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

No lecture Tue May 12!!!

1. When to use which formula.
2. Bicycling to work example, comparing two means.
3. Paired data, studying with music example.

No lecture Tue May 12!!!

HW3, due Fri, May22, 1159pm. 4.CE.10, 5.3.28, 6.1.17, and 6.3.14.
The problems are on the next 4 slides.
In 5.3.28d, use the theory-based formula. You do not need to use an applet.
Read ch7 and 10.
The course website is http://www.stat.ucla.edu/~frederic/13/S26 .

Spanking and IQ

4.CE.10 Studies have shown that children in the U.S. who have been spanked have a significantly lower IQ score on average than children who have not been spanked.

a. Is it legitimate to conclude from this study that spanking a child causes a lower IQ score? Explain why or why not.

b. Explain why conducting a randomized experiment to investigate this issue (of whether spanking causes lower IQs) would be possible in principle but ethically objectionable.

Reading *Harry Potter**

4.CE.11 You want to investigate whether teenagers in the United Kingdom (UK) tend to have read more *Harry Potter* books, on average, than teenagers in the United States (US).

a. Identify and classify (as categorical or quantitative) the explanatory and response variable.

b. Would you ideally use random sampling for this study, or random assignment, or both? Explain.

Restaurant customer behavior

h. Use an appropriate applet to find and report the following from the data:
   - The standardized statistic
   - The theory-based p-value

i. How do the simulation-based and theory-based p-values compare?

5.3.28 Recall the data from the Physicians' Health Study: Of the 11,034 physicians who took the placebo, 138 developed ulcers during the study. Of the 11,037 physicians who took aspirin, 169 developed ulcers.

a. Define the parameters of interest. Assign symbols to these parameters.

b. State the appropriate null and alternative hypotheses in symbols.

c. Explain why it would be okay to use the theory-based method (that is, normal distribution based method) to find a confidence interval for this study.

d. Use an appropriate applet to find and report the theory-based 95% confidence interval.

e. Does the 95% confidence interval contain 0? Were you expecting this? Explain your reasoning.

f. Interpret the 95% confidence interval in the context of the study.

g. Use the 95% confidence interval to state a conclusion about the strength of evidence in the context of the study.

h. Relatively speaking, is the 95% confidence interval narrow or wide? Explain why that makes sense.

5.3.29 Recall the data from the Physicians' Health Study: Of the 11,034 physicians who took the placebo, 138 developed ulcers during the study. Of the 11,037 physicians who took aspirin, 169 developed ulcers.

every week. Be sure to compare and contrast the shape, center, and spread for study hours' distributions for males and females.

6.1.16 Reconsider the data in the previous question about number of hours spent studying.

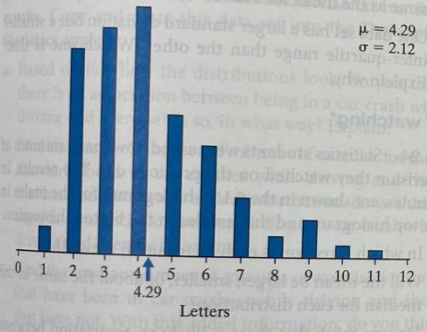
a. Find the median number of study hours for both males and females. What do these numbers tell us about the two data sets?

b. Find the inter-quartile range for the number of study hours for both males and females. What do these numbers tell us about the two data sets?

c. Construct parallel boxplots by hand for the two data sets.

Gettysburg Address

6.1.17 The graph below displays the distribution of word lengths (number of letters) in the Gettysburg Address, which you explored in Exploration 2.1A.



a. Describe the shape of this distribution.

b. Based on this shape, do you expect the median to be less than the mean, greater than the mean, or very close to the mean? Explain.

The following table lists how often each of the word lengths appears for these 268 words.

Word length	1	2	3	4	5	6	7	8	9	10	11
Number of words	7	49	54	59	34	27	15	6	10	4	3

c. Determine the median word length of these 268 words.

d. The mean word length is 4.29 letters per word. Is the median greater than, less than, or very close to the mean? Does this confirm your answer to part (b)?

e. Calculate the five-number summary of the word lengths.

College student bedtimes*

6.1.18 In a survey, 30 college students were asked what their usual bedtime was and the results are shown in the 6.1.18 dotplot in terms of hours after midnight. Negative responses are hours before midnight.

a. Determine the five-number summary for the bed times.

b. What is the inter-quartile range?

c. The earliest bedtime is 11:30 PM (represented by −0.50 on the graph). If that person's usual bedtime is actually 9:00 PM and that change was made in the dotplot, does that change the inter-quartile range? Would it change the standard deviation?

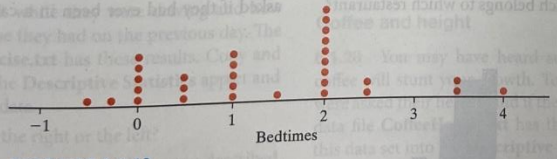
Candy bars

6.1.19 Weights of 20 Mounds* candy bars and 20 PayDay* candy bars, in grams, are shown in the 6.1.19 dotplots.

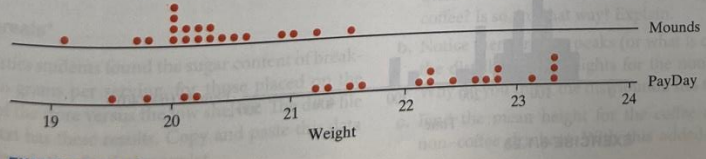
a. Describe how the distributions of weights of the two types of candy bars differ in both variability and center.

b. Based on your answers to part (a), which set of candy bar weights has the lowest standard deviation? Which has the lowest mean?

c. Would you say there is an association between the type of candy bar and the weight? Why or why not?



EXERCISE 6.1.18



EXERCISE 6.1.19

h. Summarize your conclusions about the research question of the study. Be sure to comment on statistical significance, confidence/estimation, causation, and generalization.

Perceived wealth

6.3.13 Do people tend to spend money differently based on perceived changes in wealth? In a study conducted by Epley et al. (2006), 47 Harvard undergraduates were randomly assigned to receive either a "bonus" check of $50 or a "rebate" check of $50. A week later, each student was contacted and asked whether they had spent any of that money, and if yes, how much. In this exercise we will focus on how much money they recalled spending when contacted a week later. It turned out that those in the "bonus" group spent an average of about $22, compared to $10 in the "rebate" group.

a. Identify the observational units.

b. Identify the explanatory and response variables. Identify each as either categorical or quantitative.

c. State the appropriate null and alternative hypotheses in the context of the study.

d. In the article that appeared in the *Journal of Behavioral Decision Making*, the researchers reported neither the sample size nor the sample SD of each group. In this exercise you will explore whether and how the strength of evidence is impacted by the sample size and sample SD. Complete the following table by finding the *t*-statistic and a p-value for a theory-based test of significance comparing two means under each of the four different scenarios.

e. Summarize what your analysis has revealed about the effects of the sample size breakdown and the sample standard deviations on the values of the *t*-statistic and p-value.

Nostril breathing and cognitive performance*

6.3.14 In an article titled "Unilateral Nostril Breathing Influences Lateralized Cognitive Performance" that appeared in *Brain and Cognition* (1989), researchers Block et al. published results from an experiment involving assessments of spatial and verbal cognition when breathing through only the right versus left nostril.

The subjects were 30 male and 30 female right-handed introductory psychology students who volunteered to participate in exchange for course credit. Initial testing on spatial and verbal tests revealed the following summary statistics. Note that the scores on the spatial task can range from 0 to 40, whereas those on the verbal task can go from 0 to 20. The distributions are not strongly skewed on either scale or for males or females.

	Spatial		Verbal	
Sex	Mean	SD	Mean	SD
Male	10.20	2.70	10.90	3.00
Female	7.80	2.50	15.10	3.40

a. Consider comparing males to females with regard to performance on the spatial assessment task. State the appropriate null and alternative hypotheses in the context of the study.

b. Explain why it is valid to use the theory-based method for producing a p-value to test the hypotheses stated in part (a).

c. Carry out the appropriate test to produce a p-value to test the hypotheses stated in part (a) and interpret the p-value.

d. Find a 95% confidence interval for the difference in mean scores of males and females with regard to performance on spatial assessments. Interpret the interval.

e. Based on your p-value, state a conclusion in the context of the study. Be sure to comment on statistical significance, estimation (confidence interval), causation, and generalization.

f. Repeat the investigation comparing males and females, this time on verbal performance. Be sure to address the questions asked in parts (a)–(e).

Scenario		Sample sizes	Sample means	Sample SDs	*t*-statistic	p-value
1	Bonus	24	22	5		
	Rebate	23	10	5		
2	Bonus	24	22	10		
	Rebate	23	10	10		
3	Bonus	30	22	5		
	Rebate	17	10	5		
4	Bonus	30	22	10		
	Rebate	17	10	10		

EXERCISE 6.3.13

When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws are normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} s/√n.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 s/√n.

n ≥ 30 is often considered large enough to use 1.96.

In practice, we typically do not know the draws are normal, but if the distribution looks roughly symmetrical without enormous outliers, the t formula may be reasonable.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

s = √[p(1-p)], σ = $[\pi(1-\pi)]$.

When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws ~ normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} s/√n.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 s/√n.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

s = √[p(1-p)], σ = $[\pi(1-\pi)]$.

If n is large and π is unknown, use $\bar{x}$ +/- 1.96 s/√n.

Here large n means ≥ 10 of each type in the sample.

When to use which formula.

What if n is small and the draws are not normal, and you want a theory-based test or CI?

How should you find the t multiplier for a CI or a p-value using the t-statistic, when n is small?

These are questions outside the scope of this course, but some techniques have been developed, such as the bootstrap, which are sometimes useful in these situations.

When to use which formula.

c. Numerical data from 2 samples, iid observations, want a 95% CI for $\mu_1 - \mu_2$.

If n is large and σ is unknown, use $\bar{x}_1 - \bar{x}_2 +/- 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

As with one sample, if σ_1 is known, replace s_1 with σ_1, and the same for σ_2. And as with one sample, if σ_1 and σ_2 are unknown, the sample sizes are small, and the distributions are roughly normal, then use t_{mult} instead of 1.96. If the sample sizes are small, the distributions are normal, and σ_1 and σ_2 are known, then use 1.96.

d. Binary data from 2 samples, iid observations, want a 95% CI for $\pi_1 - \pi_2$.

same as in c above, with $p_1 = \bar{x}_1$, $s_1 = \sqrt{[p_1(1-p_1)]}$, $\sigma_1 = \quad [\pi_1(1-\pi_1)]$.

Large for binary data means sample has ≥ 10 of each type.

For testing, use pooled estimate of p for the SE.

For CIs for the difference in proportions,

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

In testing the difference in proportions,

$$\text{SE} = \sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}\right)}$$

where $\hat{p}$ is the proportion in both groups combined.

T-test versus Z-test.

For 1 sample numerical data, iid observations.

$z = (\bar{x} - \mu) \div (s/\sqrt{n})$. If you know σ, use σ in place of s.

$t = (\bar{x} - \mu) \div (s/\sqrt{n})$. Same formula.

$(s/\sqrt{n})$ = SE for the mean.

The only difference is how you convert it to a p-value. With z, you use the normal distribution to find a p-value, and with t, you use the t distribution.

If n is small, population is normal and σ is unknown, should call it t.

If n is large, calling it z is correct.

If n is large and population is normal and σ is unknown, calling it z is correct and calling it t is also correct. Technically, if you really know population is normal, then calling it t would be preferable. But when n is large p-value will be essentially the same anyway.

$n \geq 30$ is often considered large sample size for quantitative data.

For 0-1 data, must have ≥10 of each type in your sample.

T-test versus Z-test.

For 1 sample categorical data, iid observations.

z = $(\bar{x}$ - μ) ÷ (s/√n) . If you know σ, use σ in place of s.

Never use t for categorical data because the population cannot be normal.

For 0-1 data,

p = $\bar{x}$, and

s = √[p(1-p)], σ = $[\pi(1-\pi)]$.

For 0-1 data, must have ≥10 of each type in your sample.

For testing the difference between means of 2 groups for quantitative data, still use

(observed difference - expected difference under Ho) / SE,

where now

$$\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\,.$$

$s_1 =$ standard deviation of group 1,

$s_2 =$ standard deviation of group 2.

Here expected difference under Ho is always 0.

For testing the difference in proportions for 2 groups, still use (observed difference - expected difference under Ho) / SE, where now

$$\text{SE} = \sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}\right)}$$

$\hat{p}$ is the pooled proportion.

It is the proportion of 1's in both groups combined.

Again, with 2 groups, expected difference under Ho is always 0.

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.

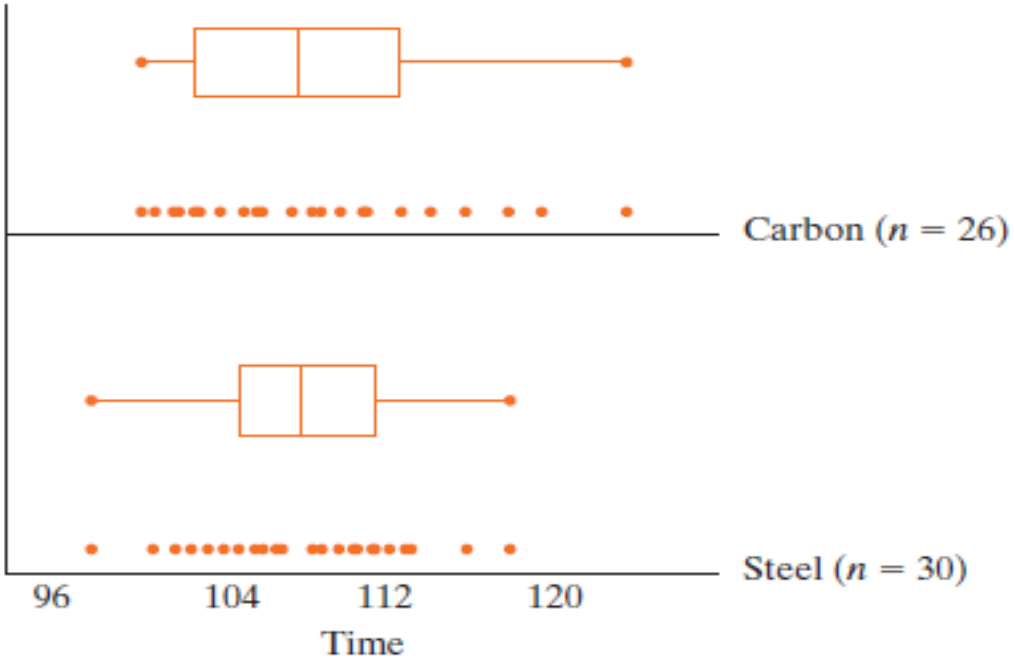
- $\mathbf{H_0}$: $\mu_{carbon} = \mu_{steel}$
- $\mathbf{H_a}$: $\mu_{carbon} \neq \mu_{steel}$.

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 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				

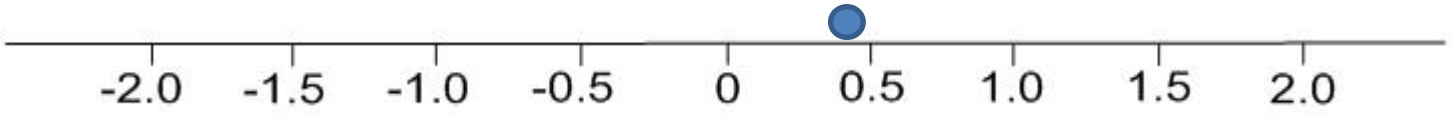
mean = 108.27

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				

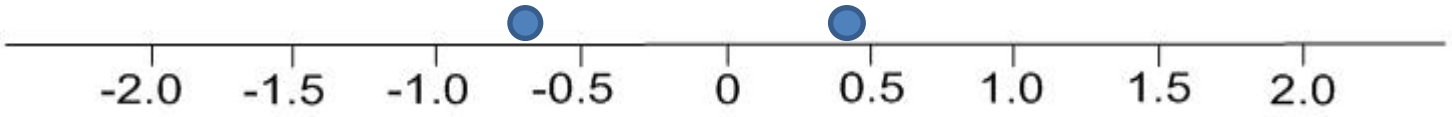
mean = 108.37

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 = 107.69

107.69 – 108.37 = -0.68



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				

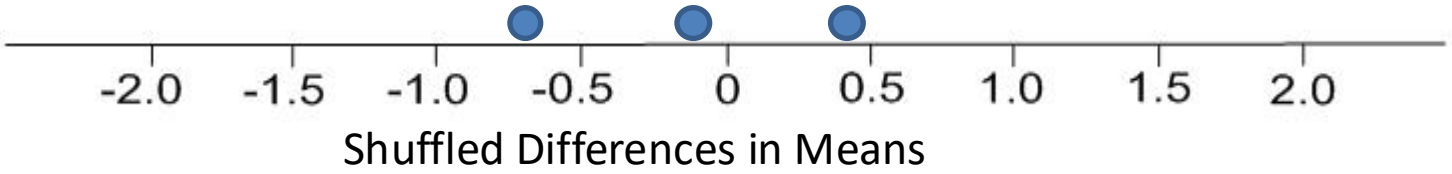
mean = 107.97

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

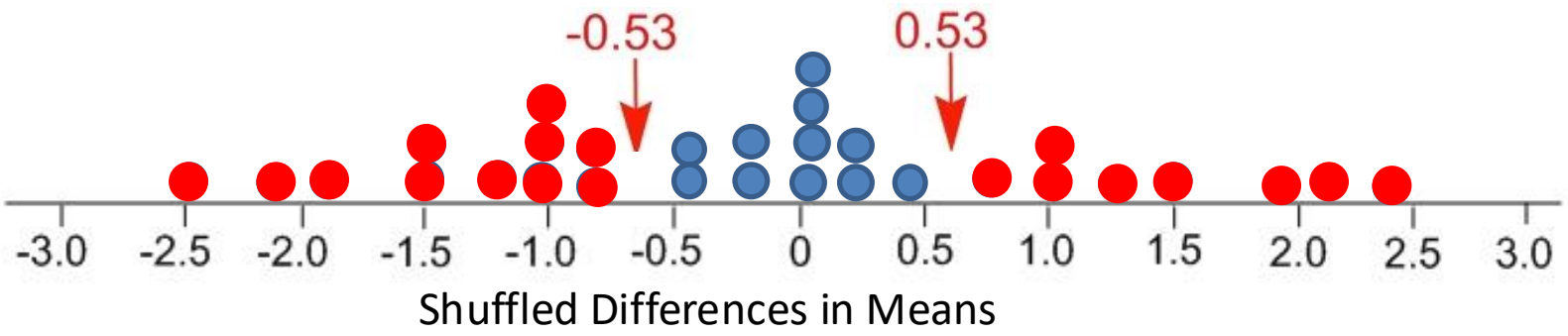
mean = 108.13

107.97 – 108.13 = -0.16



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.

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

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.

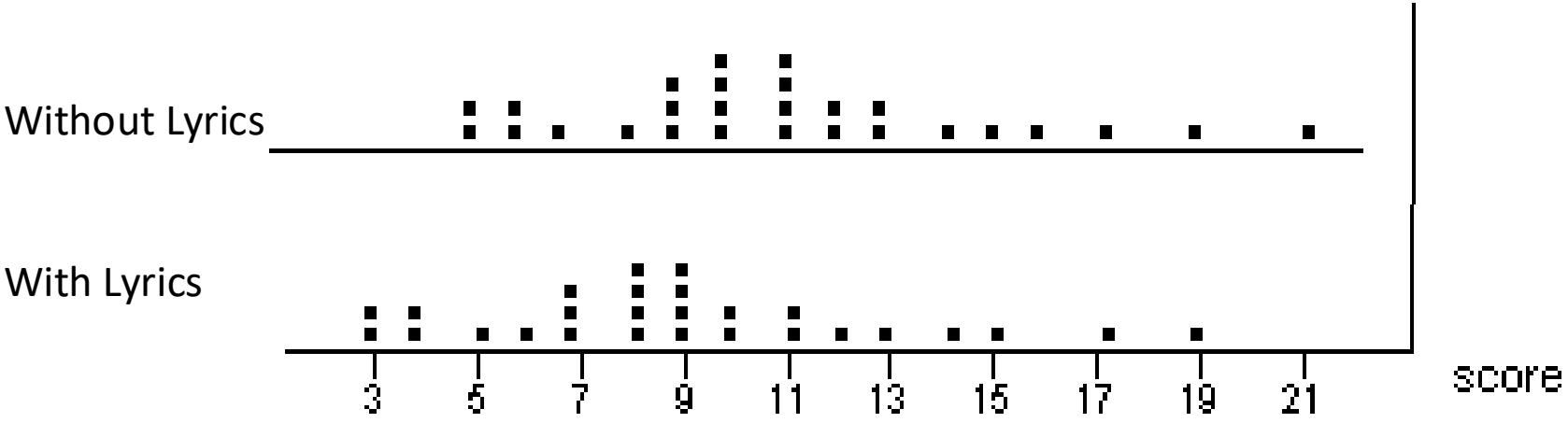
Experiment C — Paired design using matching

- 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

Pairing 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.