Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

0. Midterms and HW4.
1. Comparing two means and bicycling to work example.
2. Paired data and studying with music example.
3. Simulation approach with paired data and baseball example.
4. Theory based approach for paired data and M&M example.

Read ch7 and 10.
Hw4 is due Tue Sep12, 10am, again by email to statgrader or statgrader2, and is prob.s 10.1.8, 10.3.14, 10.3.21, and 10.4.11.
http://www.stat.ucla.edu/~frederic/13/sum23 .

0. Midterms and hw.

Hw4 is 10.1.8, 10.3.14, 10.3.21, and 10.4.11.
10.1.8 starts "Which of the following statements is correct? A. Changing the units of measurements of the explanatory or response variable".
10.3.14 starts "Consider the following two scatterplots based on data gathered in a study of 30 crickets".
10.3.21 starts "The book *Day Hikes in San Luis Obisbo County*".
10.4.11 starts "In a study to see if there was an association between weight loss and the amount of a certain protein in a person's body fat".

On the midterm,
the mean $\bar{x}$ was 87%. s = 16%.
The grading is the standard scale, i.e. 90-100 = A range, 80-89.9 = B range, etc.

10.1.8 Which of the following statements is correct?

A. Changing the units of measurements of the explanatory or response variable does not change the value of the correlation.

B. A negative value for the correlation indicates that there is no relationship between the two variables.

C. The correlation has the same units (e.g., feet or minutes) as the explanatory variable.

D. Correlation between y and x has the same number but opposite sign as the correlation between x and y.

10.1.9* If two variables are negatively associated, then we know that:

A. Above-average values in one variable correspond to below-average values in the other variable.

B. Above-average values in one variable correspond to above-average values in the other variable.

C. Below-average values in one variable correspond to below-average values in the other variable.

D. Below-average values in one variable correspond either above-average or below-average values in the other variable.

10.1.10 For each of the following statements, say what, if anything, is wrong.

a. Because the correlation coefficient between

10.3.12 Reconsider the previous five exercises and the **Legos** data file. The last product listed in the data file has 415 pieces and a price of $49.99.

a. Determine the predicted price for such a product.

b. Determine the residual value for this product.

c. Interpret what this residual value means.

d. Does the product fall above or below the least squares line in the graph? Explain how you can tell, based on its residual value.

10.3.13 Reconsider the previous six exercises and the **Legos** data file. This is very unrealistic, but suppose that one of the products were to be offered at a price of $0.

a. Would you expect this change to affect the least squares line very much? Explain.

b. For which one product would you expect this change to have the greatest impact on the least squares line? Explain how you choose this product.

c. Change the price to $0 for the product that you identified in part (b). Report the (new) equation of the least squares line and the (new) value of r^2. Have these values changed considerably?

Crickets

10.3.14 Consider the following two scatterplots based on data gathered in a study of 30 crickets, with temperature measured in degrees Fahrenheit and chirp frequency measured in chirps per minute.

a. If the goal is to predict temperature based on a cricket's chirps per minute, which is the appropriate scatterplot to examine—the one on the left or the one on the right? Explain briefly.

One of the following is the correct equation of the least squares line for predicting temperature from chirps per minute:

A. predicted temperature = 35.78 + 0.25 chirps per minute

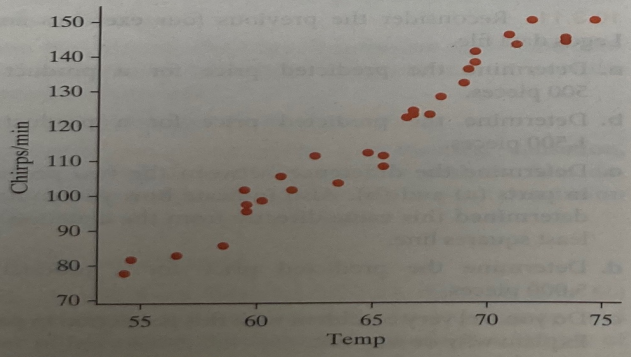
B. predicted temperature = −131.23 + 3.81 chirps per minute

C. predicted temperature = 83.54 − 0.25 chirps per minute

b. Which is the correct equation? Circle your answer and explain briefly.

c. Use the correct equation to predict the temperature when the cricket is chirping at 100 chirps per minute.

d. Interpret the value of the slope coefficient, in this context, for whichever equation you think is the correct one.



EXERCISE 10.3.14

Cat jumping*

10.3.15 Harris and Steudel (2002) studied factors that might be associated with the jumping performance of domestic cats. They studied 18 cats, using takeoff velocity (in centimeters per second) as the response variable. They used body mass (in grams), hind limb length (in centimeters), muscle mass (in grams), and percent body fat in addition to sex as potential explanatory variables. The data can be found in the **CatJumping** data file. A scatterplot of takeoff velocity vs. body mass is shown in the figure for Exercise 10.3.15.

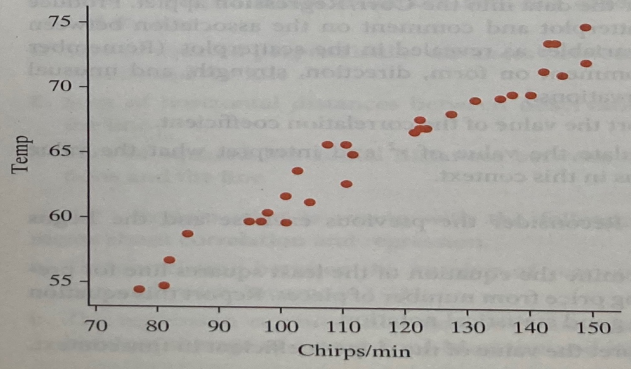
a. Describe the association between these variables.

b. Use the **Corr/Regression** applet to determine the equation of the least squares line for predicting a cat's takeoff velocity from its mass.

c. Interpret the value of the slope coefficient in this context.

d. Interpret the value of the intercept coefficient. Is this a context in which the intercept coefficient is meaningful?

e. Determine the proportion of variability in takeoff velocity that is explained by the least squares line with mass.



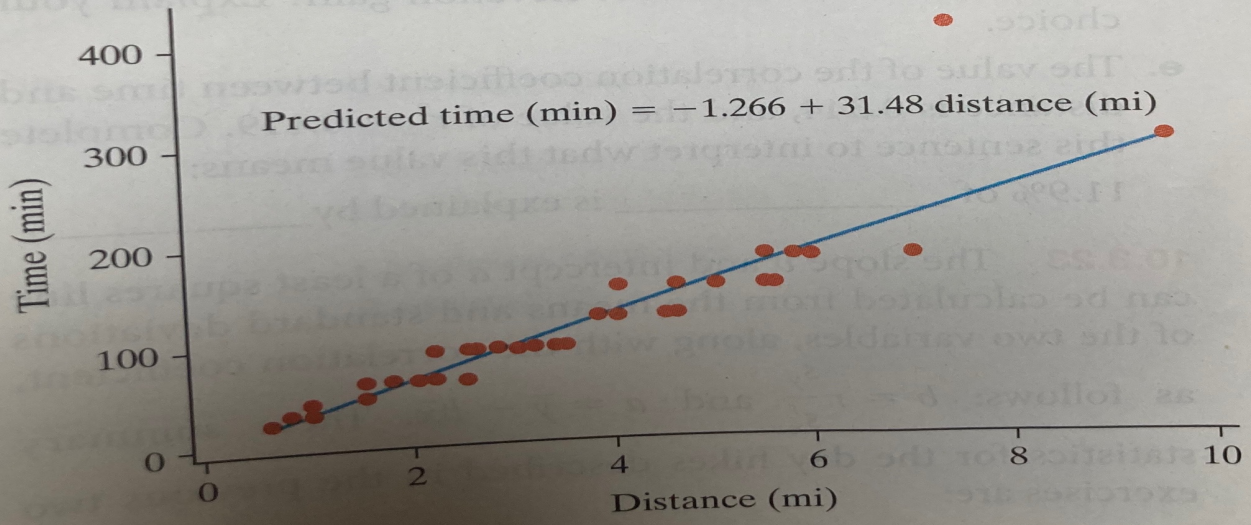
500-page textbook. Then do the same for a 1,500-page textbook. Which prediction would you have more confidence in? Explain.

d. Interpret what the slope coefficient means in this context.

e. Determine the proportion of variability in textbook prices that is explained by knowing the number of pages in the book.

Day hikes

10.3.21 The book *Day Hikes in San Luis Obispo County* lists information about 72 hikes, including the distance of the hike (in miles), the elevation gain of the hike (in feet), and the time that the hike is expected to take (in minutes). Consider the scatterplot below, with least squares regression line superimposed:



a. Report the value of the slope coefficient for predicting time from distance.

b. Write a sentence interpreting the value of the slope coefficient for predicting time from distance.

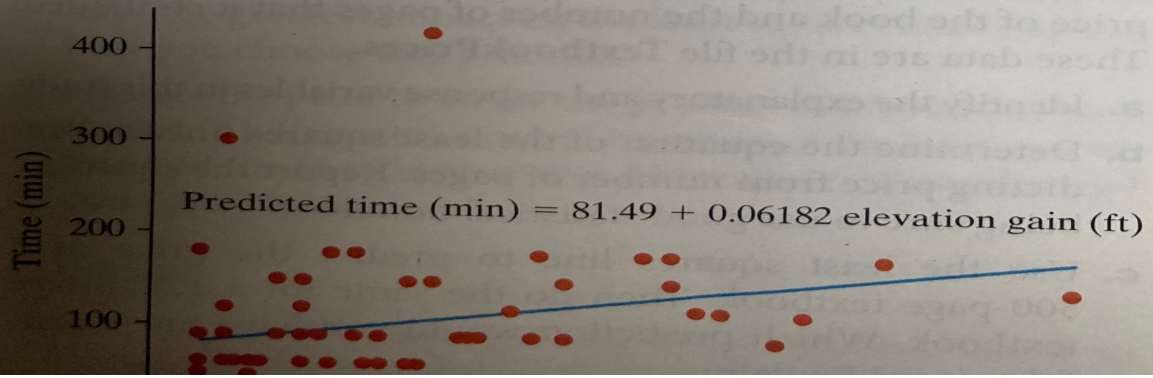
c. Use the line to predict how long a 4-mile hike will take.

d. Would you feel more comfortable using the line predict the time for a 4-mile hike or for a 12-mile hike? Explain your choice.

e. The value of the correlation coefficient between time and distance is 0.916, and the value of $r^2 = 0.839$. Complete this sentence to interpret what this value means:

83.9% of _____ is explained by _____.

10.3.22 Reconsider the previous exercise. The following scatterplot displays hiking time vs. elevation gain, with the least squares line superimposed:



Distan
Elevat
Time

a. Use *a* fo

b. Use *a* fo

c. Use hike

d. Use a hik tion

e. Wha and

Introdu

10.3.24 Test 2 a sample

a. In th

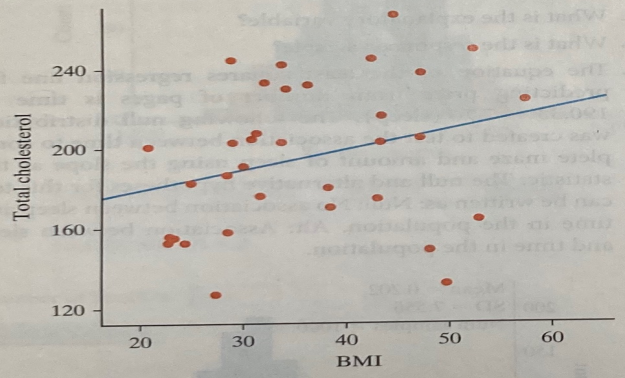
10.4.10 Reconsider the previous exercise about the amount of sleep (in hours) obtained in the previous night and time to complete a paper and pencil maze (in seconds). The equation of the least squares regression line for predicting price from number of pages is $\widehat{time} = 190.33 - 7.76\,(sleep)$.

a. Interpret what the slope coefficient means in the context of sleep and time to complete the maze.

b. Interpret the intercept. Is this an example of extrapolation? Why or why not?

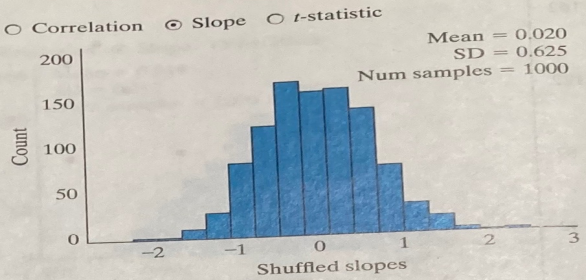
Weight loss and protein

10.4.11 In a study to see if there was an association between weight loss and the amount of a certain protein in a person's body fat, the researchers measured a number of different attributes in their 39 subjects at the beginning of the study. The article reported, "These subjects were clinically and ethnically heterogeneous." Two of the variables they measured were body mass index (BMI) and total cholesterol. The results are shown in the scatterplot along with the regression line.



a. What are the observational units in the study?

b. The equation of the least squares regression line for predicting total cholesterol from BMI is $\widehat{cholesterol} = 162.56 - 0.9658\,(BMI)$. The following null distribution was created to test the association between people's total cholesterol number and their BMI using the slope as the statistic. The null and alternative hypotheses for this test can be written as: Null: No association between cholesterol and BMI in the population. Alt: Association between cholesterol and BMI in the population.



i. Based on information shown in the null distribution, how many standard deviations is our observed statistic below the mean of the null distribution? (That is, what is the standardized statistic?)

ii. Based on your standardized statistic, do you have strong evidence of an association between a people's total cholesterol and their BMI? Explain.

10.4.12 Reconsider the previous exercise about the cholesterol and BMI. The equation of the least squares regression line obtained was $\widehat{cholesterol} = 162.56 - 0.9658\,(BMI)$.

a. Interpret what the slope coefficient means in the context of cholesterol and BMI.

b. Interpret the intercept. Is this an example of extrapolation? Why or why not?

Honda Civic prices*

10.4.13 The data in the file **UsedHondaCivics** come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car's age (calculated as 2006 minus year of manufacture) and price. Consider conducting a simulation analysis to test whether the sample data provide strong evidence of an association between a car's price and age in the population in terms of the population slope.

a. State the appropriate null and alternative hypotheses.

b. Conduct a simulation analysis with 1,000 repetitions. Describe how to find your p-value from your simulation results and report this p-value.

c. Summarize your conclusion from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

10.4.14 Reconsider the previous exercise on prices of Honda Civics.

a. Find the regression equation that predicts the price of the car given its age.

b. Interpret the slope and intercept of the regression line.

1. Comparing Two Means: Simulation-Based Approach and bicycling to work example.

Section 6.2

Similar to proportions.

- We will be comparing means, much the same way we compared two proportions using randomization techniques.
- The difference here is that the response variable is quantitative (the explanatory variable is still binary though). So if cards are used to develop a null distribution, numbers go on the cards instead of words.

Bicycling to Work

Example 6.2

Bicycling to Work

- Does bicycle weight affect commute time?
- British Medical Journal (2010) presented the results of a randomized experiment done by Jeremy Groves, who wanted to know if bicycle weight affected his commute to work.
- For 56 days (January to July) Groves tossed a coin to decide if he would bike the 27 miles to work on his carbon frame bike (20.9lbs) or steel frame bicycle (29.75lbs).
- He recorded the commute time for each trip.

Bicycling to Work

- What are the observational units?
  - Each trip to work on the 56 different days.
- What are the explanatory and response variables?
  - Explanatory is which bike Groves rode (categorical – binary)
  - Response variable is his commute time (quantitative)

Bicycling to Work

- **Null hypothesis:** Commute time is not affected by which bike is used.
- **Alternative hypothesis:** Commute time is affected by which bike is used.

Bicycling to Work

- In chapter 5 we used the difference in proportions of "successes" between the two groups.
- Now we will compare the difference in averages between the two groups.
- The parameters of interest are:
  - μ_{carbon} = Long term average commute time with carbon framed bike
  - μ_{steel} = Long term average commute time with steel framed bike.

Bicycling to Work

- μ is the population mean. It is a parameter.
- Using the symbols μ_{carbon} and μ_{steel}, we can restate the hypotheses.

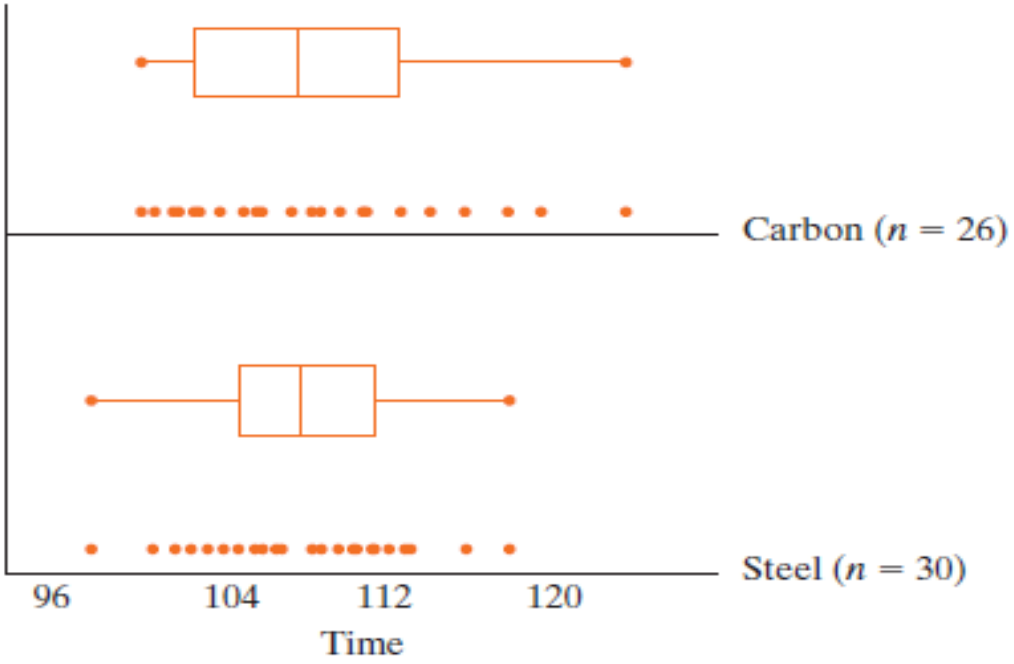
- **H_0:** $\mu_{carbon} = \mu_{steel}$
- **H_a:** $\mu_{carbon} \neq \mu_{steel}$.

Bicycling to Work

Remember:

- The hypotheses are about the longterm association between commute time and bike used, not just his 56 trips.
- Hypotheses are always about populations or processes, not the sample data.

Bicycling to Work



	Sample size	Sample mean	Sample SD
Carbon frame	26	108.34 min	6.25 min
Steel frame	30	107.81 min	4.89 min

Bicycling to Work

- The sample mean was higher for the carbon framed bike.
- Does this indicate the bike is better?
- Or could a higher average just come from the random assignment? Perhaps the carbon frame bike was randomly assigned to days where traffic was heavier or weather slowed down Dr. Groves on his way to work?

Bicycling to Work

- **Statistic:**
- The observed difference in average commute times

$$\bar{x}_{\text{carbon}} - \bar{x}_{\text{steel}} = 108.34 - 107.81$$

$$= 0.53 \text{ minutes}$$

Bicycling to Work

Simulation:

- We can imagine simulating this study with index cards.
  - Write all 56 times on 56 cards.
- Shuffle all 56 cards and randomly redistribute into two stacks:
  - One with 26 cards (representing the times for the carbon-frame bike)
  - Another 30 cards (representing the times for the steel-frame bike)

Bicycling to Work

Simulation (continued):

- Shuffling assumes the null hypothesis of no association between commute time and bike
- After shuffling we calculate the difference in the average times between the two stacks of cards.
- Repeat this many times to develop a null distribution

Carbon Frame

116	114	119	123	113
111	113	106	118	109
103	103	104	112	110
101	102	100	102	107
105	103	111	106	102
108				

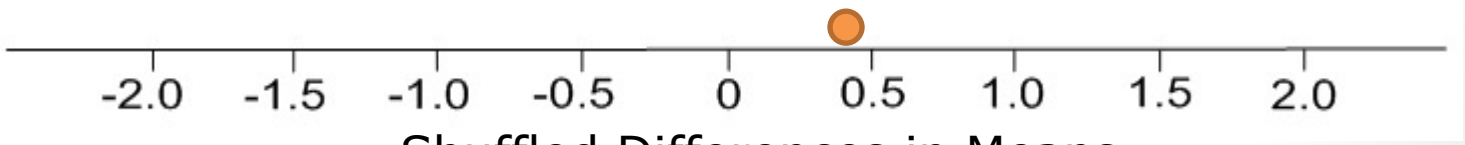
mean = 108.27

Steel Frame

116	116	109	118	113
110	113	104	113	105
111	111	110	105	106
103	102	98	109	108
102	112	101	106	102
105	105	106	107	106

mean = 107.87

108.27 – 107.87 = 0.40



Shuffled Differences in Means

Carbon Frame

116	114	119	123	113
111	113	106	118	109
103	103	104	112	110
101	102	100	102	107
105	103	111	106	102
108				

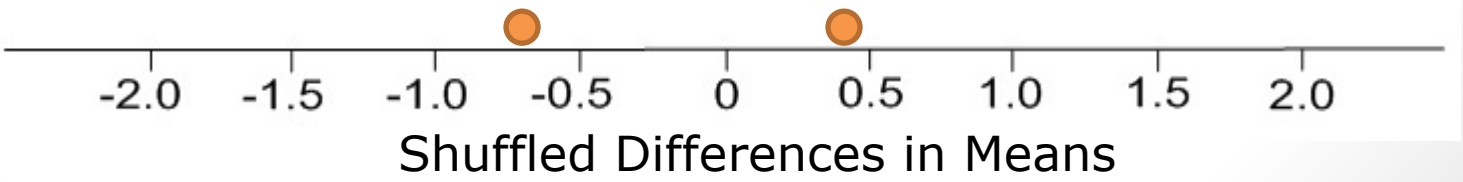
mean = 107.69

Steel Frame

116	116	109	118	113
110	113	104	113	105
111	111	110	105	106
103	102	98	109	108
102	112	101	106	102
105	105	106	107	106

mean = 108.37

107.69 – 108.37 = -0.68



Carbon Frame

116	114	119	123	113
111	113	106	118	109
103	103	104	112	110
101	102	100	102	107
105	103	111	106	102
108				

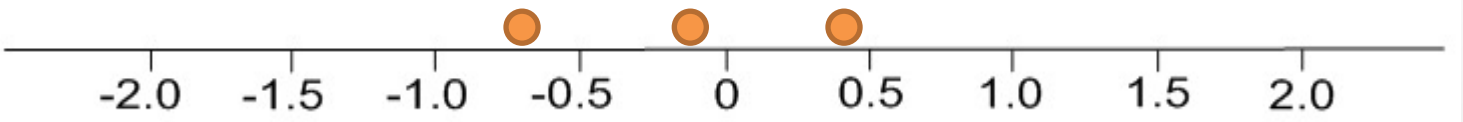
mean = 107.97

Steel Frame

116	116	109	118	113
110	113	104	113	105
111	111	110	105	106
103	102	98	109	108
102	112	101	106	102
105	105	106	107	106

mean = 108.13

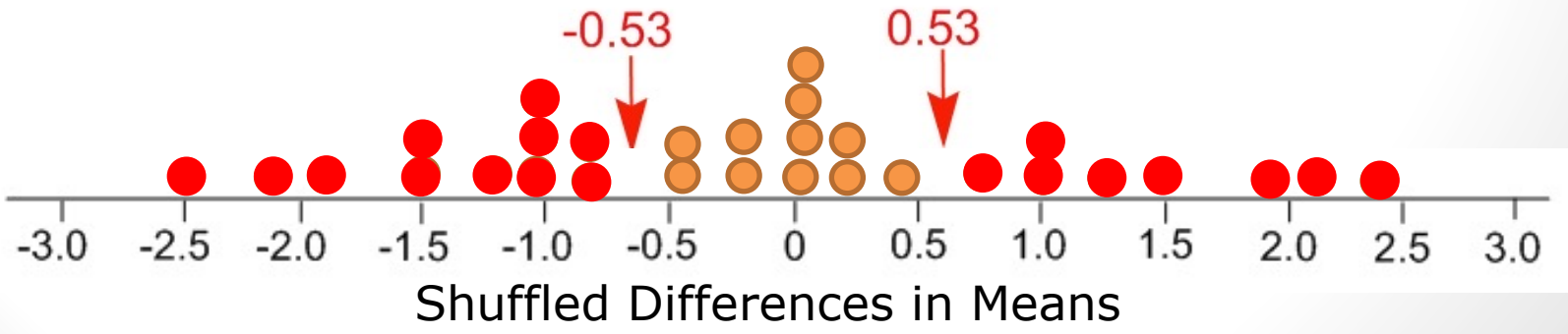
107.97 – 108.13 = -0.16



Shuffled Differences in Means

More Simulations

Nineteen of our 30 simulated statistics were as or more extreme than our observed difference in means of 0.53, hence our estimated p-value for this null distribution is 19/30 = 0.63.



Bicycling to Work

- Using 1000 simulations, we obtain a p-value of 72%.
- What does this p-value mean?
- If mean commute times for the bikes are the same in the long run, and we repeated random assignment of the carbon bike to 26 days and the steel bike to 30 days, a mean difference as extreme as 0.53 minutes or more would occur in about 72% of the simulations.
- Therefore, we do not have strong evidence that the commute times for the two bikes will differ in the long run. The difference between bikes observed by Dr. Groves is not statistically significant.

Bicycling to Work

- Have we proven that the bikes are equivalent? (Can we conclude the null is true?)
  - No, a large p-value is not “strong evidence that the null hypothesis is true.”
  - It suggests that the null hypothesis is consistent with the data.
  - There could be no long-term difference. But there also could be a small long-term difference.

Bicycling to Work

- Imagine we want to generate a 95% confidence interval for the long-run difference in average commuting time.
  - Sample difference in means ± 1.96×SE for the difference between the two means
- From simulations, the SE = standard deviation of the simulated differences between sample means = 1.47.
- 0.53 ± 1.96(1.47)= 0.53 ± 2.88
- -2.35 to 3.41.
- What does this mean?

Bicycling to Work

- We are 95% confident that the true longterm difference (carbon – steel) in average commuting times is between -2.41 and 3.47 minutes.
- We are 95% confident the carbon framed bike is between 2.41 minutes faster and 3.47 minutes slower than the steel framed bike.
- Does it make sense that the interval contains 0, based on our p-value?

Bicycling to Work

- Was the sample representative of an overall population?
- What about the population of all days Dr. Groves might bike to work?
  - No, Groves commuted on consecutive days in this study and did not include all seasons.
- Was this an experiment? Were the observational units randomly assigned to treatments?
  - Yes, he flipped a coin for the bike.
  - We can probably draw cause-and-effect conclusions here.

Bicycling to Work

- We cannot generalize beyond Groves and his two bikes.
- A limitation is that this study is not ***double-blind.***
  - The researcher and the subject (which happened to be the same person here) were not blind to which treatment was being used.
  - Dr. Groves knew which bike he was riding, and this might have affected his state of mind or his choices while riding.

2. Paired Data.

Chapter 7

Introduction

- The paired data sets in this chapter have one *pair* of quantitative response values for each obs. unit.
- This allows for a comparison where the other possible confounders are as similar as possible between the two groups.
- Paired data studies remove individual variability by looking at the difference score for each subject.
- Reducing variability in data improves inferences:
  - Narrower confidence intervals.
  - Smaller p-values when the null hypothesis is false.
  - Less influence from confounding factors.

- The main idea is to look at the difference between responses, and then analyze these differences the way we analyzed one variable previously.

Paired data and studying with music example.

Example 7.1

Studying with Music

- Many students study while listening to music.
- Does it hurt their ability to focus?
- In “Checking It Out: Does music interfere with studying?” Stanford Prof Clifford Nass claims the human brain listens to song lyrics with the same part that does word processing.
- Instrumental music is, for the most part, processed on the other side of the brain, and Nass claims that listening to instrumental music has virtually no interference on reading text.

Studying with Music

Consider the experimental designs:

Experiment A — Random assignment to 2 groups

- 27 students were randomly assigned to 1 of 2 groups:
  - One group listens to music with lyrics.
  - One group listens to music without lyrics.
- Students play a memorization game while listening to the particular music that they were assigned.

Studying with Music

Experiment B — Paired design using repeated measures

- All students play the memorization game twice:
  - Once while listening to music with lyrics
  - Once while listening to music without lyrics.

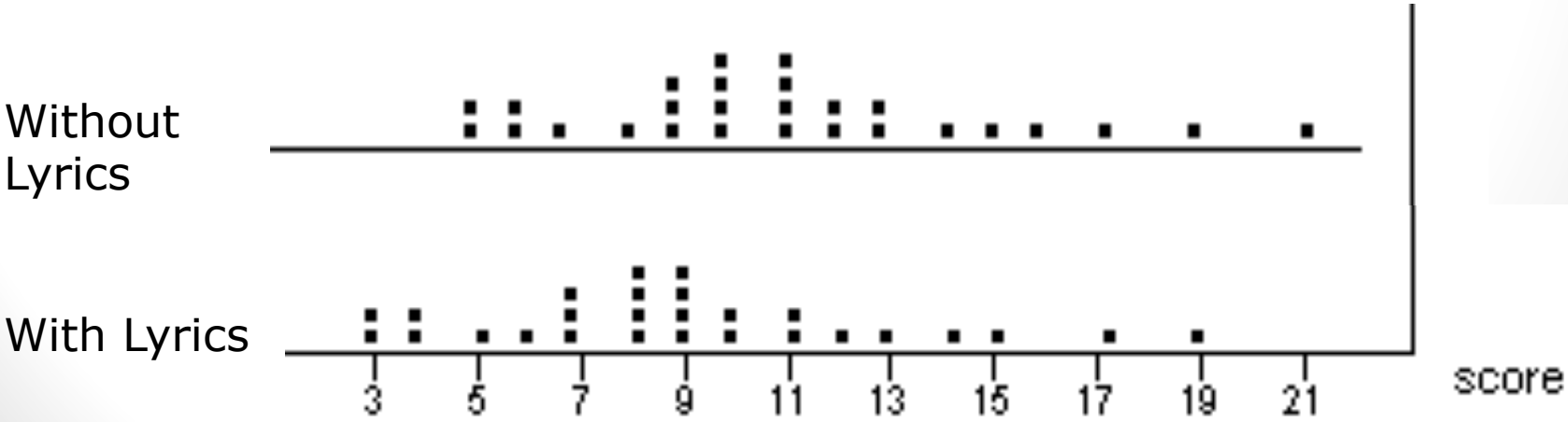
Experiment C — Paired design using matching

- Sometimes repeating something is impossible (like testing a surgical procedure) but we can still pair.
  - Test each student on memorization.
  - Match students up with similar scores and randomly:
    - Have one play the game while listening to music with lyrics and the other while listening to music without lyrics.

Studying with Music

We will focus on the repeated measures type of pairing.

- What if everyone could remember exactly 2 more words when they listened to a song without lyrics?
- Using Experiment A, there could be a lot of overlap between the two sets of scores and it would be difficult to detect a difference, as shown here.



Studying with Music

- Variability in people's memorization abilities may make it difficult to see differences between the songs in Experiment A.
- The paired design focuses on the *difference* in the number of words memorized, instead of the number of words memorized.
- **By looking at this difference, the variability in general memorization ability is taken away.**

Studying with Music

- In Experiment B, there would be no variability at all in our hypothetical example.
- While there is substantial variability in the number of words memorized between students, there would be no variability in the *difference in the number of words memorized.* All values would be exactly 2.
- Hence we would have extremely strong evidence of a difference in ability to memorize words between the two types of music.

Pairing and Random Assignment

Pairing often increases power, and makes it easier to detect statistical significance.

In our memorizing with or without lyrics example:

- If we see significant improvement in performance, is it attributable to the type of song?
- What about experience? Could that have made the difference?
- What is a better design?
  - Randomly assign each person to which song they hear first: with lyrics first, or without.
  - This cancels out an “experience” effect

Paring and Observational Studies

You can often do matched pairs in observational studies, when you know the potential confounder ahead of time.

If you are studying whether the portacaval shunt decreases the risk of heart attack, you could match each patient getting the shunt with a patient of similar health not getting the shunt.

If you are studying whether lefthandedness causes death, and you want to account for age in the population, you could match each leftie with a rightie of the same age, and compare their ages at death.

3. Simulation based Approach for Analyzing Paired Data, and rounding first base example.

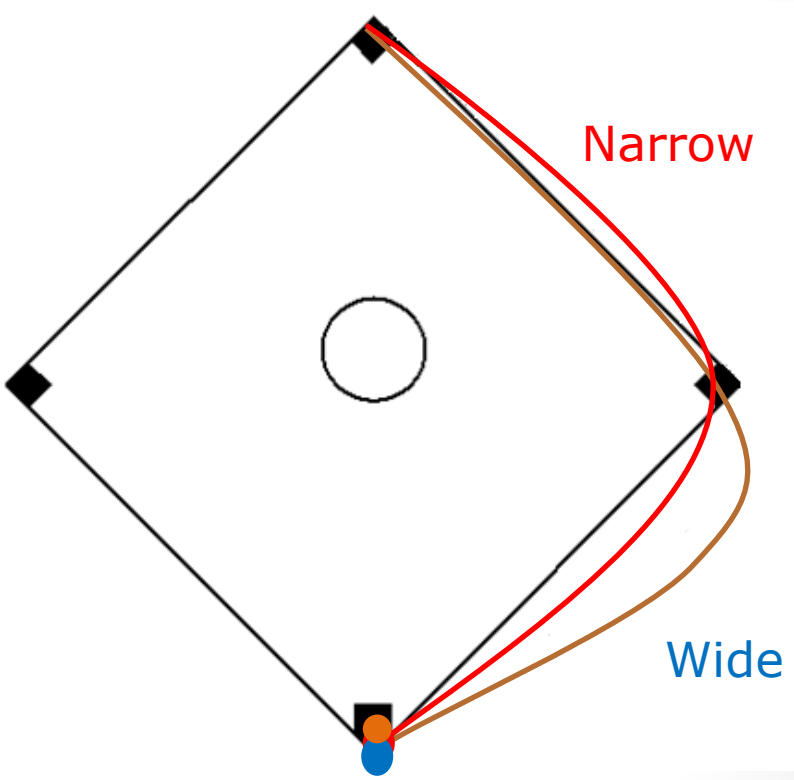
Section 7.2

Rounding First Base

Example 7.2

Rounding First Base

- Imagine you've hit a line drive and are trying to reach second base.
- Does the path that you take to round first base make much of a difference?
  - Narrow angle
  - Wide angle



Rounding First Base

- Woodward (1970) investigated these base running strategies.
- He timed 22 different runners from a spot 35 feet past home to a spot 15 feet before second.
- Each runner used each strategy (paired design), with a rest in between.
- He used random assignment to decide which path each runner should do first.
- **This paired design controls for the runner-to-runner variability.**

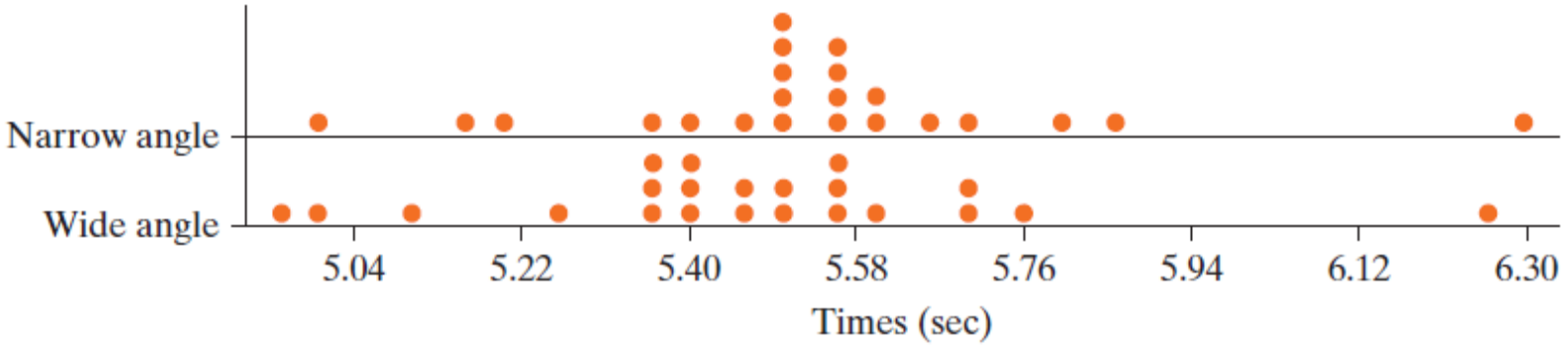
First Base

- What are the observational units in this study?
  - The runners (22 total)
- What variables are recorded? What are their types and roles?
  - Explanatory variable: base running method: wide or narrow angle (categorical)
  - Response variable: time from home plate to second base (quantitative)
- Is this an observational study or an experiment?
  - Randomized experiment.

The results

TABLE 7.1 The running times (seconds) for the first 10 of the 22 subjects

Subject	1	2	3	4	5	6	7	8	9	10	
Narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	...
Wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	...



The Statistics

- There is a lot of overlap in the distributions and substantial variability.

	Mean	SD
Narrow	5.534	0.260
Wide	5.459	0.273

- It is difficult to detect a difference between the methods when these is so much variation.

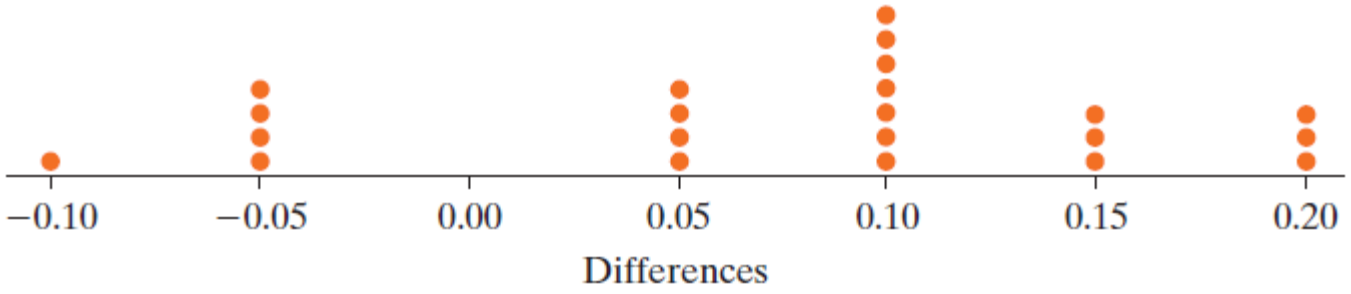
Rounding First Base

- However, these data are clearly paired.
- The paired response variable is time difference in running between the two methods and we can use this in analyzing the data.

The Differences in Times

TABLE 7.2 Last row is difference in times for each of the first 10 runners (narrow – wide)

Subject	1	2	3	4	5	6	7	8	9	10	
Narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
Wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
Difference	−0.05	−0.05	0.10	0.10	0.15	−0.05	0.05	0.15	0.15	0.10	…

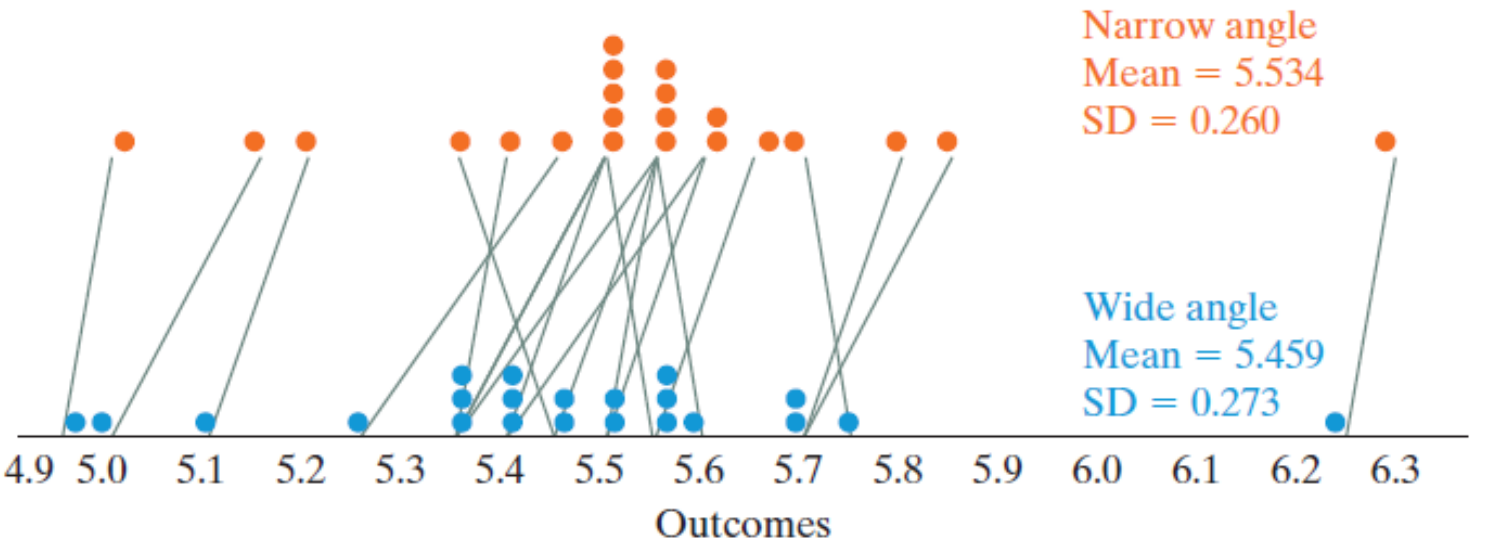


The Differences in Times

- Mean difference is $\bar{x}_d = 0.075$ seconds
- Standard deviation of the differences is $SD_d = 0.0883$ sec.
- This standard deviation of 0.0883 is smaller than the original standard deviations of the running times, which were 0.260 and 0.273.

Rounding First Base

- Below are the original dotplots with each observation paired between the base running strategies.
- What do you notice?



Rounding First Base

- Is the average difference of $\bar{x}_d = 0.075$ seconds significantly different from 0?
- The parameter of interest, μ_d, is the long run mean difference in running times for runners using the narrow angled path instead of the wide angled path. (narrow – wide)

Rounding First Base

The hypotheses:

- H_0: $\mu_d = 0$
  - The long run mean difference in running times is 0.
- H_a: $\mu_d \neq 0$
  - The long run mean difference in running times is not 0.

- The statistic $\bar{x}_d = 0.075$ is above zero.
- *How likely is it to see an average difference in running times this big or bigger by chance alone, even if the base running strategy has no genuine effect on the times?*

Rounding First Base

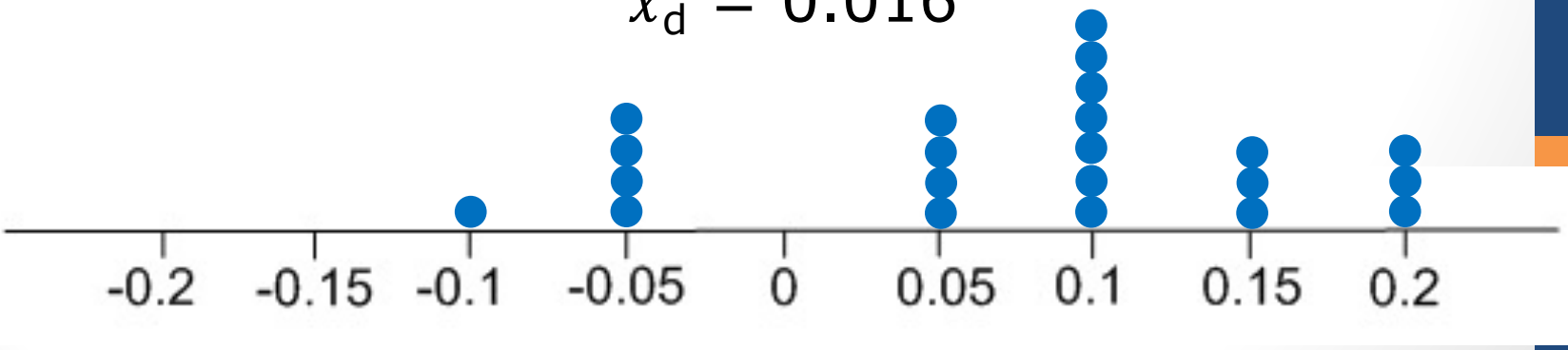
How can we use simulation-based methods to find an approximate p-value?

- The null hypothesis says the running path does not matter.
- So we can use our same data set and, for each runner, randomly decide which time goes with the narrow path and which time goes with the wide path and then compute the difference. (Notice we do not break our pairs.)
- After we do this for each runner, we then compute a mean difference.
- We will then repeat this process many times to develop a null distribution.

Random Swapping

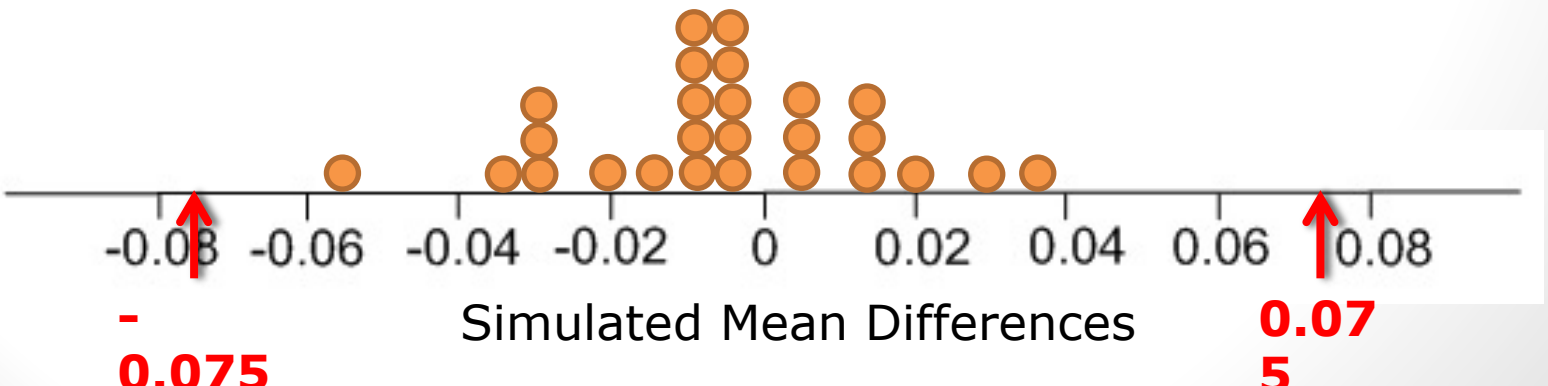
Subject	1	2	3	4	5	6	7	8	9	10	
narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
diff	0.05	-0.05	-0.10	0.10	0.15	0.05	0.05	0.15	0.15	-0.10	…

$\bar{x}_d = 0.016$



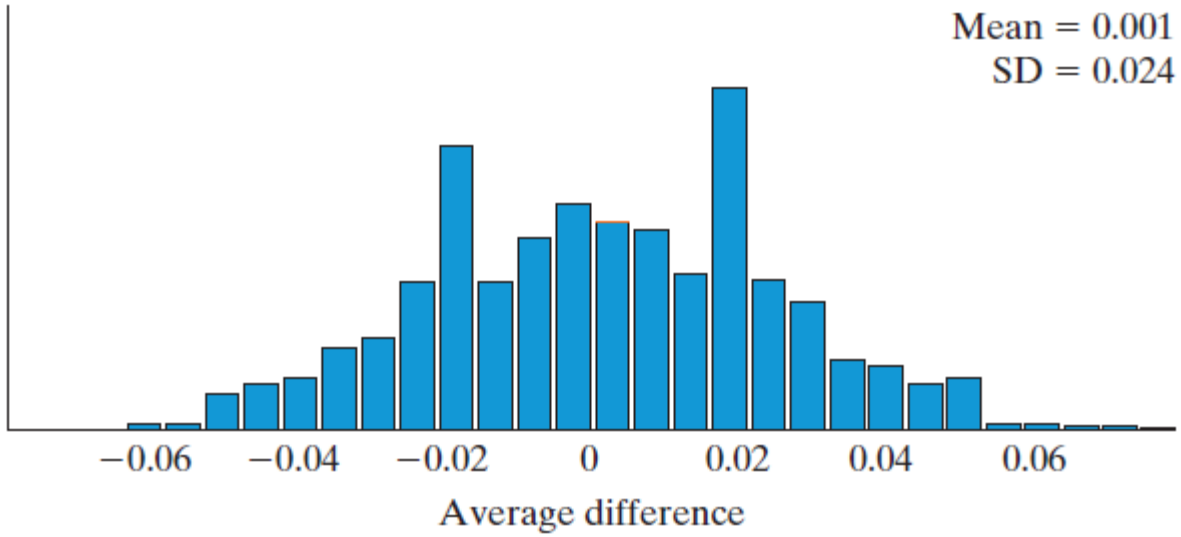
More Simulations

With 26 repetitions of creating simulated mean differences, we did not get any that were as extreme as 0.075.



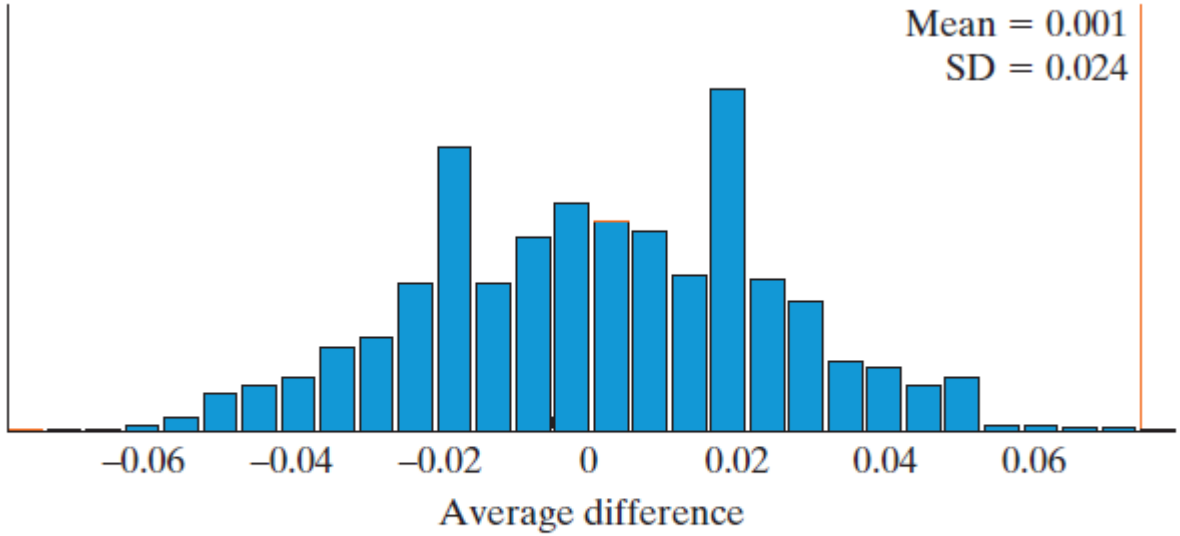
First Base

- Here is a null distribution of 1000 simulated mean differences.
- Notice it is centered at zero, which makes sense in agreement with the null hypothesis.
- Notice also the SD of these MEAN DIFFERENCES is 0.024. This is the SE.
- SD of time differences was 0.0883. SE = SD of mean time diff.s = .024.
- Where is our observed statistic of 0.075?



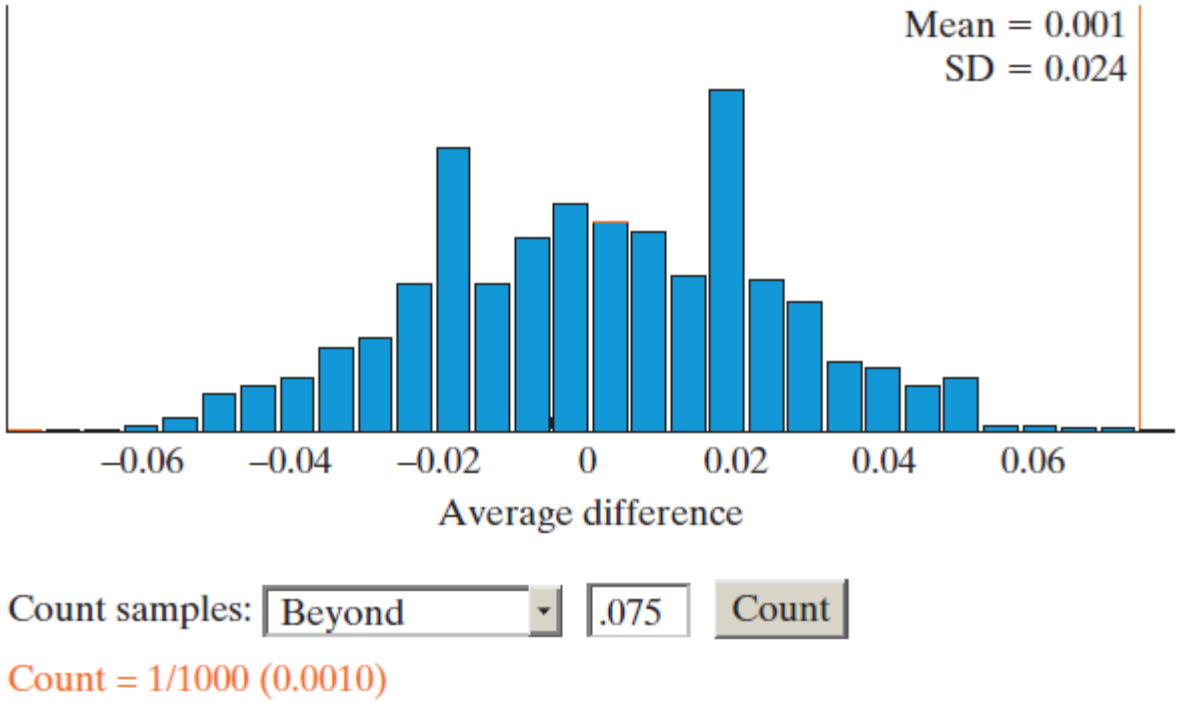
First Base

- Only 1 of the 1000 repetitions of random swappings gave a $\bar{x}_d$ value at least as extreme as 0.075.



First Base

- We can also standardize 0.075 by dividing by the SE of 0.024 to see our standardized statistic = $\frac{0.075}{0.024} = 3.125$.



Rounding First Base

- With a p-value of 0.1%, we have very strong evidence against the null hypothesis. The running path makes a statistically significant difference with the wide-angle path being faster on average.
- We can draw a cause-and-effect conclusion since the researcher used random assignment of the two base running methods for each runner.
- There was not much information about how these 22 runners were selected though so it is unclear if we can generalize to a larger population.

3S Strategy

- **Statistic:** Compute the statistic in the sample. In this case, the statistic we looked at was the observed mean difference in running times.
- **Simulate:** Identify a chance model that reflects the null hypothesis. We tossed a coin for each runner, and if it landed heads we swapped the two running times for that runner. If the coin landed tails, we did not swap the times. We then computed the mean difference for the 22 runners and repeated this process many times.
- **Strength of evidence:** We found that only 1 out of 1000 of our simulated mean differences was at least as extreme as the observed difference of 0.075 seconds.

First Base

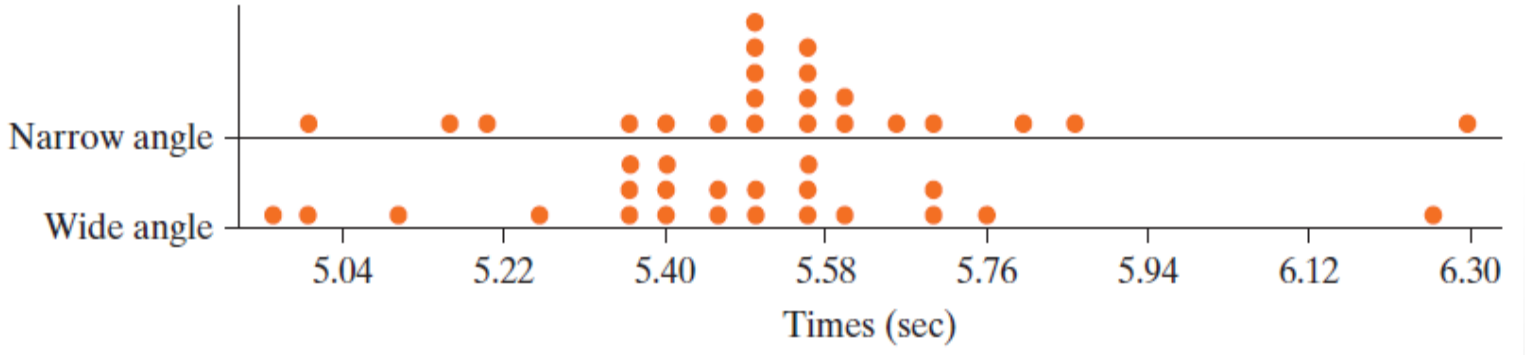
- Approximate a 95% confidence interval for μ_d:
  - 0.075 ± 1.96(0.024) seconds.
  - (0.028, 0.122) seconds.
- What does this mean?
  - We are 95% confident that, if we were to keep testing this indefinitely, the narrow angle route would take somewhere between 0.028 to 0.122 seconds longer on average than the wide angle route.

Since n = 22 here, the sample size is pretty small and the multiplier of 1.96 is not quite correct. If we assume the population of differences is normal, we should use a t multiplier, which here would be 2.08, so the 95% CI would be (.025, .125).

First Base

Alternative Analysis

- What do you think would happen if we wrongly analyzed the data using a 2 independent samples procedure? (i.e. The researcher selected 22 runners to use the wide method and an independent sample of 22 other runners to use the narrow method, obtaining the same 44 times as in the actual study.

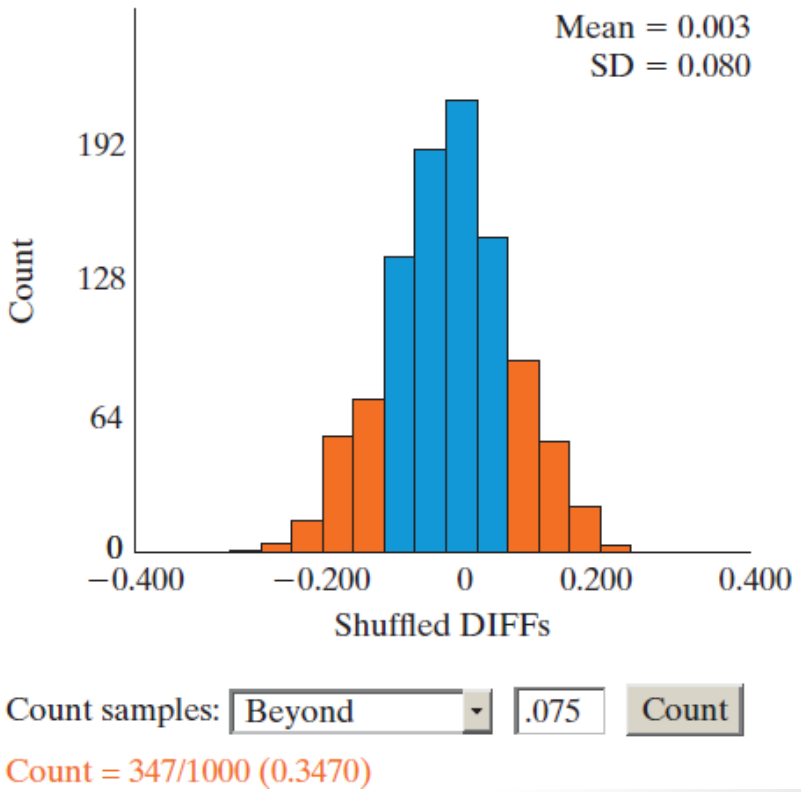


First Base

Ignoring the fact that it is paired data,

we get a p-value of 0.3470.

Does it make sense that this p-value is larger than the one we obtained earlier?



4. Theory based approach for Analyzing Data from Paired Samples, and M&Ms.

Section 7.3

How Many M&Ms Would You Like?

Example 7.3

How Many M&Ms Would You Like?

- Does your bowl size affect how much you eat?
- Brian Wansink studied this question with college students over several days.
- At one session, the 17 participants were assigned to receive either a small bowl or a large bowl and were allowed to take as many M&Ms as they would like.
- At the following session, the bowl sizes were switched for each participant.

How Many M&Ms Would You Like?

- What are the observational units?
- What is the explanatory variable?
- What is the response variable?
- Is this an experiment or an observational study?
- Will the resulting data be paired?

How Many M&Ms Would You Like?

The hypotheses:

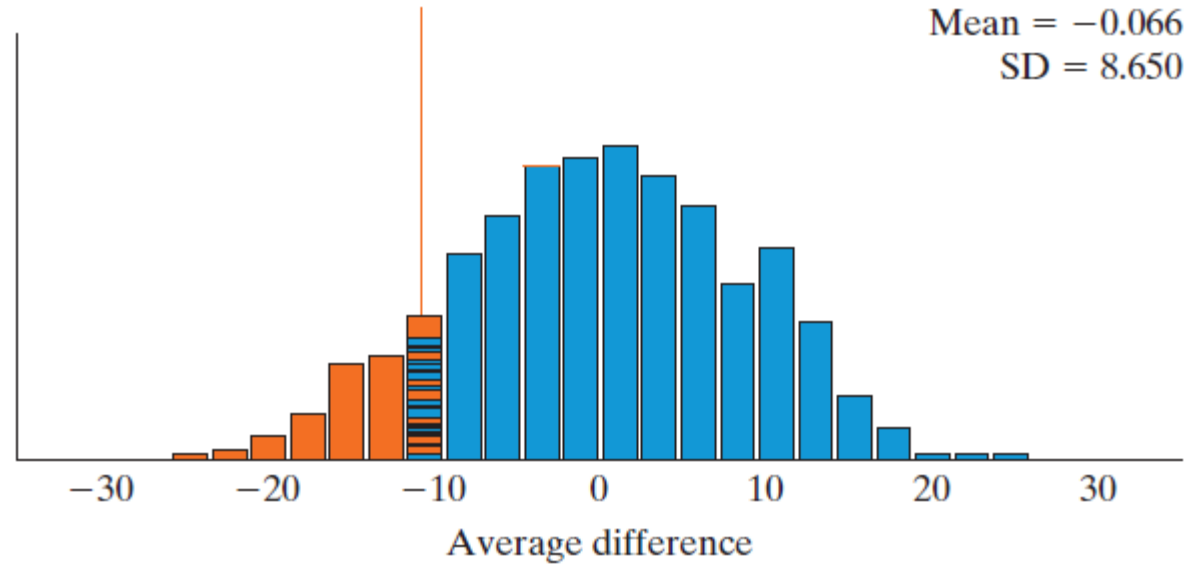
- H_0: $\mu_d = 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is 0.
- H_a: $\mu_d < 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is less than 0.

TABLE 7.5 Summary statistics, including the difference (small – large) in the number of M&Ms taken between the two bowl sizes

Bowl size	Sample size, *n*	Sample mean	Sample SD
Small	17	$\bar{x}_s = 38.59$	$s_s = 16.90$
Large	17	$\bar{x}_l = 49.47$	$s_l = 27.21$
Difference = small − large	17	$\bar{x}_d = -10.88$	$s_d = 36.30$

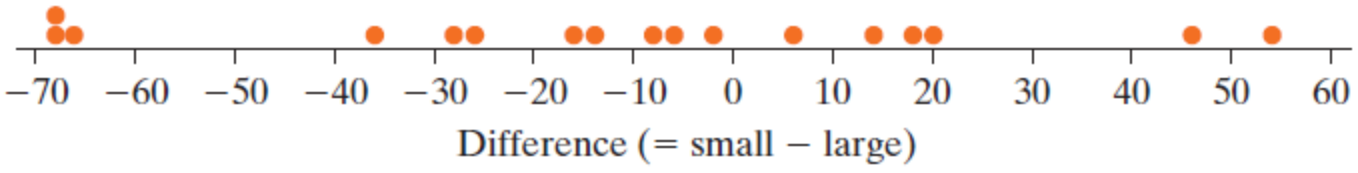
How Many M&Ms Would You Like?

- Here are the results of a simulation-based test.
- The p-value is quite large at 0.1220.



How Many M&Ms Would You Like?

- Our null distribution was centered at zero and fairly bell-shaped.
- Theory-based methods using the t distribution should be valid if σ is unknown and the population distribution of differences is normal (we can guess at this by looking at the sample distribution of differences). Alternatively, we can use the normal distribution if our sample size is at least 30.
- Our sample size was only 17, but this distribution of differences looks pretty normal, so we will proceed with a t-test.



Theory-based test

$$t = \frac{\bar{x}_d}{s_d/\sqrt{n}}$$

- This kind of test is called a paired *t*-test.

Theory-based results

Scenario: One mean

☐ Paste data

n: 17

mean, $\bar{x}$: −10.88

sample sd, s: 36.3

Calculate

☑ Confidence interval

confidence level 95 % Calculate CI

(−29.5435, 7.7835)

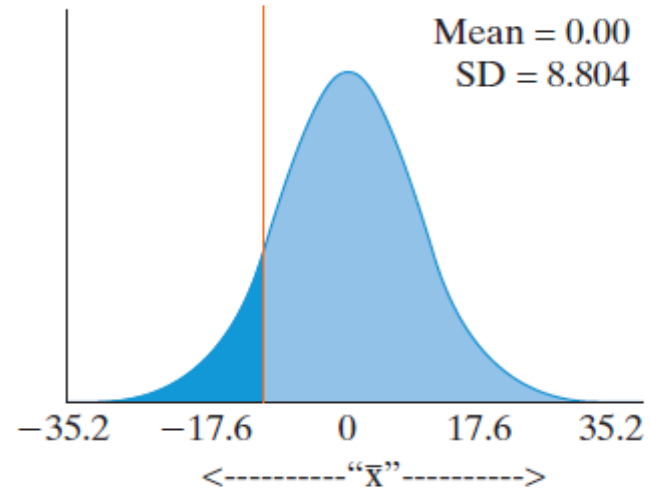
Theory-based inference

☑ Test of significance

H_0: μ = 0

H_a: μ < 0

Calculate



Standardized statistic t = −1.24 df = 16

p-value 0.1172

Conclusion

- The theory-based test gives slightly different results than simulation, 11.7% instead of 12.2% for the p-value, but we come to the same conclusion. We do not have strong evidence that the bowl size affects the number of M&Ms taken.
- We can see this in the large p-value (0.1172) and the confidence interval that included zero (-29.5, 7.8).
- The confidence interval tells us that we are 95% confident that when given a small bowl, people will take somewhere between 29.5 fewer M&Ms to 7.8 more M&Ms on average than when given a large bowl.

Why wasn't the difference statistically significant?

- There could be a number of reasons we didn’t get significant results.
  - Maybe bowl size doesn’t matter.
  - Maybe bowl size does matter and the difference was too small to detect with our small sample size.
  - Maybe bowl size does matter with some foods, like pasta or cereal, but not with a snack food like M&Ms.

Strength of Evidence

- We will have stronger evidence against the null (smaller p-value) when:
  - The sample size is increased.
  - The variability of the data is reduced.
  - The effect size, or mean difference, is farther from 0.
- We will get a narrower confidence interval when:
  - The sample size is increased.
  - The variability of the data is reduced.
  - The confidence level is decreased.