac{\widehat{p}_A - \widehat{p}_B - 0}{SD} = .467/.184 = 2.54$$

(f) Use the http://www.rossmanchance.com/applets/NormCalc.html app as before to find the p-value.
Check the proper direction we need $\boxed{>}$ and enter the z from above. The p-value is:
.0055

(g) Rewrite your results (they should be quite similar to those from the last class) including:

- The type of test used in the analysis (Z-test does not have a "number of trials");
- The null model assumed in the test;
- The observed result based on the data;
- The p-value for the test; and
- The appropriate scope of inference based on the p-value and the study design.

We used a z-test to compare the proportion of depressed patients who got better after swimming with dolphins to the proportion improved in a control group (swimming only). The test statistic was 2.54 for a one sided p–value of 0.0055. Based on the small p–value, we reject the null hypothesis that the treatment is ineffective ($p_1 = p_2$), and conclude that (due to randomized treatment assignment) the dolphin therapy caused the increase in improvement. The patients used were a convenience sample, so we cannot extend the causal inference back to some larger population of depressed patients.

Birth Weights

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

Discuss

1. Could there be some physiological reason why birthweights for the children of the 28 smokers might differ from the birth weights of the babies born to nonsmokers? Write down what you and your group know about smoke and nicotine to hypothesis a connection to birthweight.
2. If the connection you are thinking about is real, would it tend to increase or decrease birth weights of babies born to smokers? Or could the effect go either way?
   (a) What is the response variable in this study? Is it quantitative or categorical?
   birth weight, quantitative
   (b) Is there an explanatory variable in this study? If so, name it and tell which type of variable it is.
   Yes, habit, and it is categorical
   (c) Enter the data (on D2L) into the Rossman–Chance applet http://www.rossmanchance.com/applets/Dotplot.html Compute means for weight by habit and compute the difference in means.
   Smoker: 6.306, nonsmoker: 7.084, difference: -.778
   (d) Is the difference between the means large enough to convince you that babies born to smoking mothers are lighter than those born to nonsmokers? Why or why not?
   It does not seem like a difference of .78 is that large because of the spread of the plots (range from 1.7 to 8 for for smokers, 1.4 to 10 for nonsmokers).

Studies that Use Random Sampling

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

A random sample is one in which the method used to choose the sample from the population of interest is based on chance. Although there are many types of random sampling, the term is often associated with simple random sampling, in which all possible samples of a given fixed size are equally likely. All single units have the same probability of being selected. All pairs of units have the same probability of being selected, etc.

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

With data on birth weights, can we assign a baby to have a smoking versus nonsmoking mother? The habit variable splits these subjects into two populations, and we have a sample from each. In studies with random **sampling**, the goal is to describe the sample data, to compare groups, and infer any differences back to the broader population(s) from which the sample(s) was/were drawn. Even though we might have reasons from animal studies to think smoking causes certain changes, we do not expect these data to provide **causal** evidence of such a connection. We might want to know how large the difference in means (the true population means) is between two groups. That's really an estimation question which we'll tackle in Unit 3.

Consider the study on dolphin therapy. We found strong evidence of a difference due to the therapy, but the inference only applies to the people in the study. It did not allow us to say who the treatment is effective for. Is it all depressed people? All depressed people between the ages of 18–65? All depressed people between the ages of 18–65 who have a clinical diagnosis of mild to moderate depression? Is it restrivted to one city or state? There is a lot we don't know about how these people were selected.

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

MODELING THE BIRTH WEIGHTS

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

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

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

   - Open `http://www.rossmanchance.com/applets/` and select Two Means under "Statistical Inference".

- Clear all the data in the box and paste in the birth weight data. It's on D2L or here: http://www.math.montana.edu/~jimrc/classes/stat216/data/babyWeight.txt
- Copy the means and difference in means here.

smoking: 6.31, nonsmoking: 7.08, difference: 0.778

7. Generate 1 trial (sample). What is the mean birth weight to smokers from this simulated trial?

   Answers will vary, I got 6.5

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

   7.05

   What is the difference in means between these two groups?

   0.53

   Evaluate the Results

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

9. What are the cases in the plot? What changes when you click on a bar?

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

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

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

11. We're not told exactly what the researchers were thinking ahead of time, but let's assume that the alternative hypothesis is that smoking moms tend to have lighter babies. What is the alternative hypothesis of interest? Do you need to count $\boxed{\text{Greater than}}$ or $\boxed{\text{Less than}}$ or both as "more extreme" results?

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

12. Put the observed difference in the little box under the plot, chose the proper direction for comparison, and report the approximate p–value (i.e., strength of evidence) based on the observed result.

    0.013

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

    Strong to very strong evidence against the null model.

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

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

15. Can the researchers generalize the results to the population of all births in North Carolina? to all births in the US? Why or why not?

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

16. Can the researchers say that the difference in the average birth weight is caused by the mother smoking habit? Explain. If not, provide an alternative explanation for the differences.

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

17. Write–up the results of the simulation study. When reporting the results of a simulation study, pertinent details from the analysis that needs to be included are:

    - The *type of test* used in the analysis (including the number of trials);
    - The *null model* assumed in the test;
    - The *observed result* based on the data;
    - The *p–value* for the test, whether it is one or two sided; and
    - The *scope of inference* based on the p–value and study design.

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

Did She Murder Patients?

In the mid 1990's, a nurse named Kristen Gilbert was working at a Veteran's Administration Hospital in Massachusetts. Other nurses at first thought she was good at saving patients who where having heart attacks, but as the number of patients with these symptoms seemed particularly high when Nurse Gilbert was working, they began to suspect she was doing something to cause such episodes. Supervisors also began to notice that epinephrine (similar to adrenaline) was missing. Injecting patients with epinephrine can cause heart failure. The question became:

Were deaths more likely to occur during shifts when Kristen Gilbert was working than on shifts when she was not working?

Gilbert was put on trial for murder, and the following table was presented as evidence. It categorizes every shift during the years when Gilbert was on staff as a shift with a death (or no death) and with Gilbert working (or not).

	Gilbert Status		
A death occurred	Working	Not Working	Total
Yes	40	34	74
No	217	1350	1567
Total	257	1384	1641

1. Among all shifts (the total column) in what proportion of shifts did a death occur?

   4.5%

2. When Gilbert was NOT working, in what proportion of shifts did a death occur?

   2.4%

3. What is the standard error of your estimate? $\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$

   .004

4. When Gilbert was working, in what proportion of shifts did a death occur?

   15.6%

5. What is the standard error of your estimate?

   0.0225

6. Was it more likely to have a death when she was or was not working?

   was

7. Subtract the proportion of shifts with a death when Gilbert was working from the proportion of shifts with a death when Gilbert was NOT working.

   .132

8. Discuss: what factors, in general, make it more likely to have a death during a shift? List 2 or 3

   Time of day? (more deaths in early morning?) clusters due to hospital conditions?

9. Is the difference in proportions by itself convincing evidence that Gilbert was causing deaths?

   No, there is nothing to tell us how unusual this result is.

10. Compare to data used in recent activities. How is this similar to and different from:

    - Sleep Deprivation study
      Similar: categorical response
      different: no assignment of treatments, not a random sample
    - Dolphin Therapy
      Similar: categorical response
      different: no assignment of treatments
    - Birth Weights
      similar: no assignment of treatments
      different: 'Death on shift' is categorical response instead of continuous

11. Look back at the research question. This is a question about a population of shifts. Explain how the data relate to that population.

    All shifts during that time period are shown. It's a census.

Observational Studies

One of the biggest differences between this study and the previous ones is that this study does not implement random sampling nor random assignment. Gilbert was not randomly assigned to the 8–hour work shifts. A situation such as the case against Kristen Gilbert, where the cases are observed as they occur naturally rather than randomly assigned by researchers, is called an observational study.

The purpose of an observational study is to describe some group or situation. You may not be able to make any inference.

You typically cannot draw cause–and–effect conclusions from observational studies, because the possibility of alternative explanations always exists. Randomized experiments (using random assignment), such as the dolphin therapy and sleep deprivation studies, do allow for cause–and–effect conclusions when the observed experimental results are found to be very unlikely to occur under the null model of no difference between the groups. Because this situation is an observational study and Kristen Gilbert was not randomly assigned to work shifts, there are many other explanations that are possible and even plausible (you thought of a few in your response to question 8).

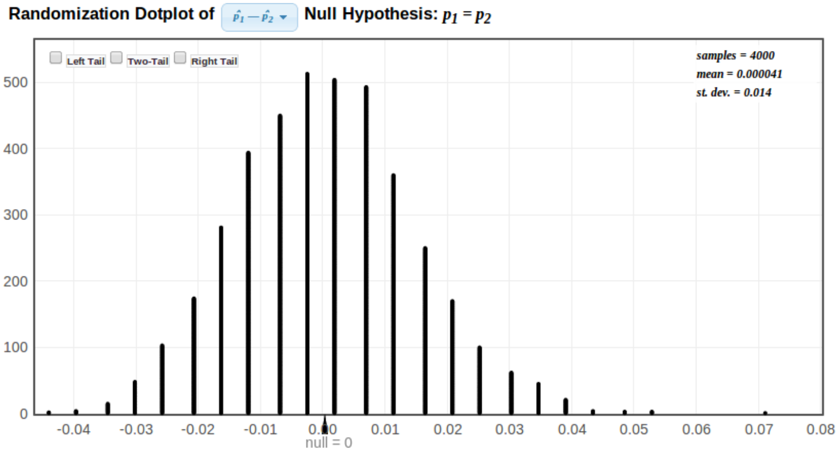
Observational studies can have random sampling, like in the Birth Weight activity, and therefore can make generalizations to the population. However, in the Gilbert study, you cannot generalize the results to the population because it is not a random sample of shifts from the population.

Modeling the Nurse Gilbert Data

Conduct a permutation test using `http://www.rossmanchance.com/applets/ChiSqShuffle.html?dolphins=1` to find out how unusual these data are, assuming there is no difference between the percent of shifts in which a death occurred when Gilbert was working and those in which she wasnt working. Follow the same steps as for the Dolphin study, but you will have to decide whether to count “Death” or “No Death” as a “success” and whether Group A is when Gilbert worked or when she didn’t.

Evaluate the Results

12. Sketch the plot.



13. Which tail or tails include stronger evidence against Gilbert than the observed result?

    Right tail (more evidence would be a greater difference or Gilbert having even more shifts were a death occurred).

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

    Less than 1/1000 = 0.001

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

    EXTREMELY strong!

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

    We have strong evidence to conclude that deaths were more likely to occur when Gilbert was on shift.

17. If we use the Z test on these data, we get a p–value very close to 0 ($< 2.0 \times 10^{-16}$). Is your conclusion the same?

    Yes. both p–values are tiny, very strong evidence against the null.

18. Can the researchers generalize the results to the population of all 8–hour work shifts? Why or why not?

    Our data are the whole population of all shifts. It doesn't make sense to think of a bigger grouop of shifts, so the answer is that results do infer back to the sample = population, and that's it.

19. Can the researchers attribute the difference in the percent of deaths to the Kristen Gilbert (when she was working vs. when she wasn't working)? Explain. If they can't, provide an alternative explanation for the differences.

    No because we did not randomly assign shifts. These are when she worked. Again, she could have consistently worked shifts were more deaths occur (like at night?).

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

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

    A randomization test for a difference in proportions with 1000 trials was used to test the null hypothesis that the proportion of shifts in which a death occurred was not associated with whether Kristin Gilbert was working or not. The proportion of shifts in which a death occurred under her watch was 15.6%, compared to just 2.5% when she was not working, giving an observed difference in proportion of shifts in which a death occurred of 13.1%. According to our null distribution, this is extremely unlikely to occur (p–value less than 0.001). This gives strong evidence to conclude that there is an association between if Kristin was working and the proportion of shifts in which a death occurred in this sample (which is of all shifts during her tenure at this hospital).

21. How well is your group working together? List three things that could be improved.

Pregnancy Tests and Errors

Pregnancy tests have evolved greatly over the years. Many of the home pregnancy tests make strong claims. For example, the First Response Gold® Digital Pregnancy Test claims it can give "results as early as 5 days before the day of your missed period".[5]

An important question is

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

To answer this question, researchers conducted a clinical trial with 215 women who were trying to become pregnant[6]. The women took the First Response Gold® Digital Pregnancy Test daily starting 5 days before the day of their expected period. We will use their data to learn how to characterize the accuracy of the test.

The results for the tests:

	Test Results:		
Truth:	Positive	Negative	Total
Pregnant	58	79	137
Not Pregnant	4	74	78
Total	62	153	215

1. What is the proportion of women in the study were actually pregnant?

   137/215 = .637

2. What is the proportion of women in the study were not actually pregnant?

   78/215 = .363

The two main measures of accuracy of a diagnostic test are known as **sensitivity** and **specificity** and a good test will have high numbers (close to one) for each.

- A test is *highly sensitive* if people who *have the disease or condition*, are usually detected and assigned a positive result.
- A test is *highly specific* if people who *do not have the disease or condition*, are usually correctly identified and are assigned a negative result.

In terms of a generic table, we'd like to see high numbers on the diagonal.

[5] Church & Dwight Co., Inc. (n.d.). First Response® – We tell you first Pregnancy Test Products. Retrieved from http://www.firstresponse.com

[6] Cole, L. A. (2011). The utility of six over–the–counter (home) pregnancy tests. *Clinical Chemistry & Laboratory Medicine*, 49(8), 1317–1322. doi: 10.1515.CCLM.2011.211

	Test Results:		
Truth:	Positive	Negative	Total
Has Disease	True Positives	False Negatives	All Diseased
No Disease	False Positives	True Negatives	All Healthy
Total	All Positives	All Negatives	Total Tested

Definitions:

- Sensitivity = probability of positive results when the person has the disease or condition.
- Specificity = probability of negative results when the person does not have the condition.

3. The sensitivity for the First Response Gold® Digital Pregnancy Test is the probability of testing positive when a patient actually is pregnant. Based on these data, the sensitivity estimate is:

   58/137

4. Describe in words how to compute the sensitivity.

   The number of people who test positive and have the disease divided by the total number of diseased people

5. The specificity for the First Response Gold® Digital Pregnancy Test is the probability of testing negative when a patient actually is not pregnant. Based on these data, the specificity estimate is:

   74/78

6. Describe in words how to compute the specificity.

   Number of people who test negative and do not have the disease divided by the total number of people who are not diseased

STOP

Compare your answers to the above questions with another group.

The **false positive rate** of a test is the probability the test result is positive if the patient does not have the disease/condition. (The test was fooled into giving a positive result when it shouldn't have).

7. Describe in words how to compute the false positive rate. Do so for the pregnancy test data.

   Number of false positives (test positive but not diseased) divided by those not diseased. 4/78

8. How is the false positive rate related to specificity in this case?

   1 – specificity

9. Will it always have this relationship to specificity?

   YES!

   The **false negative rate** is the probability the test result is negative if the patient does have the disease/condition. A false negative occurs when the pregnancy test is negative, but the patient actually is pregnant.

10. Describe in words how to compute the false negative rate.

    Number of false negatives (tested negative but do have disease) divided by the total number of diseased.

11. Find the false negative rate of the First Response Gold® Digital Pregnancy Test for pregnancy.

    79/137

12. Describe in words how the false negative rate is related to specificity and/or sensitivity.

    false negative rate = 1 – sensitivity

13. Consider a woman who takes the test hoping it will show she is not pregnant. If the test returns negative, what is the probability that she really is not pregnant (based on these data)?

    74/153

14. Is the sensitivity or specificity of more interest to her? Explain.

    Can argue either way. Specificity: the probability she tests negative when she is actually not pregnant because she wants to make sure she is not pregnant. Sensitivity because if she tests positive, she better be pregnant or else she is stressed for nothing. Usually easier to think about which error would be worse (thinking she was pregnant when she isnt stressful until she finds out otherwise – or thinking she is not pregnant when she is bad for the baby if she is drinking/doing drugs, etc.)

Justice System and Errors

You should have read the materials here:
http://www.intuitor.com/statistics/T1T2Errors.html before today's class.

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

When we reject the null hypothesis, we could be making a "Type I" error, because the null could really be true.

When we fail to reject the null hypothesis, we could be making a "Type II" error, because the null could really be false.

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

Note: We set up the null hypothesis hoping to reject it. It's like a straw man we hope to topple over. We control type I error by setting an α level (usually .10, .05, or .01)

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

1. The null hypothesis has to be wrong, but it could be wrong just by a small amount or by a large amount. For example in the shoe experiment, we did not reject the null hypothesis that treatment and control shoes were equally effective. We could be making a type II error. If in fact, if there was a small difference, it would be hard to detect, and if the treatment shoes were far better, it would be easy to detect. This is called the effect size, which is [difference between null model mean and an alternative mean] divided by standard deviation.

2. Sample size. P–values are strongly affected by sample size. With a big sample we can detect small differences. With small samples, only coarse or obvious ones.

3. Significance level. The fence, usually called $\alpha =$ alpha, is usually set at .10, .05 or .01 with smaller values requiring stronger evidence before we reject the null hypothesis.

Instead of limiting the probability of Type II error, researchers more often speak of keeping the power as large as possible. Power is one minus the probability of Type II error. Go to the Power Demo page: `http://spark.rstudio.com/jimrc/powerDemo`

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

    0.484

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

    13

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

    0.484 Smaller, down to 1.4

    What is your effect size?

    $2 - 0 = 2/1.4 = 1.429$

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

    Larger, 2.9

    What is your effect size?

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

16. How do the effect sizes in 14 and 15 compare?

    Larger effect size in 15

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

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

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

    .971

    Change it to 0.10. What is the power?

    .992

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

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

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

    33

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

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

    31

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

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

Unit 2 Wrapup

Vocabulary

- Symmetric versus skewed distributions
- Experiment versus Observational Study
- Treatment Variable or Factor
- Response Variable
- p-value and strength of evidence
- Null Model
- Random Assignment (why do we do it?)
- Anecdotal Evidence
- Simple Random Sampling
- Lurking Variables
- Randomization Test
- Sensitivity
- Specificity
- False Positive Rate
- False Negative Rate
- Type I Error probability is limited to alpha
- Type II Error how does its probability relate to power?
- What settings affect power of a study?
- What points must be included in a statistical report?

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

Study	Experiment?	Explanatory Variable?	Response?
Energy Drinks	Exp	RED–A vs Control	RBANS improvement
Sleep Deprivation	Exp	Sleep Deprived/Not	Improvement in Score
Running Shoes	Exp	Shoe type	Pain
Dolphin Study	Exp	Swim w/ or w/o dolphins	Improved Depression (Yes/No)
Birth Weights	Obs	Mexico or Other country	Ed Attainment
Murderous Nurse	Obs	Did Gilbert work? (Yes/No)	Death during shift (Yes/No)

Extensions

2. Sleep Deprivation Study

   (a) Think about how your analysis and conclusions might have changed if you had subtracted the group means in the other direction (sleep deprived mean – unrestricted sleep mean). *Our difference in means would be negative.*

   (b) What parts of your analysis would have been the same, and what parts (if any) would have turned out differently? How would they have been different (if at all)?

   (c) How would your conclusion about the study have changed (if at all)?

   The statistic would change signs and the alternative hypothesis would be < instead of >, so we would look in the left tail instead of the right to get our p–value. P–value and conclusions would not change.

   (d) Investigate your predictions by making this change and re-conducting your analysis.

   Investigate the effect that one observation can have on this analysis.

   (e) Remove the improvement score of 45.6 from the unrestricted sleep group, and re-conduct the analysis. Comment on how much impact this one observation has on your analysis and conclusion.

   Difference in means changes to 13.06 and the p-value is 0.018 after 4000 trials, so evidence got slightly weaker. However, we still make the conclusion that we have strong evidence against the null hypothesis.

   (f) Restore the 45.6 value but remove the -7.0 improvement score from the unrestricted sleep group, and investigate the effect of that change.

   Difference in means changes to 18.09 and the p-value is 0.0015 after 4000 trials, so evidence got much stronger. However, we still make the conclusion that we have strong evidence against the null hypothesis.

   (g) Notice that this research study involved slightly different numbers of people in the two groups. Suppose that you describe this study to a friend, and he argues that the study is invalid because of the unequal group sizes. Describe how you would respond to your friend, and be sure to include a description of how your analysis took these unequal group sizes into account.

   A permutation test has no trouble drawing randomized groups of different sizes, it just copies the shape of the original data. The test is valid for any sample sizes. Having equal sample sizes is preferable to get better power, but not necessary.

3. Dolphin Therapy Study

   (a) Suppose the results of the experiment had been that 11 had improved in the dolphin group (instead of 10) and only 2 had improved in the control group (instead of 3). Explain how your approximate p-value would have been different in this case. Also describe how the strength of evidence for the benefit of dolphin therapy would have changed.

   One more improved patient in the tretment group and one fewer in control make this even stronger evidence against the null hypothesis. The difference in proportions *becomes .60, and in 5000 randomization trials, I never got one sample with this large a difference in sample proportions, so p–value is* $< 1/5000 = .0002$ *Our conclusion is the same.*

(b) Suppose the results of the experiment had been that 8 had improved in the dolphin group (instead of 10) and 5 had improved in the control group (instead of 3). Explain how your approximate p-value would have been different in this case. Also describe how the strength of evidence for the benefit of dolphin therapy would have changed.

With 8 improved patients in the tretment group and 5 in control the evidence is weak to none against the null hypothesis. The difference in proportions becomes .20, and in 5000 randomization trials p–value is .22 *Our conclusion is reversed.*

(c) Suppose the study had involved exactly twice as many subjects, 60 instead of 30, and suppose that the same proportions had improved in each group (dolphin, control). Describe what would have changed in how you set up the simulation analysis. Then make a prediction, and explain your reasoning, for how the approximate p-value, and the strength of evidence for the benefit of dolphin therapy, would have changed. Finally, conduct the simulation analysis for this new situation, and comment on whether your prediction was confirmed or refuted.

A good guess is that the same proportion in the larger study provides stronger evidence. It does. When I run 5000 trials with 60 patients, the differences in means in the plot only go up to .4, never to .467, so the p–value is $< .0002$

More Examples

4. Teen Hearing Study

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

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

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

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

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

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

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

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

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

5. Mammography Study

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

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

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

160/10000 = .016 = 1.6%

(b) Describe in words what the sensitivity is in the context of this problem.

Sensitivity is the ability of the mammogram to detect cancer in the breast of a woman with breast cancer.

(c) Find the sensitivity of the mammography test.

70/160 = .438 = 43.8% *That seems poor!*

(d) Describe in words what the specificity is in the context of this problem.

Specificity is the ability of the mammogram to come back clean for a woman who does not have breast cancer.

(e) Find the specificity of the mammography test.

9140/9840 = 0.929 = 92.9% *Not bad.*

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

(f) Given the consequence of a false test result, is the sensitivity or specificity more important in this case? Explain.

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

(g) Find the false positive rate of the mammography test for breast cancer.

700/9840 = 0.071 = 7.1%

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

(h) Find the false negative rate of the mammography test for breast cancer.
$90/160 = .563 = 56.3\%$

6. Blood Pressure Study

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

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

(a) Do the data in this study come from a randomized experiment or an observational study? Explain.
Randomized experiment because the treatment (talk or count) was assigned randomly.

(b) Calculate the difference in the means.
2.625

(c) Write out the null model for this study.
Mean blood pressure is the same for people talking or counting.

(d) Do the appropriate test to determine if a difference this large could reasonably occur just by chance. Comment on the strength of evidence against the null model.
Running 5000 trials of a randomization test, I got a p–value of .101 which gives only weak evidence against the null hypothesis of equal means.

7. Social Fibbing Study

A student investigated "social fibbing" (the tendency of subjects to give responses that they think the interviewer wants to hear) by asking students "Would you favor a policy to eliminate smoking from all buildings on campus?" She randomly assigned half the subjects to be questioned by an interviewer smoking a cigarette and the other half were interviewed by the same student but not while she was smoking. The results are displayed in the following table.

	Favor Ban	Not Favoring Ban	Total
Smoking	43	57	100
Not smoking	79	21	100
Total	122	78	200

Does the behavior of an interviewer affect the responses of the people being surveyed? Do the appropriate test to determine if a difference this large could reasonably occur just by chance. Comment on whether the difference in the percents provides strong evidence against the null model.
Running 5000 trials of a randomization test, I got a p–value of less than .002 which gives very strong evidence against the null hypothesis of equal means.

Unit 3

What Are Plausible Values for p?

So far in this course we have

1. Looked at random mechanisms for generating data (Unit 1) and we saw that we cannot predict an individual outcome of a random process, We can predict a pattern of possible outcomes.

   True or False: "Random" means all outcomes are equally likely.

2. Used the hypothesis of "no difference" to compute the strength of evidence against the hypothesis that two means or proportions were equal. (Unit 2)

   Which provides **stronger evidence**: a smaller or larger p-value?

Now we will combine both to use a **statistic** to estimate a **parameter**.
Problem: when giving a single point as an estimate of a parameter we are almost certainly wrong (hopefully close, but not exactly right).

Solution: give an interval estimate instead of just one point. Typically interval estimates are centered at our point estimate, but also include nearby points which are also "plausible". How do we separate "close" from "far"?

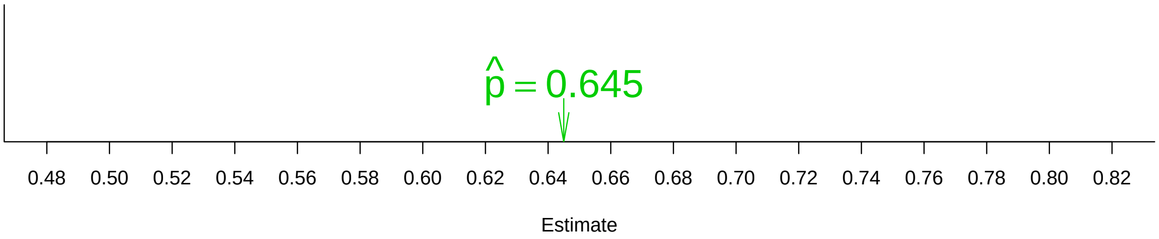
Use **p-value** (or Reject/Fail-to-Reject) as a yardstick.

Close: if the statistic is near the null hypothesis value, the p-value will be **large** and we fail to reject.

Far: if the statistic is far from our null hypothesis value, the p-value will be **small** and we reject.

Kissing the Right Way. 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.

Let's test to see if the true proportion of kissing couples leaning right is 50%. Set up the null and alternative hypothesis. Note: it's possible that more or less than 50% of kissing couples actually lean to the right, so it makes sense to use a two-sided alternative.
We will **use 10% as a cutoff** for "statistical significance. Use this number line to keep track of the plausible values:



Open the single proportion applet:
`http://www.rossmanchance.com/applets/OneProp/OneProp.htm`
and enter probability of heads = $\boxed{0.50}$ and number of tosses = $\boxed{124}$.

3. Run several thousand repetitions.

   (a) What is the center of the sampling distribution? (The summary stats options helps.)

   (b) Change "Number of heads" to "Proportion of heads". Now what is the center of the sampling distribution?

   (c) Go back to "Number of Heads" and get a count of samples as or more extreme than the one observed (80). What proportion are this or more extreme?

   (d) Click the "Two–sided" box. How does the proportion more extreme change?

   (e) For a two-sided test: $H_0: \ p = 0.50$ versus $H_A: \ p \neq 0.50$, would we reject H_0 at the 0.10 sigificance level?

   (f) If we reject, then the null value of 0.50 is not consistant with the data. If we fail to reject, then the null value of 0.50 is not too far from the observed $\hat{p} = 0.645$, so it is a plausible value.
   Write out your conclusion using a significance level of 0.10. Be sure to include in this statement the value we are using for the null hypothesis and whether we reject or "fail to reject" it. *At the 10% significance level we reject the null hypothesis that the true mean proportion of couples leaning right while kissing is 50%.*

   (g) Is p = .50 a "plausible value" given this sample? Write "yes" or "no" here and on the number line above just above 0.50. *No*

   (h) Will values further away from 0.50 be plausible? Answer here and mark the number line. *Values below 0.50 will also be implausible.*

4. Now consider another value for p, let's take $p = 0.56$, just because it's about halfway between 0.50 and our observed sample proportion of 0.645. Change the "Probability of

heads" to 0.56, click Draw Samples, and repeat the above steps to see if we Reject or Fail to reject.

(a) Generate 3000 samples with $p = .56$. Click two-tail. Do we reject this H_0? *Yes we reject* $H_o : p = 0.56$ *because the p-value of 0.0645 is less than 0.10. It is too far from 0.645 to be considered plausible.*

(b) State your conclusion using a significance level of 0.10. Be sure to include in this statement which value we are using for the null hypothesis and whether we reject or "fail to reject" it. *At the 10% significance level we reject the null hypothesis that the true mean proportion of couples leaning right while kissing is 56%.*

(c) Is $p = .56$ a "plausible value" given this sample? Again mark it as good or bad on the above number line. *No. It is too far away from 0.645 to be considered plausible.*

5. Try $p =$ 0.57 which is a bit closer to our observed sample proportion of 0.645. Change the "Probability of success" to 0.57 and repeat the above steps.

   (a) Generate 3000 samples with $p = .57$. Click Count and two-tail. Do we reject H_0? *No. P–value = 0.1003. We fail to reject* $H_0 : \ p = 0.57$.

   (b) State your conclusion using a significance level of 0.10. Be sure to include in this statement which value we are using for the null hypothesis and whether we reject or "fail to reject" it. *At the 10% significance level we fail to reject the null hypothesis that the true mean proportion of couples leaning right while kissing is 57%.*

   (c) Is $p = .57$ a "plausible value" given this sample? Mark it as good or bad on the above number line. *Yes.*

   (d) We need to keep a list of plausible values. Will values closer to 0.645 be plausible? *Yes.*

   (e) What is the smallest plausible value we've found? *0.57*

6. Next we'll look at values larger than the sample proportion, 0.645. Try $H_0 : \ p =$ 0.72. Go through the same steps to see if this is a plausible value. State your conclusion and mark "good" or "bad" on the number line. *At the 10% significance level we reject the null hypothesis that the true mean proportion of couples leaning right while kissing is 72%. We'll mark it as "bad".*

7. If 0.72 was "bad" move closer to the sample proportion. If it was plausible, move further away. Keep testing and marking points as good or bad until we have zeroed in on a dividing line between the two zones (two place accuracy is enough). *Moving down to 0.71, I get a p–value of 0.116, so at the 10% significance level we fail to reject the null hypothesis that the true mean proportion of couples leaning right while kissing is 71%.*

8. Write the plausible values as an interval estimate for p using parentheses to include the largest plausible value less than the sample proportion to the smallest value larger than the sample proportion. *(0.57, 0.71)*

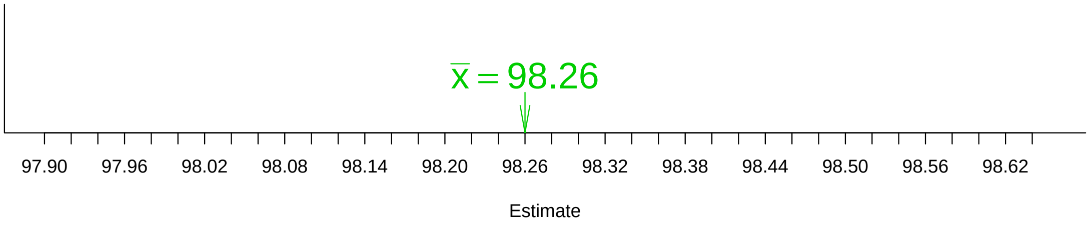
   Congratulations! You just built an interval estimate for the unknown parameter which incorporates the variability of the sampling distribution. The cutoff we used for rejection was 0.10, meaning each test had a 10% chance of making a type I error. The interval uses the center part, rather than the two tails, so we subtract from 100% and say "We are 90% confident that this interval contains the true proportion of right-leaning kissers". Use the word "confidence" not "probability".

9. Write out the meaning of this interval. (You need to repeat the sentence in quotations with the actual numbers in it instead of "this interval".) *We are 90% confident that the interval (0.57, 0.71) contains the true proportion of right-leaning kissers*

Confidence Interval for a Mean

Next we'll apply the same process to estimating mean body temperature. The guess-and-check process is a bit slow, (we'll soon speed it up) but the point of the activity is to connect hypothesis testing to confidence intervals. You know all about the first, so it will help you understand this new concept.

10. To work with a mean, use the short cut of theory (like Z test for proportions) at http://www.rossmanchance.com/applets/TBIA.html
Change $\boxed{\text{One Proportion}}$ to $\boxed{\text{One Mean}}$. We have data on body temperatures from $n = 50$ students in an intro stats class. Their average temperature was $\overline{x} = 98.26$ with standard deviation $s = .765$. Enter these values into the applet boxes.
These data come from a convenience sample of students.



11. Click "Test of Significance". We've always been told that normal body temperature is 98.6° F. Enter $\boxed{98.6}$ as the value in both null and alternative boxes. Change the alternative direction to $\boxed{\neq}$ because confidence intervals are "two–tailed". Click $\boxed{\text{Calculate}}$ to get the p-value. This time, use a significance cutoff of 0.05. (above we used $\alpha = 0.10$ and had 90% confidence.)
0.95 Is 98.6 a plausible value? *No. The p–value is 0.0028, so we reject* $H_0 : \mu = 98.6$.

12. Think about distance and fill in:
We reject the null hypothesis when it's value is ______________________ the observed mean.
far from

13. We can change the null value in the box which says $\boxed{\text{Null hypothesis: } \mu = \underline{\hspace{2cm}}}$, and then we check it against the data to see if it is a plausible value. We judge it plausible if the p–value is large and we fail to reject H_0.

Try a null hypothesis value of $\boxed{98.5^\circ}$ F. What is the p–value? Mark it as bad (reject) or good (FTR) on the number line. *Implausible, p–value = 0.03 so 98.5 is "bad".*

14. How far down the temperature scale do you have to go to find a plausible value? (We want to find the largest plausible value above 98.2.) Mark it on the number line as "good". *98.47 is close enough to 98.26 to be considered plausible. Mark it as "good".*

15. Next we need to go to values below 98.2 to find a value which is less than our sample mean, but still plausible. Try $\boxed{98.00}$. Is it plausible? If so, go a bit lower until you find a value which is too small. If not, move up a little. Keep track using the number line. *Not quite close enough. 98.00 has a p–value of 0.02, so it is "bad". I had to move up to 98.05 to get a p–value of 0.058. I fail to reject* $H_0 : \mu = 98.05$.

16. Write our 95% confidence interval of plausible values for the population mean body temperature inside parentheses. *(98.05, 98.47)*

17. Interpret the interval as you did for the 90% CI for p in number 9. Two differences here: we created a confidence interval for μ, not p, and the confidence level is different. Include a description of what group of people's mean body temperature we are aiming to cover. (The scope of inference here just depends on whether or not we have a random sample from a population, because no comparison - no causal inference – is being made.) *We are 95% confident that the interval (98.05, 98.47) contains the true mean body temperature in degrees Farenheit for the sample of students used.*

Reference

Güntürkün, O. (2003), Human behaviour: Adult persistence of head-turning asymmetry, *Nature*, 421, 711.

Margin of Error

The Gallup organization does weekly polling on many different issues. For example, they report that 62% of Americans have a "great deal" or a "fair amount" of trust in the judicial branch headed by the supreme court. The fine print explains how they conducted their poll including sample size and something called "margin of error".

> Survey Methods
> Results for this Gallup poll are based on telephone interviews conducted Sept. 5-8, 2013, on the Gallup Daily tracking survey, with a random sample of 1,510 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia.
>
> For results based on the total sample of national adults, one can say with 95% confidence that the margin of sampling error is 3 percentage points.

http://www.gallup.com/poll/165248/americans-still-divided-approval-supreme-court.aspx

1. The estimate of 62% with margin of error 3% provides a 95% confidence interval for the true proportion of Americans who trust the judicial system. What is their 95% confidence interval? *(59, 65)%*

2. How many standard errors wide is this confidence interval? Before we used a hypothesized value to compute SE. When estimating and building confidence interval, use the observed $\widehat{p}$ as in

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

For 95% CI, it should be 4 SE's wide.

The "plausible values" technique we used in the last activity illustrates the connection between confidence intervals and hypothesis tests, but is rather a cumbersome procedure requiring lots of guessing and checking. Having gone through that a few times, you can now appreciate a simplified procedure which gives nearly the same confidence intervals.

First, a reminder: collecting more data gives greater precision. Any statistic has a standard deviation which tells us how much spread there is to its sampling distribution. When we estimate that spread, we call it "standard error".

As n, the sample size, gets bigger, standard error and spread get smaller, so our estimate has greater precision.

Secondly, on Day 14 we used the normal distribution's bell-shaped curve to approximate p-values when testing proportions. That works for confidence intervals, as well.

3. Go to the Rossman–Chance site and click $\boxed{\text{Normal Probability Calculator (js)}}$.

(a) Leave mean set to 0 and SD set to 1. Below the divider line you should see

```
x    z   Probability
```

Check both boxes to the left and enter .90 and .10 in the two boxes under "Probability". What is the probability in the pink area? What z values are the cutoffs? *80%, -1.282 and 1.282*

We use these z values to get an 80% confidence level.
Alternative: It might be easier to think, "We want 10% in the upper tail", rather than 90% in the lower tail. You can put .10 in each box and change one of the $<$ boxes to $>$. The picture and cutoffs should agree.

(b) What z value do we use for a 90% confidence level? (Start by changing to .90 and .10 to the right probabilities so we have 90% in the middle. ±1.645

(c) What z value do we use for a 95% confidence level? ±1.96

(d) What z value do we use for an 98% confidence level? ±2.326

(e) Which of these is closest to 2? What approximate confidence goes with an interval (estimate −2SE, estimate +2SE)? *The 95% cutoff of 1.96 is very close to 2, so a* $\pm 2SE$ *interval is an approximate 95% CI.*

Putting together the two pieces with the fact that sample proportions have a nice symmetric distribution (when conditions given on Day 14 are met), we conclude that 95% of the time our sample proportion will be within 2 SE's of the true population proportion.

For the Gallup poll,

$$SE(\widehat{p}) = \sqrt{.62 \times .38/1510} = 0.012 = 1.2\%$$

If we double that, we get 2.4%, so their claim of 3% is conservative (they also use more complex methods). A less conservative 95% CI is $62\% \pm 2.4\% = (59.6, 64.4)\%$.

4. Let's do the same computation for the "Kissing on the Right Side" example in which n = 124 with 80 couples leaning right for a sample proportion of 0.645.

   (a) SE = $\sqrt{.645 * .355/124} = .043$

   (b) $.645 \pm 2SE =$ *(.559, .731)*

   (c) Convert to percent: *(55.9, 73.1)%*

   (d) Compare this to the interval we got by looking for plausible values. Is this method close? *should be close*

(e) What is the margin of error for this 95% confidence interval? *0.086*

5. In the same Gallup poll, 55% of Americans said they approved how Supreme Court Justice Roberts is doing his job.

   (a) Based on the sample size (1510) and the 55% approval rating, what is the standard error of the approval rating? *0.0128*

   (b) What does this SE tell us about the sampling distribution of $\widehat{p}$? Specifically, what percent of sample proportions will fall within 1, 2, or 3 SE's of the true proportion? *68% within 0.0128, 95% within 0.256, and 99.7% within 0.0384.*

   (c) Build a 95% CI for the true approval rating of Justice Roberts. $.55 \pm 2 * 0.0128 = (0.524, 0.576)$

6. Go to the Rossman–Chance applet site and pick "Simulating Confidence Intervals for Population Parameter", http://www.rossmanchance.com/applets/ConfSim.html. Suppose we know that the true percentage of US population who have graduated from college is 27.5%. We want to take a sample of US citizens and build a confidence interval for that parameter (as if we did not really know it).
   This is odd reasoning. We know the true value, but we are still talking about estimating it. Why? Well, it's a check to see how well our method works. Hang on to any doubts, and we'll come back to it later.

   (a) Change the number after π to our true population value, 0.275 and set n to 50. Click Sample. What proportion of your 50 people have a college degree (that's labeled p-hat)? Does the interval contain the true parameter? If so, it's shown in green and the green and black lines intersect, if not, it will be red. *AWV. If $\widehat{p}$ is close to 0.275, we get containment and green color.*
   Note: If we did know the true value is 27.5%, we would not be able to check our method.

   (b) Click sample a few times to see how the sample changes and the interval with it. Now set "Intervals" to 100 and click Sample. The simulation shows 100 confidence intervals.

      i. Click on a red line and write down those little numbers that show in black. (If no segments are red, click sample again until you have some red intervals.) What is the sample proportion, $\widehat{p}$, for this sample? Where does this $\widehat{p}$ fall in the plot of 100 $\widehat{p}$'s at lower right? *AWV. I have a p-hat of .46 and interval estimate (0.322, 0.598) in red. The .46 is my largest $\widehat{p}$ in the lower plot.*

      ii. Look at all the other red intervals. What do they have in common in the lower right plot (besides their color)? *Each mean from a red sample is far from the population value of 0.275*

      iii. Write down the proportion of intervals which contain π. Ideally, this would be 95%, but it might not be very close.

   AWV

   (c) Click Sample five more times and write down each proportion of intervals containing π. What is happening to the running total? *AWV. Running total coverage should get closer to 0.275, but it may take a long time.*

(d) Change confidence level to $\boxed{99}$% and watch what happens to your intervals. Do they get longer or shorter? Is your coverage close to 99%? *longer. I'm at 98%*

(e) Go back to 95% confidence, and then to 90% confidence. What happens to interval length as you decrease the confidence level? *It gets shorter.*

(f) Set confidence to 99% and look at the intervals in red. Where do their $\widehat{p}$'s fall? (If you have no red, get another sample). Now change to 95% and watch the same intervals. Do they stay red? Does $\widehat{p}$ move? Describe what happens. *The samples in red for 99% have the most extreme $\widehat{p}$'s. When you decrease the confidence level, the intervals get shorter and pull further from 0.275. The ones that were red stay red, and some other $\widehat{p}$'s turn red because we don't have to be as far from the true value now to miss it. The $\widehat{p}$'s don't move.*

What does it mean to say we have 95% confidence in an interval we just computed? The applet shows us (for the artificial case where true p is known) that over many many repetitions of the process, we capture p with our interval estimation technique 95% of the time.

Our confidence is in the process. When repeated many times, in the long run 95% of computed intervals will contain the true parameter of interest. That also means that 5% of the time our intervals will miss the target. Unfortunately, when we do it once, our interval might be one of the misses. We have to be willing to trade the diasadvantage of certainty for the advantage of being able to estimate.

This is a direct application of the idea of randomness where we started this course. Finish these statements using "predict", "can", and "can't".

With a random process you ________________________ cannot predict what will happen in one instance, but you ________________________ can predict the overall pattern.

When we compute a confidence interval on real data with unknown parameter, we cannot say that our interval contains the parameter for sure, but we can be confident that the process works 95% of the time overall.

There is a temptation to use a probability statement about confidence intervals. Before you compute the interval, probability makes some sense because the interval is centered at a random value. However, after collecting data and nailing down the endpoints with a formula, there is nothing random about it. Either the interval contains the unknown parameter or it doesn't, and we don't know which.

In the Day 20 activity, we worked with continuous measurement data as well as with proportions. With large sample sizes, one can use the ±2 SE method to build an approximate 95% confidence interval for a mean as well. We need the standard error of the statistic:

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

We'll get to more accurate methods soon which work also for small sample sizes, but let's use the same method for means and see how well it does.

7. Body temperatures.
   $n = 50$, $\overline{x} = 98.26$, standard deviation: $s = 0.765$

   (a) $SE(\overline{x}) = 0.765/\sqrt{50} =$ *0.218*

   (b) 95% CI is $98.26 \pm 2SE = 98.26 \pm .216 = (98.04, 98.48)$

   (c) Again, look back at the last activity and compare this interval to the interval based on plausible values. How close are they? *Should be close.*

   (d) What is the margin of error for this 95% confidence interval? *0.216*

8. On Day 16 we looked at birth weights of babies. For the 28 babies whose mothers smoked, mean weight was achievement was 6.3 with standard deviation of an individual, $s = 1.65$ lb.

   (a) What is the standard error of the sample mean? *0.31*

   (b) What percentage of sample means in samples from this population fall within about .62 lbs of the true population mean birth weight? *95%*

   (c) Build a 95% confidence interval for the population mean . $6.3 \pm 0.62 = (5.67, 6.92)$ *lb*

   (d) To what group of people does this confidence interval inference extend (scope of inference)? *The sample was randomly selected, so inference extends to babies whose mothers smoked during pregnancy in North Carolina (the population from which our sample was drawn).*

With sample size over 100, the ±2 SE method is a decent approximate 95% confidence interval. Next we'll look at methods which work for any sample size.

If you have time, go back to the Rossman–Chance simulation and see what happens when you change $n = 50$ to $n = 100$. Do coverage rates change? Does interval width change?

Bootstrap Resampling

To this point we have built intervals with the "plausible value" method and with estimate±2×SE method. We've seen that the first method is cumbersome, but allows us set the confidence level to $1 - \alpha$ for whatever significance level, α, we happen to like. The $2 \times SE$ method is quicker, but only works for proportions or for large sample sizes. There are other methods for building confidence intervals.

In traditional introductory statistics classes we teach people to read z and t-tables to get a multiplier for each different confidence level. Using the z (standard normal table when variance is known) table, the exact multiplier for 95% confidence is 1.96 instead of 2 if we want 95% confidence. We could multiply SE by this smaller value and still have 95% confidence that the interval captures the unknown parameter. A more modern approach is to use the power of computers to do the work for us.

Problem:
We need to know the sampling distribution to know how far away our statistic might be from our parameter. However, the sampling distribution depends on the unknown parameter.

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

Important fact:

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

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

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

Go to this bootstrapping applet: `http://www.math.montana.edu/~jimrc/randomization/BootDemo.html`

1. The counts shown are all the values in the population, which are counts of the numbers of species of songbirds scientists recorded at different sites in a forest.

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

(b) Click 1 Resample and watch what happens. Click slower 2 or 3 times and watch it again. What is this button doing? *It selects 8 values from the sample with replacement, pulls each down to the next line, and leaves a colored spot on each one it grabbed. The resample then gets combined (averaged) to a single value and that is plotted on the dotplot scale.*

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

(d) Click 100 in the "Many Resamples" choices. Explain what is being plotted. Write down the interval estimate. Count (approximately) how many circle centers are outside the red lines.

It takes a 100 resamples, computes the mean of each, and plots them. Red lines are a confidence interval, but it doesn't say how it found it. I'm guessing 3 circles are below and 1.5 above the interval.

Repeat twice more. Write down each confidence interval and guess how many points fall inside each. *AWV*

(e) What does one point represent? If in doubt, click 1 Resample again. *The mean of one resample of 8 values sampled with replacement from our data sample.*

(f) Click 500, 1000, and 5000 in turn. Write down three CIs for each. Compare the CI's. Are some groups off-center compared to others? More variable?

The smaller numbers of resamples give more variability in CI.

(g) Currently this web app does not let you change the confidence level from 95%. Based on what we saw from last class, predict what will happen to lengths when we change confidence going from 95% to 99% confidence intervals get longer
going from 95% to 90% confidence intervals get shorter

2. When we started, we saw the whole population of counts, and I happen to know that its true mean is $\mu = 19.45$

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

(b) Reload the page, collect a new sample. Compute a 95% confidence interval for the mean using 1000 bootstrap iterations. Show the interval and write "green" if it contains 19.45, "red" if it does not. Do 1000 again, write the interval and "red" or "green" *AWV. Mine was green.* Repeat 8 more times to get a total of 10 samples with 2 intervals for each. Does coverage depend more on the sample or on the particular resample?

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

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

Is College Worth It?

1. In 2011 the Pew Center for the Public Interest conducted a survey in which they interviewed a large sample of American adults. Results:

   - 1221 of 2142 American adults surveyed say the higher education system in the United States provides students a poor to fair value for the money spent.
   - 1606 of 2142 say college is too expensive for most Americans to afford.
   - 651 of the 757 college graduates in the survey say that college was a good investment for them personally.

   (a) Compute the percentages for each number above. Does it seem that those who finished a 4-year degree have a different perspective from general American adults?
   57% think college is a poor to fair value, and 75% think it's too expensive. Of those who graduated, 86% think it was worth it. The 86% does seem to be quite a bit greater than the 57% of the general public who think it a poor–fair value.

   (b) We want to build a bootstrap-based 99% confidence interval for the true proportion of American adults who think college is a poor or fair value for the money spent. We can't use the ±2 SE approach because that's not the right confidence level. (What confidence level does the ±2 SE approach assume?)
   95%

   We saw how bootstrapping worked in the applet we used in the last class period, but that was just a demo and didn't let us put in the data. For real data, we'll use StatKey: http://www.lock5stat.com/statkey to build confidence intervals for each true proportion. Go to their site (or load their app in the Google Chrome browser) and click [Confidence Interval for a Single Proportion].

   (c) [Edit Data] and put in the number of "successes" and the number of people sampled.
   1221 of 2142

   (d) Generate [one bootstrap resample]. How many successes were there in this resample? What is the sample proportion, $\widehat{p}$?
   AWV. I got 1215 for a $\hat{p} = 0.567$.

   (e) Again, this page does not show the resample. What does it mean to resample 2142 of these people with replacement? Explain using the numbers in the table on the right labeled Bootstrap Sample.
   It means that we draw 2142 people "at random" with replacement from a group in which 57% are of the fair–poor opinion, and 43% say a college education is more worthwhile.

   (f) Generate several thousand trials. Click [two-sided], and change the confidence level to [.99]. What is your interval?
   AWV. I got (0.543, 0.599) using 3000 resamples.

(g) Are we confident that the true proportion is over 50%? Explain.
Yes, our 99% CI does not contain 0.50, so one–half is not a plausible number for the true proportion who think college is not worth it. The p–value for $H_0: \ p = .5$ must be less than 0.01.

(h) Is .50 a plausible value for p?
No. Not at the 99% confidence level (which is the 0.01 significance level).

(i) The subjects were called using random digit dialing on both land lines and cell phones. To what group does this confidence interval apply?
All US adults with cell phones or land lines.

2. Change to the number of adults who think college is too expensive. (1606 of 2142 surveyed)

   (a) Find a 99% bootstrap confidence interval for the true proportion.
   AWV. I got (0.725, 0.774) with 4000 bootstrap resamples.

   (b) Interpret the interval explaining what it means in the context of this problem.
   We are 99% confident that the true mean proportion of US adults who think college is too expensive is inside the interval (0.725, 0.774).

3. Change the numbers to $\boxed{651}$ of $\boxed{757}$ and compute a 90% CI for the true proportion of

   ________________________. You fill in the blank and interpret the interval.
   US adults with a college degree who think it was worthwile to get their degree.
   We are 90% confident that the true proportion is in the interval (0.84, 0.881)

4. As you well know, books are an expensive part of a college education. In 2009 students at UCLA randomly sampled 73 books required for classes and compared the UCLA bookstore price to a price they would pay Amazon for the same book.

   (a) Are prices categorical or quantitative?
   quantitative

   (b) If a book has a high price on Amazon, is likely to have a high or low price in the bookstore? And similarly, if it's cheap from one vendor, do you expect it to be expensive or cheap from the other?
   high, cheap

   (c) Which better describes these data? (circle the best answer).

   - We have two independent samples, one of Amazon prices, another of UCLA prices.
   - We have one sample of books with each book measured in two ways (one Amazon, the other UCLA).

   (d) The correct way to work with these data (due to your answer just above), is to take the difference in prices for each book and work with just that string of nunmbers. We will find a confidence interval estimate of the true differnce in mean cost by subtracting Amazon price from UCLA price. Here are the first few rows of data, find the differences:

```
  deptAbbr course           ibsn uclaNew amazNew
1   Am Ind   C170 978-0803272620   27.67   27.95
2   Anthro      9 978-0030119194   40.59   31.14
```

(e) Go to the Statkey site http://www.lock5stat.com/statkey and click Confidence Interval for a Mean, Median, Std. Dev. Click Edit Data, and highlight and delete the data contained there. Get the data from D2L or http://www.math.montana.edu/~jimrc/classes/stat216/data/textbookDiffs.csv and paste it into the data window. Click "Data has a header row" since we have PriceDiff at the top, but don't click "First column is identifier". Then click OK. The original data are shown in the top right window. What are the mean and standard deviation? Is the distribution symmetric? Skewed?
$\overline{x} = 12.762$, $s = 14.25$. *Right skewed.*

(f) Click Generate 1 sample. The resample appears in the window just below the data.
Is the largest cost value included in your bootstrap sample?
Is the smallest?
Are either of them in more than once?
AWV. Some resamples will contain the largest value (or smallest) and some will not. I got one that had the largest 3 times.

(g) What does it mean to see that most of the values are positive? Which source of books is cheaper for the majority of books?
Amazon is generally cheaper, so when we subtract UCLA - Amazon prices, the difference is usually > 0.

(h) Generate 1000 resamples several times. Click Two-tail, click the middle blue box which says 0.95 and change it to .99, and record your interval.
AWV

Note: In statkey you can mouse over any point in the big plot, and the smaller plot to the right changes to show you that particular resample.

(i) Prediction: Before going to the next step, predict what will happen if you change the confidence level to .90. The interval will get:
A. Wider
B. Narrower

Be sure your group all agrees before proceeding.

(j) Now go ahead and switch to .90. What is your new 90% CI?
AWV, but it will be narrower than the 99% CI.

Were you right? Explain why it changed in this way.
Requiring greater confidence means something else "has to give", and we have to extend the interval to include more points in the tails. That makes it wider.

(k) Interpret the meaning of the interval.
We are 90% confident that the true mean difference in cost of textbooks at UCLA in 2009 (local bookstore price minus Amazon price) is in the interval given above.

(l) To what group does this confidence interval apply?
They took a random sample, so it covers the true mean amount saved per book by

using Amazon over UCLA bookstore for the population of all books used at UCLA that year.

(m) Ask your instructors what is needed for the writeup.
Typically we want the same 5 points as for testing.

Note: Please do not assume that all university–associated bookstores are similar to UCLA's in comparison to Amazon. Our local MSU bookstore is student owned and run, and they try very hard to find good deals for students.

Comparing Means

We often have two groups, for example treatment and control or men and women and want to compare means based on independent sample from each group. This is not the same as"Matched Pairs" or "One sample measured twice", which we did in the last class.

We will use StatKey: Confidence interval for the difference in means.
http://lock5stat.com/statkey/bootstrap_1_quant_1_cat/bootstrap_1_quant_1_cat.html
to get bootstrap confidence intervals.

1. Choose the data titled Employed ACS (Income by Sex) from the drop down data menu.

   (a) What are the mean incomes for each gender?
   $50.96K for males, $32.16K for females.

   (b) What is the difference in income? Which mean gets subtracted?
   -18.8 K$, so it subtracts men's mean from women's.

   (c) How big is each sample?
   217 men, 214 women.

   (d) When you click Generate 1 Sample, how is Statkey creating the bootstrapped resample?
   It resamples 217 of the men's salaries with replacement, and 214 of the women's salaries with replacement, then computes means of the new data and subtracts men's resampled mean from women's resampled mean.

   What is the largest male salary in your resample? Largest female salary?
   AWV 563K for men, 382K for women. These happen to be the largest value for men and the max for women, but it's quite possible to get a smaller max for either group.

   Can you find values in the sample which are not used in the resample?
   AWV. 318 is in the sample, but not in my resample for men, and 176 is a woman's salary which is not in my resample.

   Can you spot any values which show up more than once in the resample? (Mousing over in the resample plot might help.)
   AWV. It might take a lot of mousing around.

   (e) Generate several 1000 resamples. Examine resamples in each tail of the distribution. How do they differ?
   On the left we see resamples where the men's mean is relatively low and women's mean is relatively high. It's the opposite in the right tail.

   (f) Compute confidence intervals for these confidence levels:
   AWV. I got:

Confidence	Interval
80%	(-24.98 , -12.56)
90%	(-26.93, -10.79)
95%	(-28.48, -9.53)
99%	(-31.90, -6.14)

(g) Explain the pattern in the lengths of the intervals above. Which are longer? Explain why they are longer.
The longer intervals show up when we ask for greater confidence levels.

(h) Just from the confidence intervals above, would we reject the null hypothesis: $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$ at any of the commonly used (1, 5, or 10%) significance levels? Explain.
Yes. None of the CI's contain 0, so with significance levels 0.20, 0.10, 0.05, and 0.01, we reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$.

(i) Does this study fit the definition of an experiment, or is it observational?
We cannot assign gender to people, so it must be an observational study.

(j) Name three or more possible lurking variables.
AWV. Certainly education level, Socio–Economic–Status, and type of job, part– versus full– time will affect salary.

(k) What is the scope of inference for this study?
Inference is not causal because no treatments were assigned. We don't know if the people in the study are a random sample, so we must assume they are not. Therefore, we can only infer an association between gender and income for the people in this sample.

2. Last year we asked Stat 216 students how much they spent on books for the semester. We've extracted those numbers for two of the colleges on campus, Nursing and Agriculture. See the link on D2L or get the data from http://www.math.montana.edu/~jimrc/classes/stat216/data/bookCost-AgNurs.csv

(a) Discuss and predict: Who spends more on textbooks, nursing or ag students?
AWV

(b) Put the data into StatKey and go through the above steps to compare mean expenditure on books. Which group has the higher average? By how much?
AWV. In Fall 2013 they were quite close.

(c) Is that average large because of a few unusual people, or is it a more general split between groups?
AWV

(d) Find a 95% CI for the true difference in means. Give the interval and explain how you found it.
AWV

(e) What inference do we obtain from this interval which we cannot directly obtain by doing two separate CI's, one for nursing mean and another for the ag mean?
We can ask if we would reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$ at the 5% significance level.

(f) Write up your results. Include the confidence interval, your interpretation of what that means in this setting, and the scope of inference.
We are 95% confident that the true difference in mean amounts spent on books for ag and nursing students falls in the interval $(\cdots, \cdots)$. This inference is not causal, but associative in nature and applies only to those who filled out the web form.

3. In the 2008 Olympics swimming competitions, some swimmers wore full body wet suits which were designed to increase their speed in the water (they have since been banned from the Olympics). In a randomized trial, 12 Olympic class swimmers and triathletes (a convenience sample) swam 1500 m twice: once with the wetsuit, once with a regular swimming suit. The order of the treatment (suit) was randomized for each swimmer.

   (a) Explain in your own words what you expect to see happen with two speeds of a really fast swimmer compared to two speeds of a slower swimmer if the full bodysuit is effective. In other words, is it possible for there to be a real treatment effect, but it might not be strong enough to make all bodysuit speeds larger than all swimming suit speeds?
   This is a paired comparison. A fast swimmer might be faster than the overall mean both with and without the bodysuit, and a slower swimmer might be at the bottom end for both.

   (b) If we take two random samples from different populations, we can say that the two samples are independent, meaning that there is not common effect linking the two. Are these two measurements independent?
   No. Times of the same swmmer with or without the bodysuit will be more similar (correlated).

   (c) Here are the maximum velocities we wish to analyze:

swimmer	1	2	3	4	5	6	7	8	9	10	11	12
bodysuit	1.57	1.47	1.42	1.35	1.22	1.75	1.64	1.57	1.56	1.53	1.49	1.51
swimming suit	1.49	1.37	1.35	1.27	1.12	1.64	1.59	1.52	1.50	1.45	1.44	1.41
difference	.08	.10	.07	.08	.10	.11	.05	.05	.06	.08	.05	.10

   Do swimmers "on average" do better in the full body suit? Does each swimmer do better in the full bodysuit? Explain.
   No. Velocities from the same swimmer are affected by that swimmer's speed relative to the rest of the swimmers, and are likely to be more similar than times from different swimmers.

   (d) We want to build a 99% confidence interval for the **improvement** in speed due to use of the full body suit. Which numbers in the above table are relevant? In other words, do we need to enter both speeds for each swimmer, or is it enough to enter just the differences? Explain.
   Use just the individual differences because these subtract away the individual's speed and give a cleaner comparison.

   (e) Use the appropriate part of Statkey, either $\boxed{\text{CI for a single mean}}$ or $\boxed{\text{CI for difference in means}}$. Enter the data by hand and compute a bootstrap 99% confidence interval.
   AWV I got (.062, .093) for a 99% CI.

   (f) What is the scope of inference for this confidence interval?
   Because order was randomized within each swimmer, we can make causal inference back to this convenience sample of swimmers.

Other Tools

Modern computers and internet access make it easy to simulate distributions as we've done with StatKey. Before such tools were available we had to rely on tables of values of standard distributions. Most stat intro courses still rely on z and t tables to do confidence intervals and hypothesis tests. This activity will help you see how the simulation tools we've used relate to the older methods.

You've used Statkey, and that's a useful skill because it's freely available and you can actually use it to analyze real data in the future. However, if you want to learn more powerful techniques – regression for example - you will need to take more statistics and learn other software better suited to higher powered methods of analysis. We are happy to visit with you about possibilities for more statistics (Stat 217 is a great second course). Another motivation for this lesson is to make it easier to continue your statistical adventures.

The final motivation is that in your field of study, you will need to read research articles which discuss statistical results. Some of the most common are z and t confidence intervals and hypothesis tests.

Proportions

On day 14 we used the normal distribution to compute p-values for a test of a single proportion. We computed

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

then found the tail probabilities from a Normal distribution calculator web applet.

On day 21 we saw that we could input the desired confidence level in the same app and read off the correct z multiplier to build a confidence interval

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

using these z values:

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

1. On day 23 we computed a bootstrap 99% CI for the proportion of adults who think college is a "poor to fair" value. Of 2142 the American adults surveyed, 1221 agreed with that statement.

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

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

(c) Compute a 99% CI using the multiplier you found above.
$0.57 \pm 2.575 \times 0.0107 = 0.57 \pm = (0.542, 0.598)$

(d) Go back to that activity and compare to your bootstrapped CI to this one. Which interval is longer? Are the centers similar?
AWV. I got $(0.543, 0.599)$ *which has very similar center and width.*

As we noted on Day 14, using the normal distribution requires:

- A population of size at least $10n$ (So does the simulation approach).
- Representative sample (So does the simulation approach).
- Independent trials (So does the simulation approach).
- Large enough sample size to expect at least 10 successes (np_0) and at least 10 failures ($n(1 - p_0)$). (Simulation does not require this.)

Means

To build confidence intervals for means and test hypotheses about means, we have an additional problem. We want to learn about the mean, but we also don't know the standard deviation. As you would guess, we can use the sample standard deviation to substitute in for the unknown true standard deviation, but this adds more variability.

2. Take a guess: will our standard deviation estimate be closer to the true standard deviation if we have a sample of size 20 or a sample of size 100?
No wrong answer – it's an opinion. In fact, larger samples will give better estimates.

(a) In StatKey: Theoretical Distributions, click on **t** input 19 for df. What multiplier does it use for a 95% CI?
± 2.093

(b) Edit parameters and change df to 99. Now what multiplier do we use for 95% CI?
± 1.984

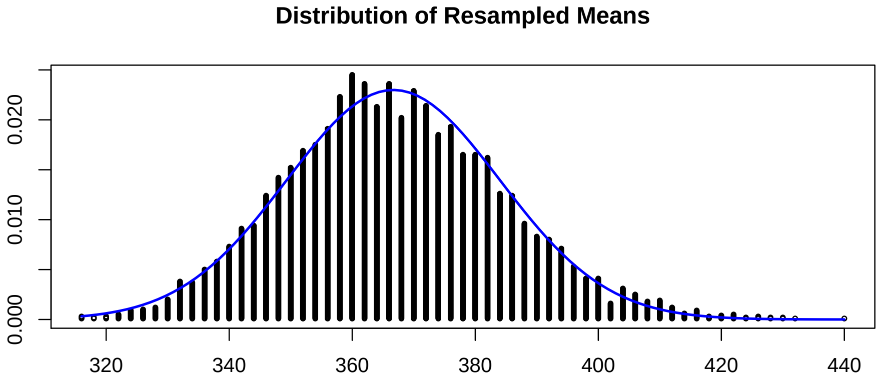
(c) Compare the above values with the standard normal multipliers you found above.
The 99 df t is quite close to the normal 1.96. For 19 df, it's not very close.

The general form of a confidence interval for any parameter is

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

We first used 2 as a generic multiplier to give an approximate 95% CI. When estimating one mean, we need to use a t-distribution with $n - 1$ degrees of freedom. Larger sample size (n) means we have better information about the standard deviation, and hence the distribution looks more like a normal distribution.

3. The cost of books we collected from Stat 216 students had
   $\overline{x} = 366.7$ $s = 243.5$, $n = 197$
   If we bootstrap the mean cost of books 5000 times, we get this picture, where the curve shows a t distribution with 196 degrees of freedom:

   Compute a 95% CI for true mean cost of textbooks.

   (a) Find the standard error of the mean.

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

   $243.5/\sqrt{197} = 17.349$

   (b) Find the t multiplier using n-1 = ____________ df and the StatKey theoretical t page.
   Using 196 df, the t–multiplier is 1.972.

   (c) Create the CI:
   The 95% CI for mean book cost is: $366.7 \pm 1.972 \times 17.349 = 366.7 \pm 34.212 = (332.49, 400.91)\$$

   (d) To what group does this inference apply?
   It only applies to the students who filled out the form.

   Note:
   We can use t distributions for the mean

   - for even small sample sizes (less than 15) if the data are normally distributed.
   - for moderate sample sizes (up to about 30) if the data show only mild skewness.
   - For large sample sizes (30) without worrying about the shape of the distribution.

4. On Day 20 we built an interval of “plausible values” for mean body temperature based on 50 measurements. Suppose the experiment is repeated by randomly selecting records of 50 MSU students who visited the Student Health Center yesterday. We get a sample

mean of 98.44° F with standard deviation 1.06. We want to test the null hypothesis that true mean temperature is 98.6° F versus a two-sided alternative.

(a) Write out the null and alternative hypotheses.
$H_0: \ \mu = 98.6, \quad H_a: \ \mu \neq 98.6$

(b) Find the standard error of the mean.
$SE(\bar{x}) = 1.06/\sqrt{50} = 0.1499$

(c) Compute the test statistic:

$$t = \frac{\overline{x} - \mu_0}{SE(\overline{x})} =$$

$(98.44 - 98.6)/0.1499 = .14/.1499 = -1.067$

(d) What degrees of freedom do we use?
49

(e) In this case, we find the p-value by clicking the tail corresponding to the sign of the test statistic (positive z or t means right-tail, negative z or t means left-tail) and set the x-axis value to your test statistic. But because you are interested in both tails in this alternative hypothesis and because the Normal and t distributions are symmetric, we can obtain our p-value by doubling the area in one tail. Look up the p-value in Statkey.
$0.146 \times 2 = 0.292$

(f) Interpret the strength of evidence and give your decision at the 10% significance level.
Our p–value is large: 0.292, so there is no evidence to support the claim that mean body temperature is different from $98.6°F$ *at the 10% significance level.*

Are we better people in a pleasant environment?

A study from the *Journal of Social Psychology*[9] reports on a study of people's behavior under two different conditions. One treatment was the pleasing smell of fresh bread outside a bakery, while the other was the neutral smell outside a clothing store. Researchers dropped a glove or a package of tissues and recorded whether or not passers–by helped retrieve the lost article. Outside the bakery, 154 of 200 people told the person they lost something, whereas outside the clothing store, 105 of 200 people took the time.

Our goal is to estimate the difference in true proportion of helpers associated with the setting. Use StatKey: Confidence interval for the difference in proportions.
http://lock5stat.com/statkey/bootstrap_2_cat/bootstrap_2_cat.html

1. Enter the data Edit Data
   Write the two proportions (and write labels for them), then the difference in proportions.
   bakery: $\hat{p}_1 = 0.77$
   non-bakery: $\hat{p}_2 = 0.525$
   difference: $\hat{p}_1 - \hat{p}_2 = 0.245$

2. When you click Generate 1 Sample, what difference in proportions do you get?
   I got 0.275

3. Explain how a "Group 1" bootstrap sample is generated. (The same procedure is used for the Group 2 sample, then the difference in proportions is computed.)
   From the bakery sample (which is 77% "yes") select 200 "subjects" at random with replacement. Of those, I got 158 "yes"'s Similarly, sample 200 independent subjects with replacement from the non–bakery which has 52.5% "yes"'s I got 103 positives.

4. Compute confidence intervals for these confidence levels:

   80% (.185, .305) 2nd: same
   90% (.17, .32) 2nd time: (.165,.322)
   95% (.155, .330) 2nd time: (.155, .335)

   99% (.135, .355) 2nd time: (.125, .360)

5. Explain the pattern in the lengths of the intervals above. Why is it that the longer ones must be longer?
   When we ask for more confidence, we have to be willing to take a longer interval.

6. Just from the confidence intervals above, would we reject the null hypothesis: $H_0 : p_1 = p_2$ in favor of $H_A : p_1 \neq p_2$ at any of the commonly used significance levels? Explain.
   Yes. At the 1% significance level, we reject: $H_0 : p1 = p2$ *because the 99% CI does not contain 0.*

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

7. We can create similar intervals using the z multiplier approach.
   First, we need the combined proportion of successes (ignoring treatment), the *marginal* proportion $\widehat{p_m} = \frac{\text{total successes}}{\text{total trials}}$. Secondly, we need the standard error of the statistic: $SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\widehat{p_m}(1-\widehat{p_m})(\frac{1}{n_1}+\frac{1}{n_2})}$.

   (a) Compute $\widehat{p}_m$, the overall, or marginal proportion to improve (ignoring treatment).
   259/400 = 0.6475

   (b) From that, compute $SE(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\widehat{p_m}(1-\widehat{p_m})(\frac{1}{n_1}+\frac{1}{n_2})} = \sqrt{.6475 * .3525/100} = 0.0478$

   (c) Construct a 99% confidence interval for the difference in proportions and compare it to those from above.
   (.122, .368) this is slightly wider than the bootstrapped interval.

8. The researchers gave this description of their methods:

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

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

   (a) Was a random mechanism used to select the person studied? Explain.
   Yes, the coin flip would pick about half the subjects at random.

   (b) What was the "treatment" and how was it applied to a subject?
   Bakery (with smell) or clothing store (no smell). It was not assigned to a subject.

[10] ibid

(c) Does this study fit the definition of an experiment, or is it observational? Explain.
Observational. The people studied happened to be in front of a bakery or a non–bakery. They were not assigned at random to the stores.

(d) Name three or more possible lurking variables.
socio-economic status, race, How hurried the person felt.

(e) What is the scope on inference for this study?
We can infer association to the sample of people observed.

9. The dataset `babyWeights.txt` contains data about a random sample of 999 births in North Carolina in 2003. We want to compare mean birth weight of babies whose mother smoked during pregnancy with mean birth weight of babies whose mothers did not smoke. The data are available on D2L or at `http://www.math.montana.edu/~jimrc/classes/stat216/data/babyWeight.txt`

   (a) Put the data into StatKey and find the means and the sample sizes for each group of babies.
   nonsmokers: $n = 873,\ \overline{x}_2 = 7.14lb$
   smokers: $n = 126,\ \overline{x}_2 = 6.83lb$

   (b) Use StatKey to compute a 95% CI for the difference in means.
   I get $(0.060, 0.585)$ *lb based on 5000 bootstrap resamples..*

   (c) Use $s = 1.50$ as the estimated standard deviation to build the standard error for the difference in means:
   $$SE(\overline{x}_1 - \overline{x}_2) = s\sqrt{\frac{1}{126} + \frac{1}{873}} =$$
   0.143

   (d) Find the t–multiplier with 997 degrees of freedom to sue for a 95% CI.
   1.962

   (e) Build the 95% CI for $\mu_1 - \mu_2$, where the first group is non–smoking moms and the second group is smoking moms.
   $0.3155 \pm 1,962 \times 0.143 = (0.035, 0.597)$ *lb.*

   (f) Is zero in either interval? What does that indicate? (Refer to plausible values and the associated hypothesis test.)
   At the 5% significance level, we reject $H_0 : \ \mu_1 = \mu_2$ *and conclude that mean birth weight is lower for mothers who smoke than for mothers who don't smoke.*

   (g) What is the scope of inference for your CI's?
   This is a random sample of birth weights for the state of North Carolina in 2003, so we can make inference back to all babies born in the state that year. It is not causal inference because we cannot assign a smoking habit to some moms. One could get a better estimate of the smoking association if we accounted for other important factors, like length of pregnancy (separating out premature births).

Unit 3 Wrapup
Vocabulary

Define each term:

- Precision
- Parameter
- Statistic
- Standard Deviation
- Standard Error (of a sample proportion and a sample mean)
- Sampling Variation/Sampling Error
- Margin of Error
- Sample Estimate
- Plausible Value
- Paired sample versus Independent samples (means)
- Interval Estimate
- Bootstrap Resample
- Z-test (for single and two-sample proportion)
- T-test

- Interpretation of a confidence interval

Practice Problems

1. Rating Chain Restaurants[11]

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

   (a) Do you think that the sample here was chosen randomly from the population of Consumer Report readers? Explain.

   (b) Why do the authors of the article make this disclaimer about not necessarily representing the general population?

   (c) To what group of people would you feel comfortable generalizing the results of this study? Explain.

2. Emotional Support[12]

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

   (a) Comment on whether Hite's sampling method is likely to be biased in a particular direction. Specifically, do you think that the 96% figure overestimates or underestimates

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

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

the proportion who give more support in the population of all American women? What type of bias could be occurring here?

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

3. Balsa Wood

Student researchers investigated whether balsa wood is less elastic after it has been immersed in water. They took 44 pieces of balsa wood and randomly assigned half to be immersed in water and the other half not to be immersed in water. They measured the elasticity by seeing how far (in inches) the piece of wood would project a dime into the air. Use the data file located on D2L.

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

(b) The observed difference in mean elasticity between the two groups is 4.16 inches. Explain why its more appropriate to produce a bootstrap interval, as opposed to simply reporting this value, for estimating the actual treatment effect of immersing balsa wood in water.

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

4. MicroSort® Study

   The Genetics and IVF Institute is currently studying methods to change the odds of having a girl or boy2. MicroSort® is a method used to sort sperm with X- and Y-chromosomes. The method is currently going through clinical trials. Women who plan to get pregnant and prefer to have a girl can go through a process called X-Sort®. As of 2008, 945 have participated and 879 have given birth to girls[13]

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

   (b) Interpret your interval.

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

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

      ii. What is the multiplier that should be used?

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

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

      v. Give the interval estimate.

      vi. How is this method similar to the 2SE method? How does it differ?

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

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

(d) Suppose more data has been collected since 2008. If the number of women had increased to 3000 but the observed percent of girls stayed the same, what would you expect to happen to your interval?

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

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

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

5. Anorexia nervosa is an eating disorder characterized by immoderate food restriction and irrational fear of gaining weight, as well as a distorted body self-perception. Several psychotherapy methods have been tested as a way to treat individuals suffering from anorexia. The data set available on D2L gives the results of a study using cognitive behavioral therapy (CBT) and the pre-and post-treatment weights of 29 patients, along with the change in weight.

   (a) Enter the data into Statkey to create a 99% bootstrap confidence interval. Should the single or two-sample mean applet be used? Explain, then give the confidence interval.

   (b) Using the interval you created in 5a, use the plausible values method to create a 99% confidence interval. Round to one decimal place. Below list the values you tried, the p-value of the test, and whether the value is plausible or not. Then give the confidence interval.

   (c) Compare the two intervals above. Are they similar in center and width?

(d) Would a 95% confidence interval be wider or narrower? Use both the bootstrap and plausible values methods to explain your choice.

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

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

6. Across the U.S. there has been increasing awareness of the dangers of concussions among high school, college, and professional athletes. One study investigated whether concussions are more common among male or female soccer players. The study took a random sample of college soccer players from 1997 1999. Of 75,082 exposures (practices or games) for female soccer players, 158 resulted in a concussion while 75,734 exposures for men resulted in 101 concussions. Does this show a gender difference in concussions rates among collegiate soccer players?

   (a) Write the null and alternative hypotheses.

   (b) Use the theoretical distribution to conduct this test.

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

      ii. Calculate the z test statistic.

      iii. Find the p-value.

      iv. Write-up your results using all five components required. Use a 10% significance level.

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

i. What confidence level should be used?

ii. Give and interpret the interval.

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

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