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

1. Comparing two means and biking to work example, continued.
2. Paired data, music and studying example.
3. Simulations for paired data, and rounding first base example.

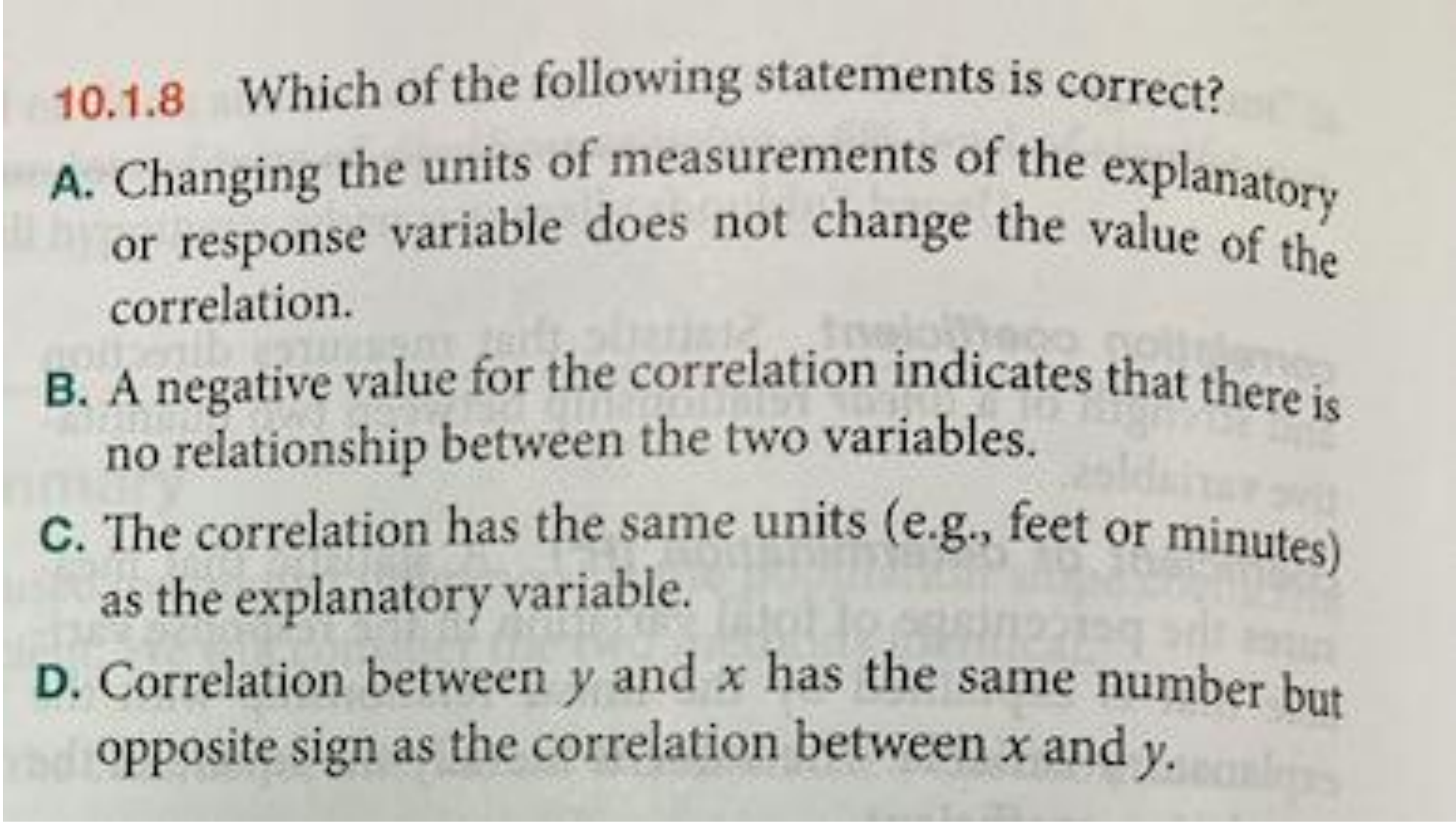
Read chapter 7.

HW4 is due Wed, Mar12, 1159pm. 10.1.8, 10.3.14, 10.3.21, and 10.4.11.
The problems are on the next 5 slides.

The course website is http://www.stat.ucla.edu/~frederic/13/W25 .

If I haven't given your midterm back to you yet, I can do so after class.

10.1.8.



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.

10.3.12 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.

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.

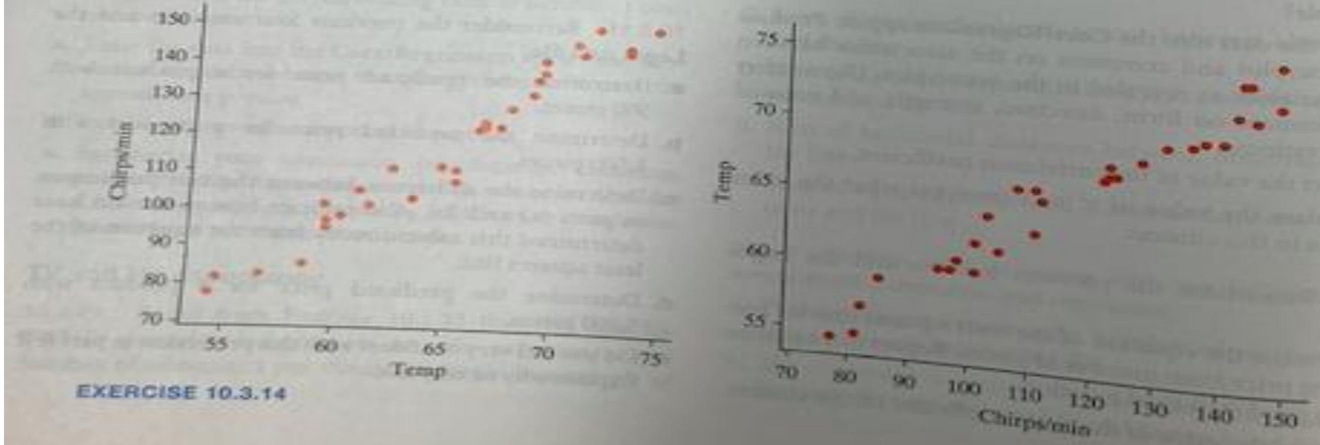
a. Describe the association between these variables.

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.

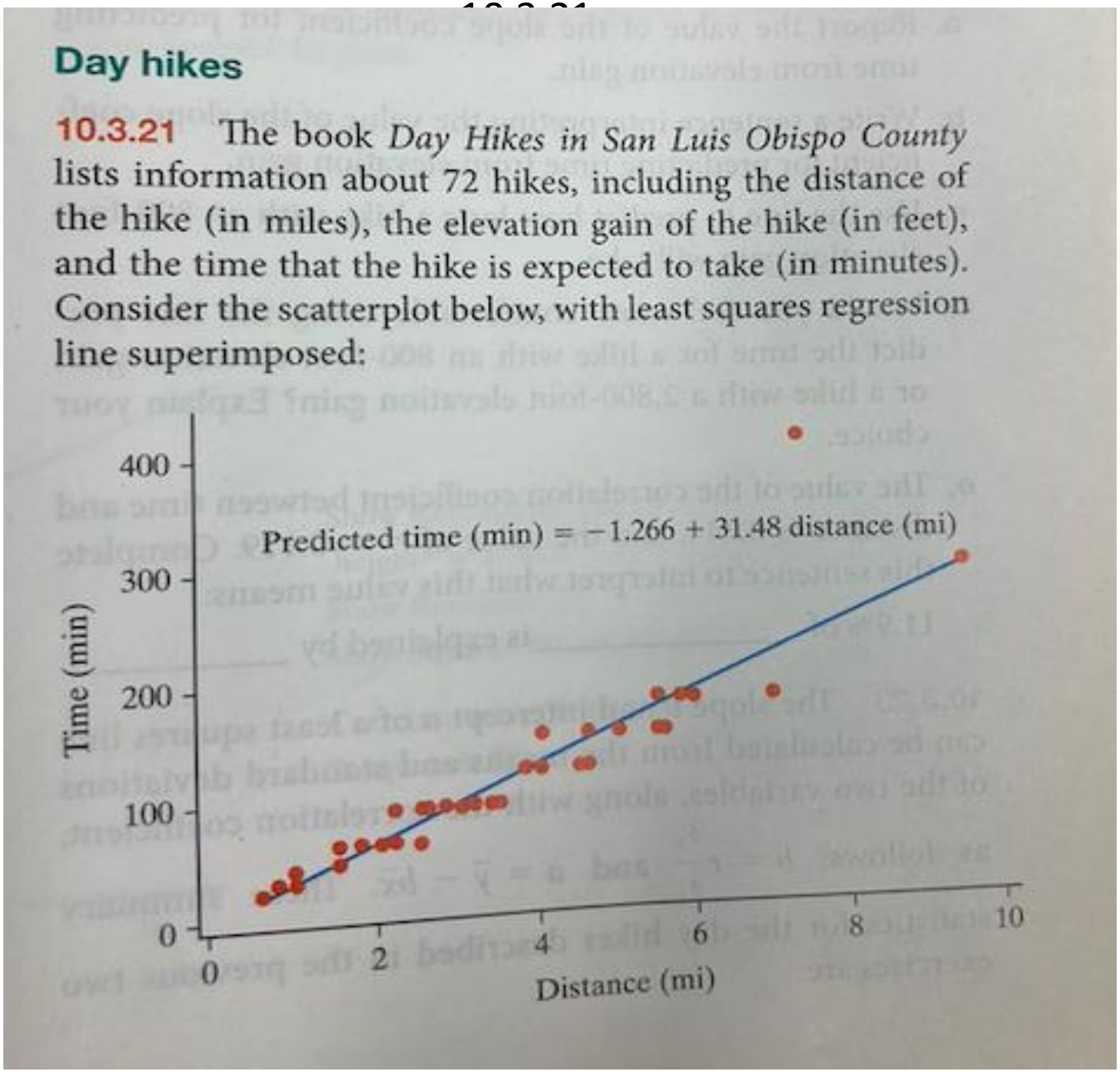
d. Interpret the value of the intercept coefficient. Is this a context in which the intercept coefficient is meaningful?

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



EXERCISE 10.3.14

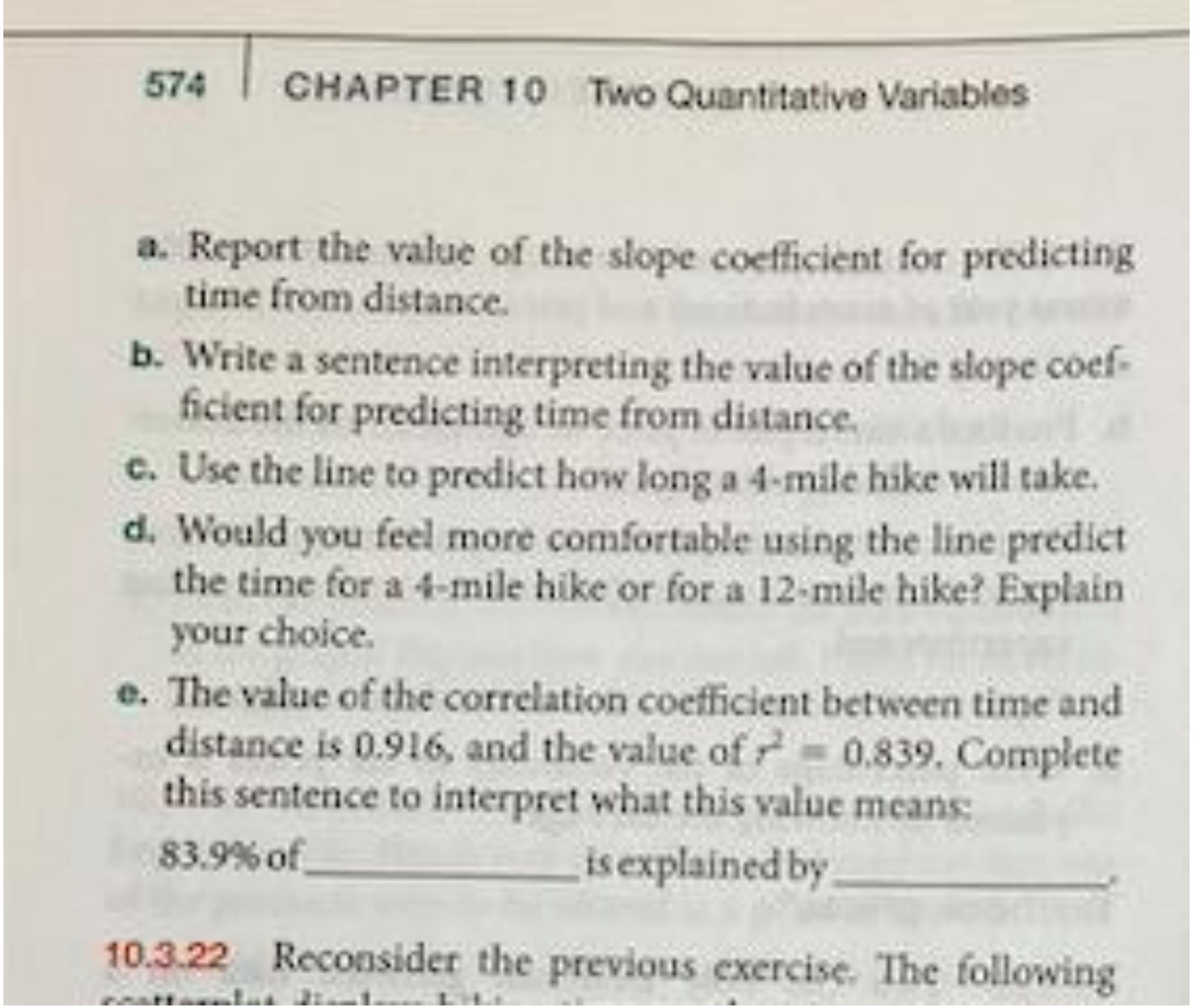
10.3.21



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:

10.3.21.



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.

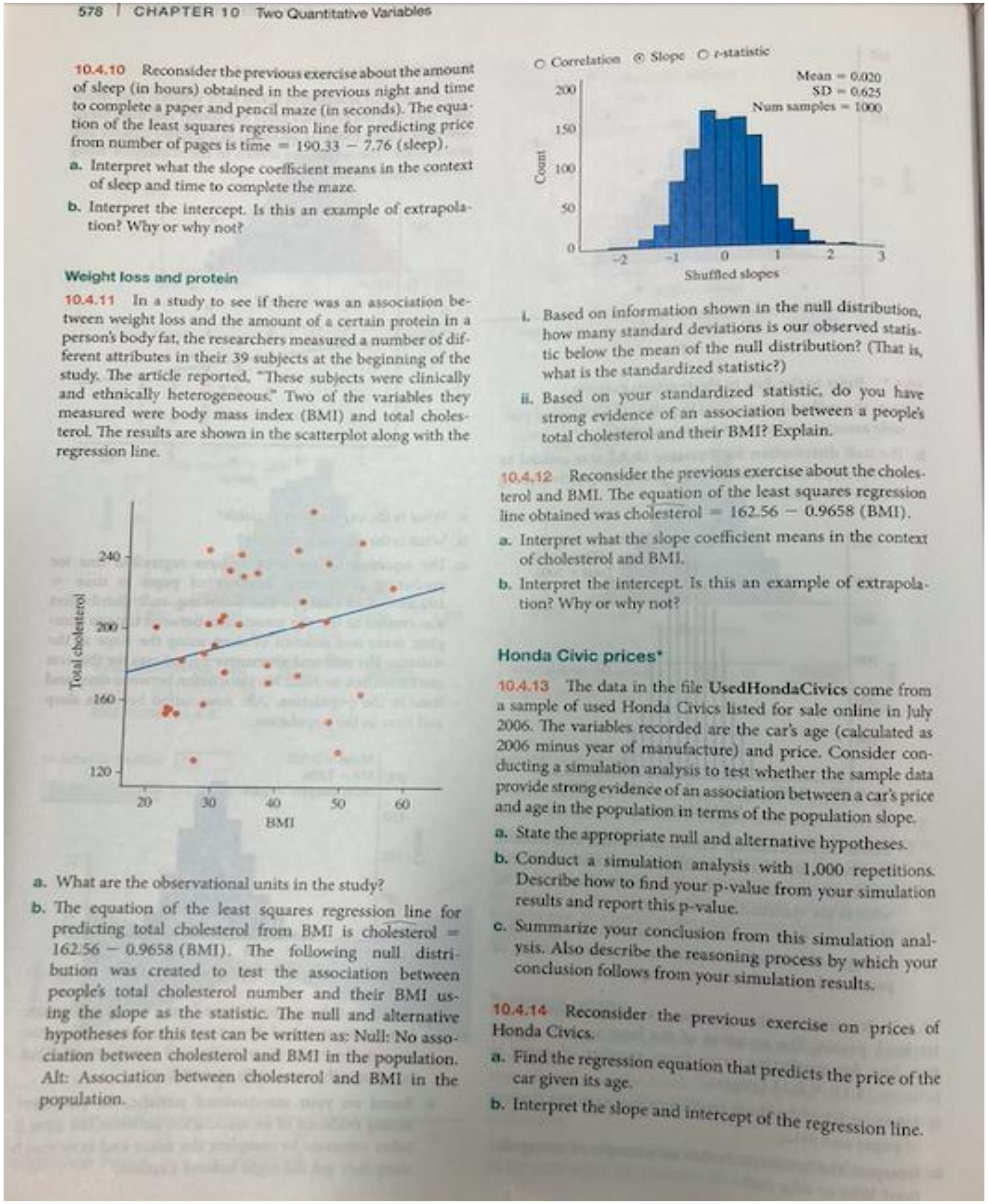
c. Use the line to predict how long a 4-mile hike will take.

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



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 $\widehat{\text{time}}$ = 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 $\widehat{\text{cholesterol}}$ = 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 $\widehat{\text{cholesterol}}$ = 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

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

- 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.

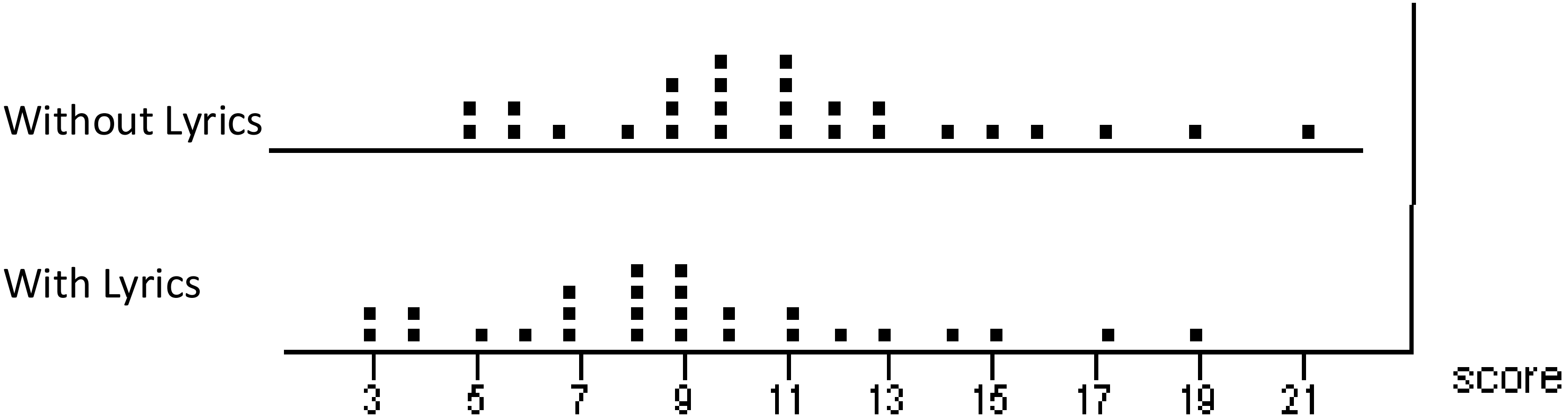
Experiment C — Paired design using matching

- 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

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.

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

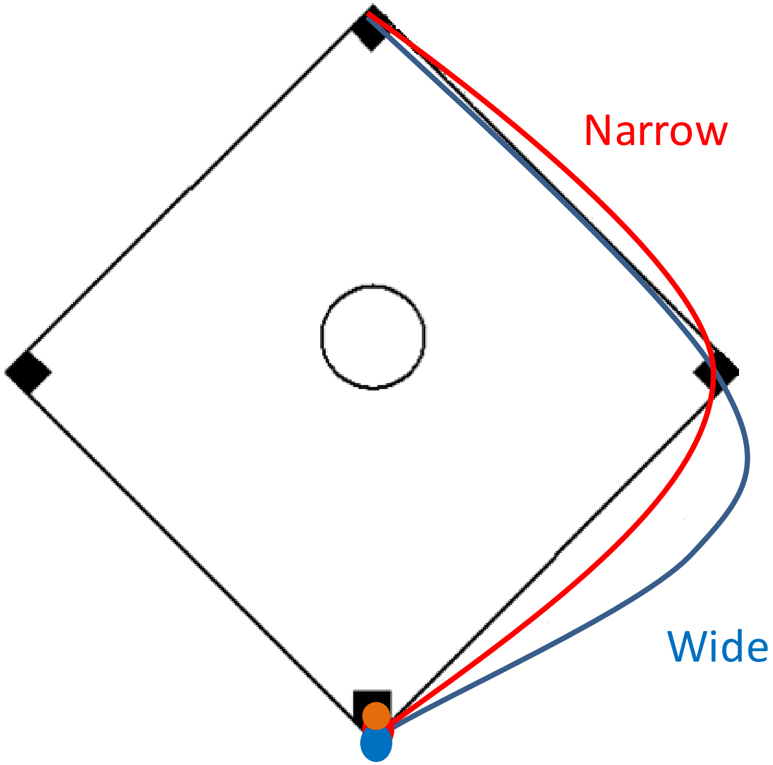
Section 7.2

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

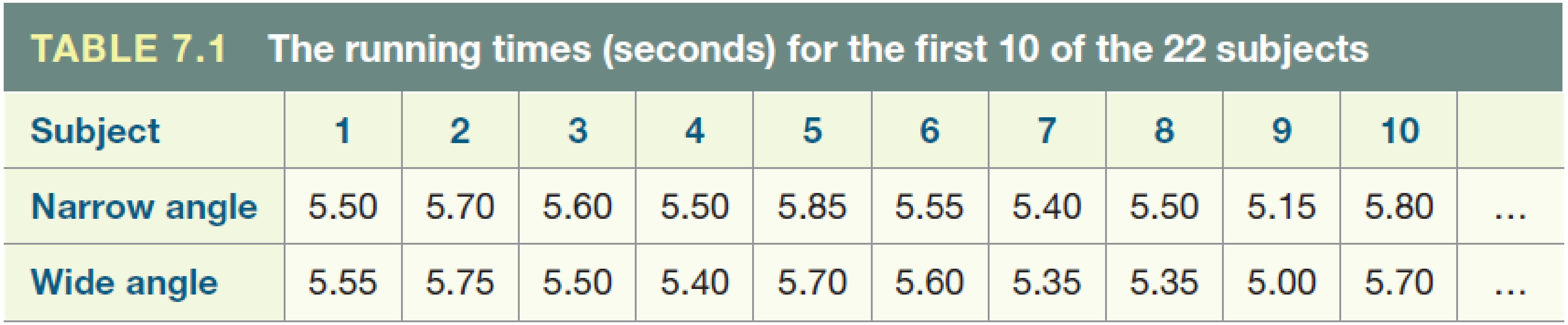
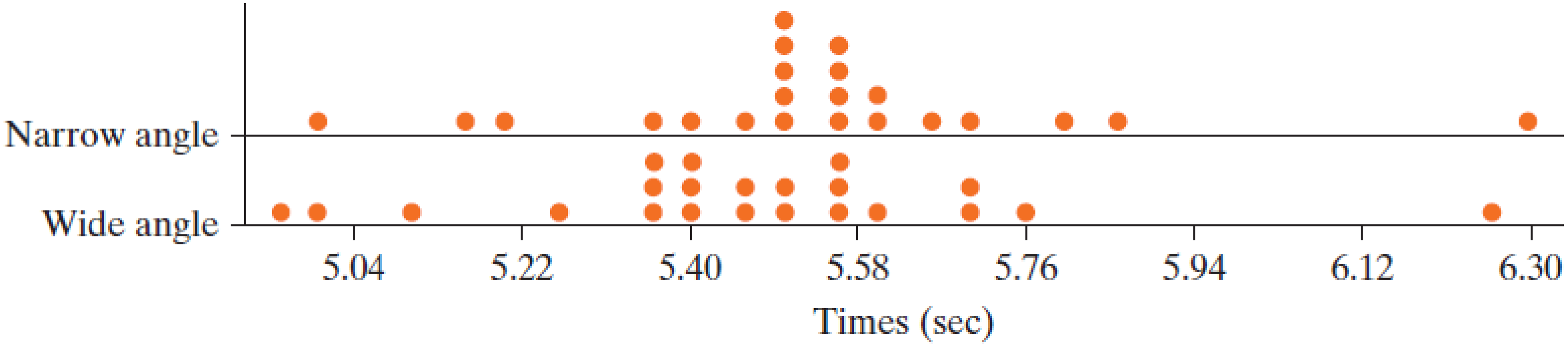


TABLE 7.1 The running times (seconds) for the first 10 of the 22 subjects

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

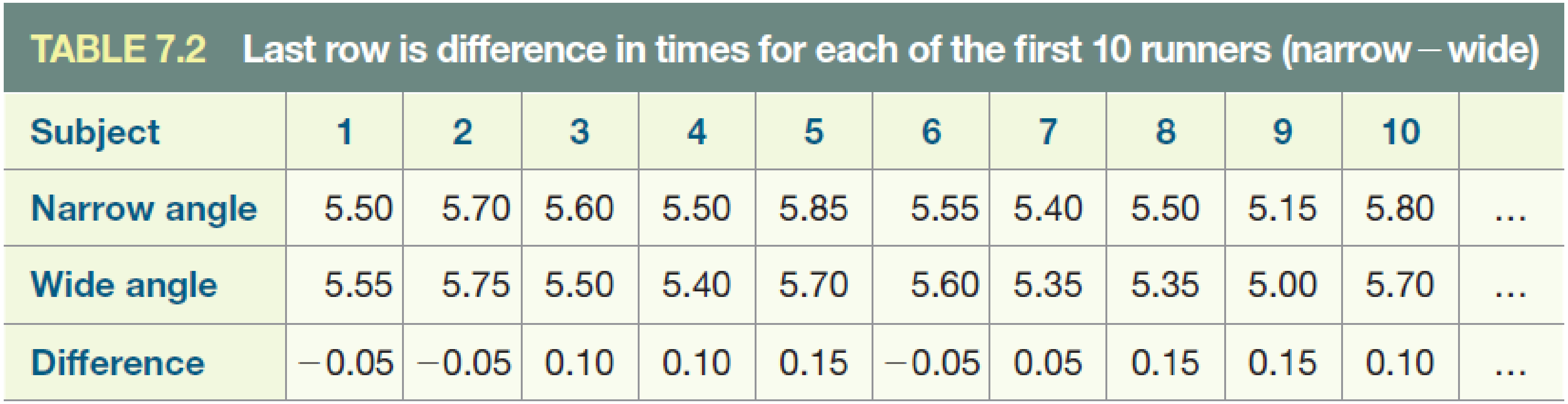
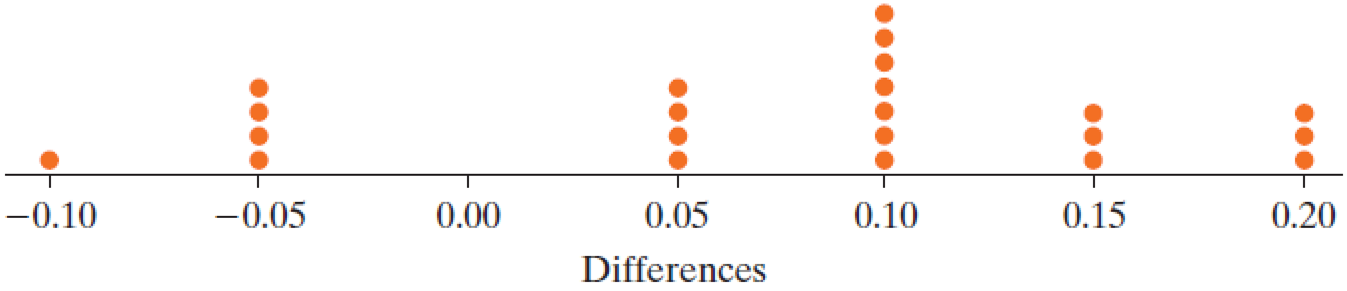


TABLE 7.2 Last row is difference in times for each of the first 10 runners (narrow − wide)

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	...

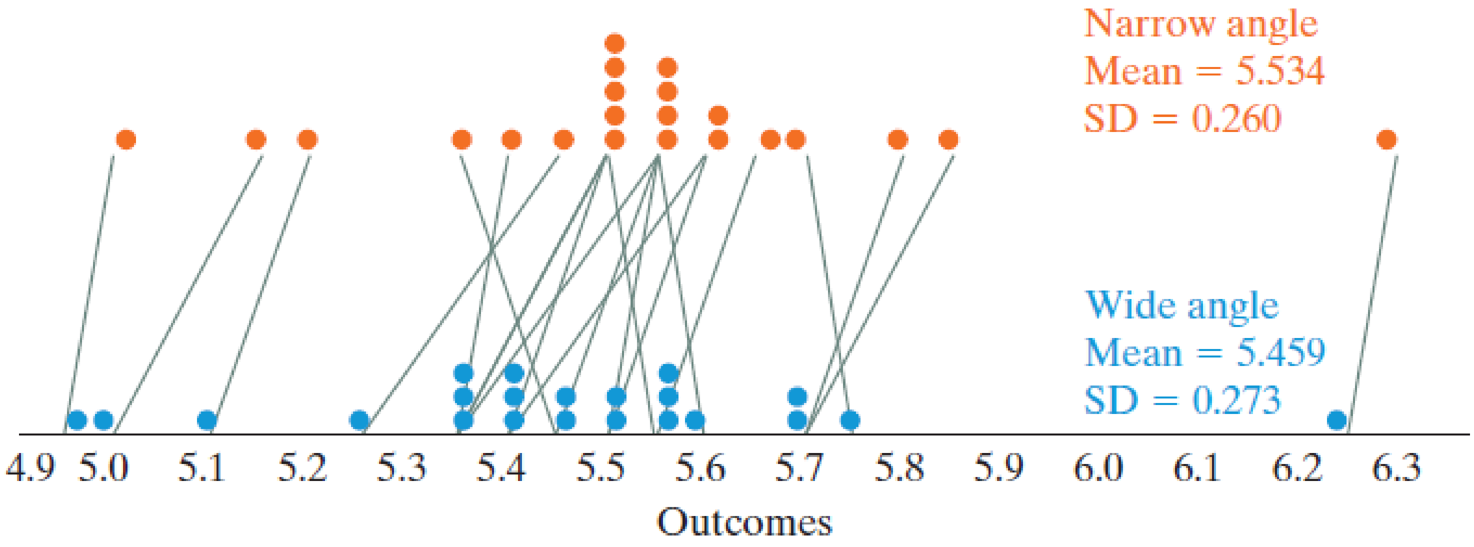


The Differences in Times

- Mean difference is $\bar{x}_{\mathrm{d}}$ = 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

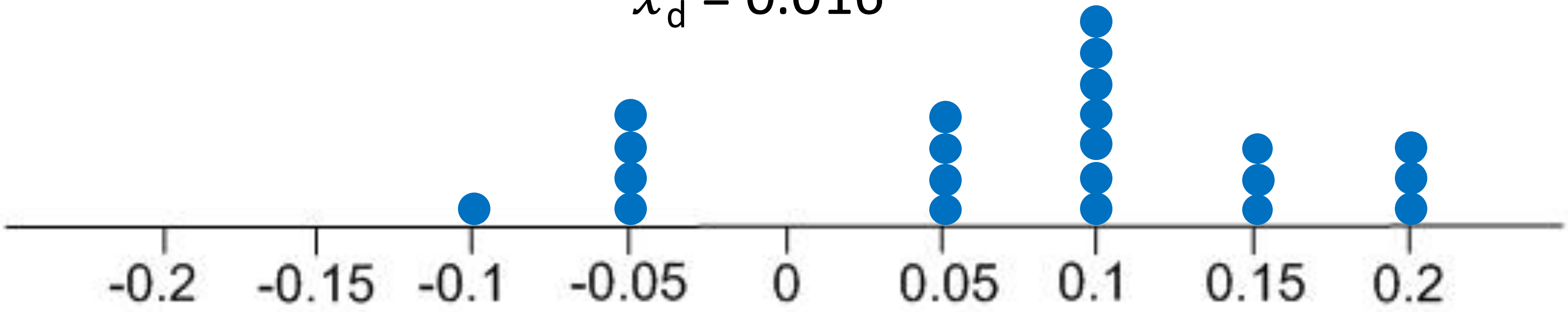
How can we use simulation-based methods to find an approximate p-value?

- 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	…

$\bar{x}_d = 0.016$


-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2