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

1. Simulations for paired data and rounding first base example continued.
2. Theory based approach for paired data, M&Ms example.
3. Multiple test, publication bias, and Reboxetine.
4. Two quantitative variables, scatterplots and correlation.

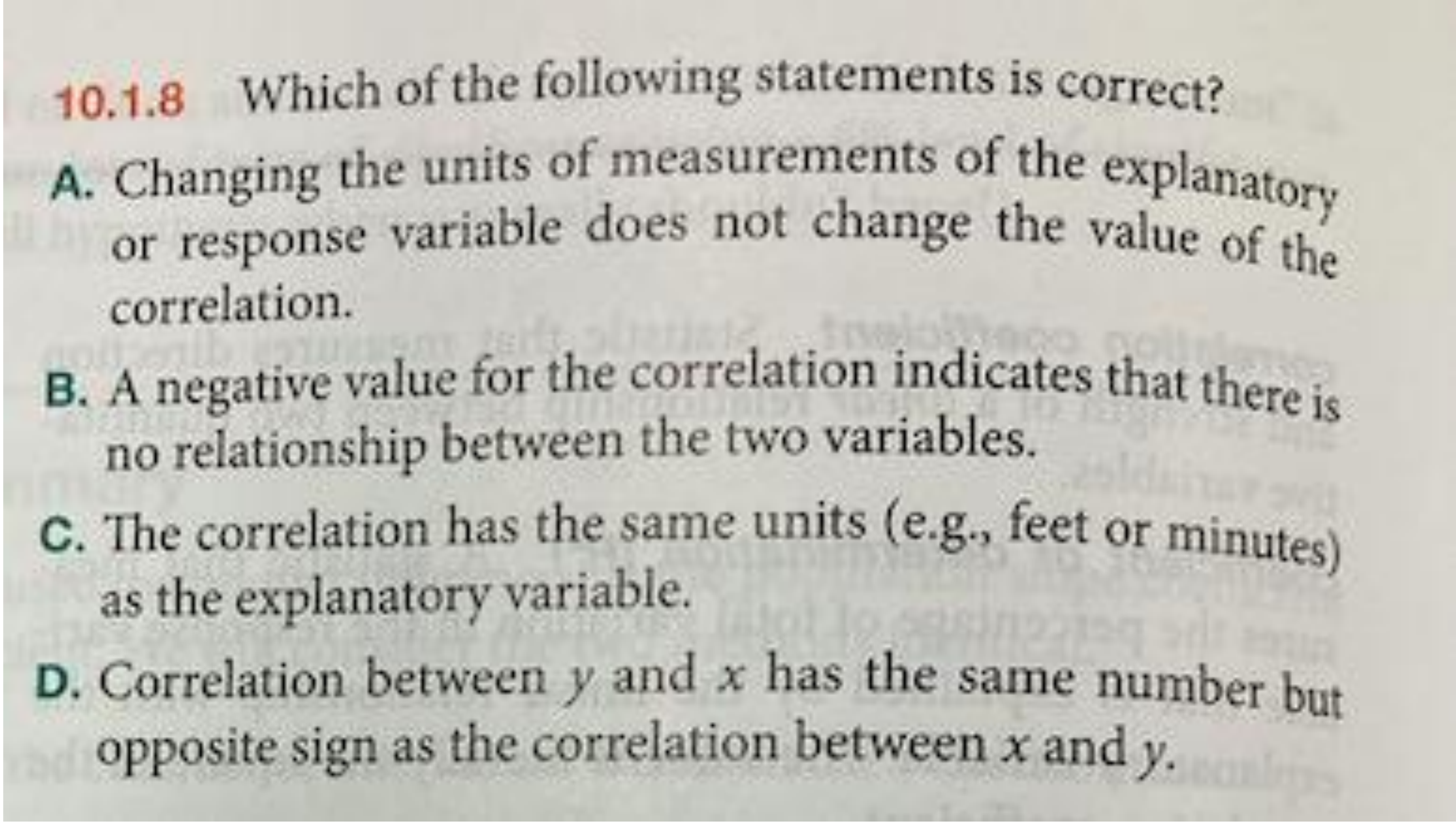
Read chapters 7 and 10.

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.

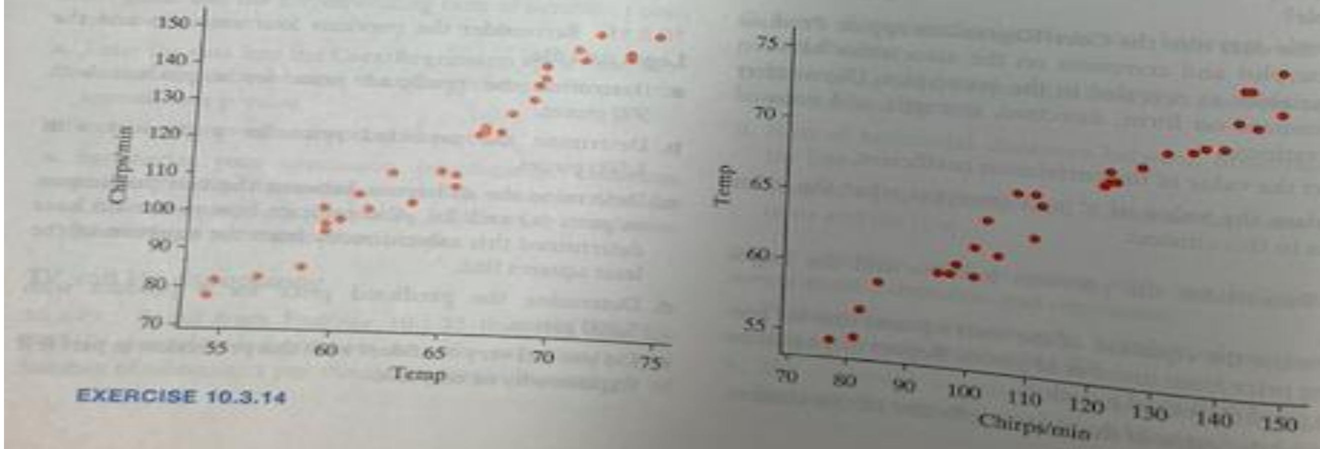
a. Describe the association between these variables.

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.

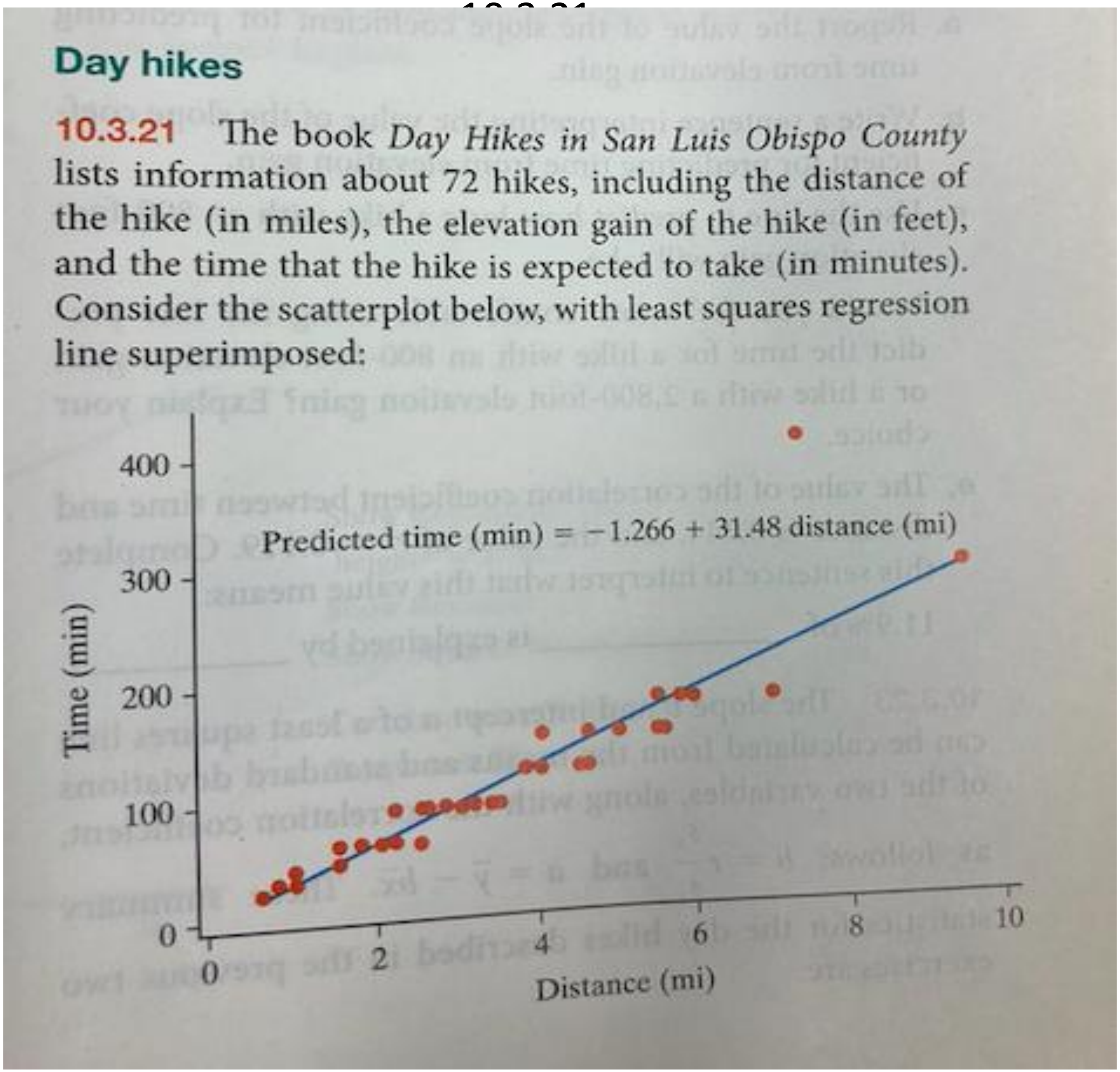
d. Interpret the value of the intercept coefficient. Is this a context in which the intercept coefficient is meaningful?

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

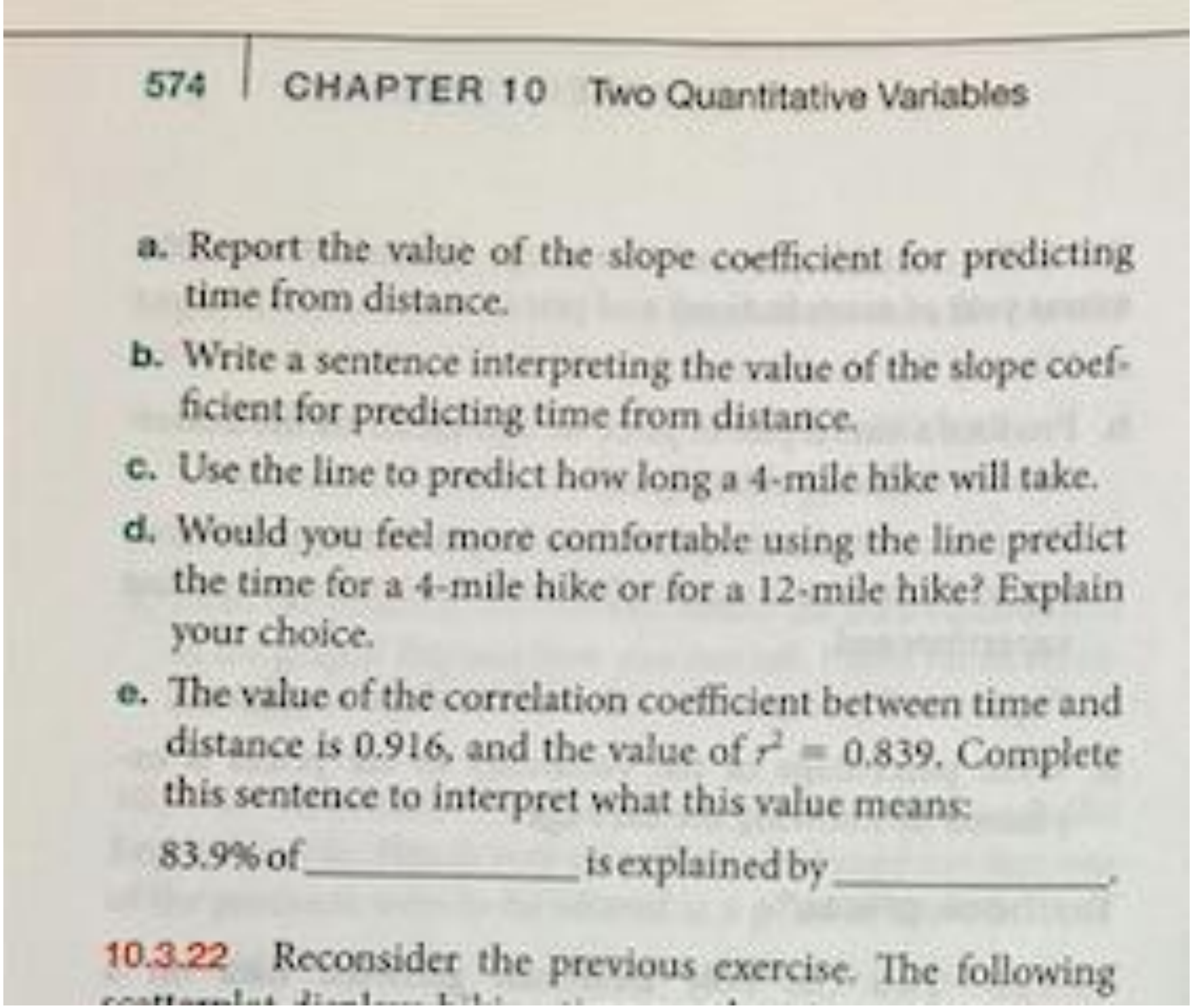


Day hikes

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:

Predicted time (min) = −1.266 + 31.48 distance (mi)
Time (min)
400
300
200
100
0
Distance (mi)
0
2
4
6
8
10

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.

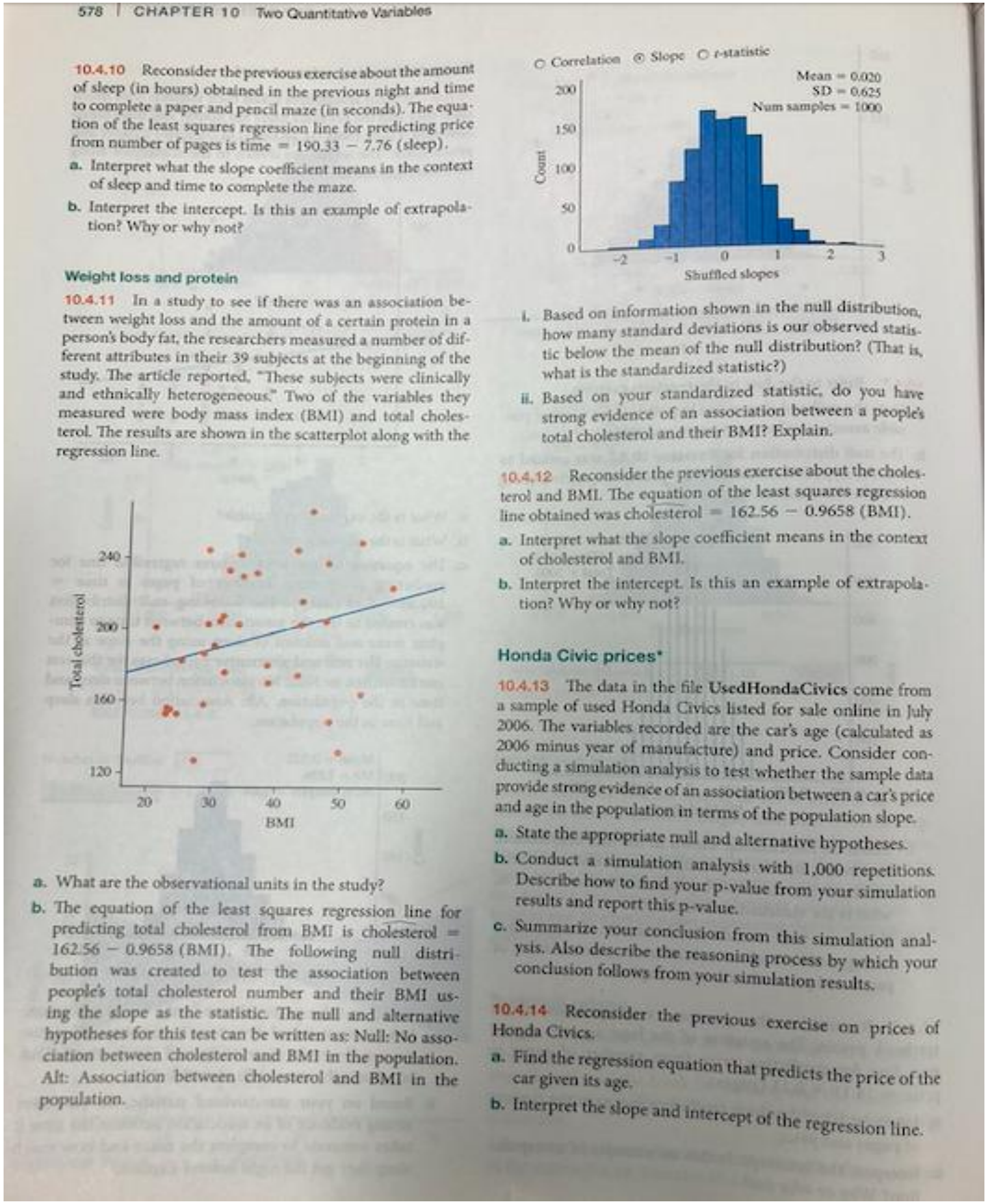
c. Use the line to predict how long a 4-mile hike will take.

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



578 | CHAPTER 10 Two Quantitative Variables

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 $\widehat{\text{time}} = 190.33 - 7.76\ (\text{sleep})$.

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 $\widehat{\text{cholesterol}} = 162.56 - 0.9658\ (\text{BMI})$. 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 $\widehat{\text{cholesterol}} = 162.56 - 0.9658\ (\text{BMI})$.

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.

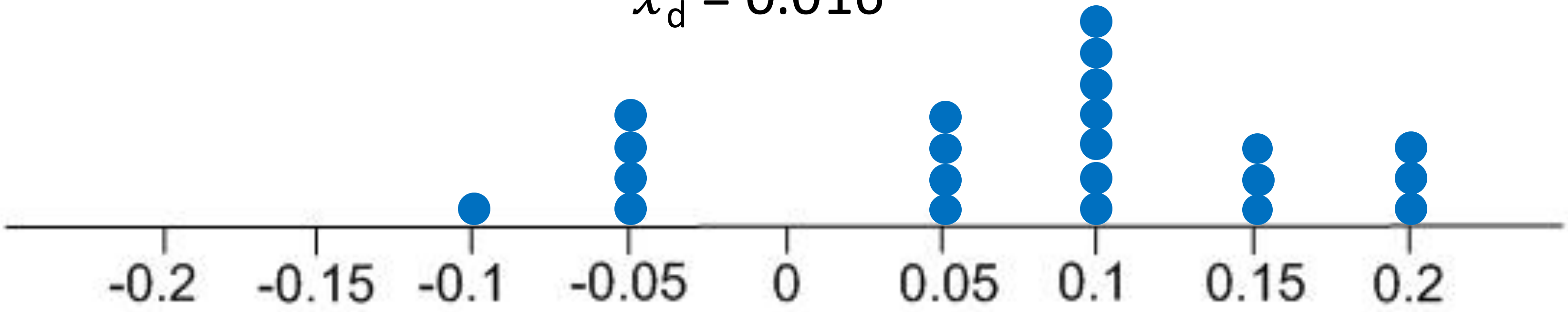
a. Find the regression equation that predicts the price of the car given its age.

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

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

-0.002 0.030 -0.011 -0.007

0.002

-0.002 -0.016 0.016 -0.007

-0.067 0.002 0.020 -0.007 -0.002

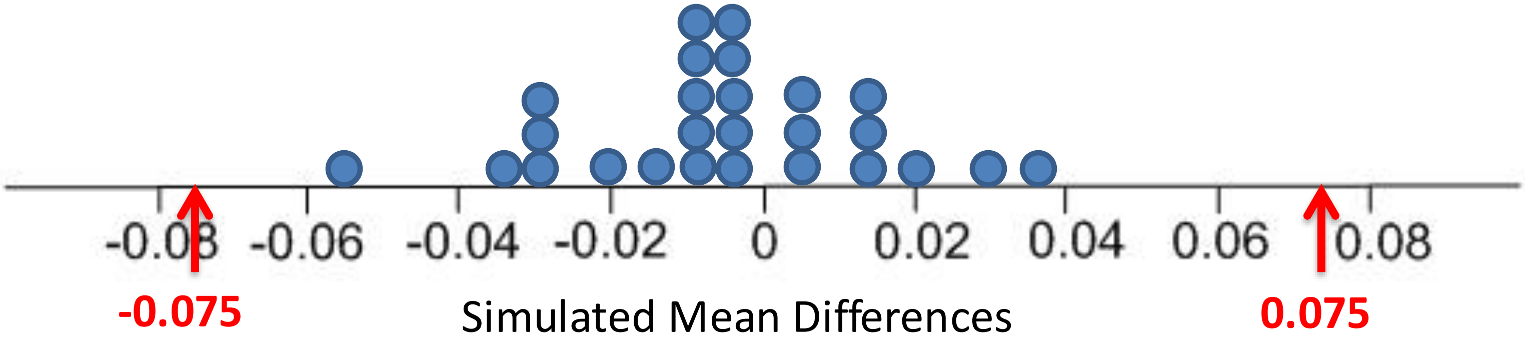
0.007 -0.034 0.467

-0.030 -0.016 0.002

0.020

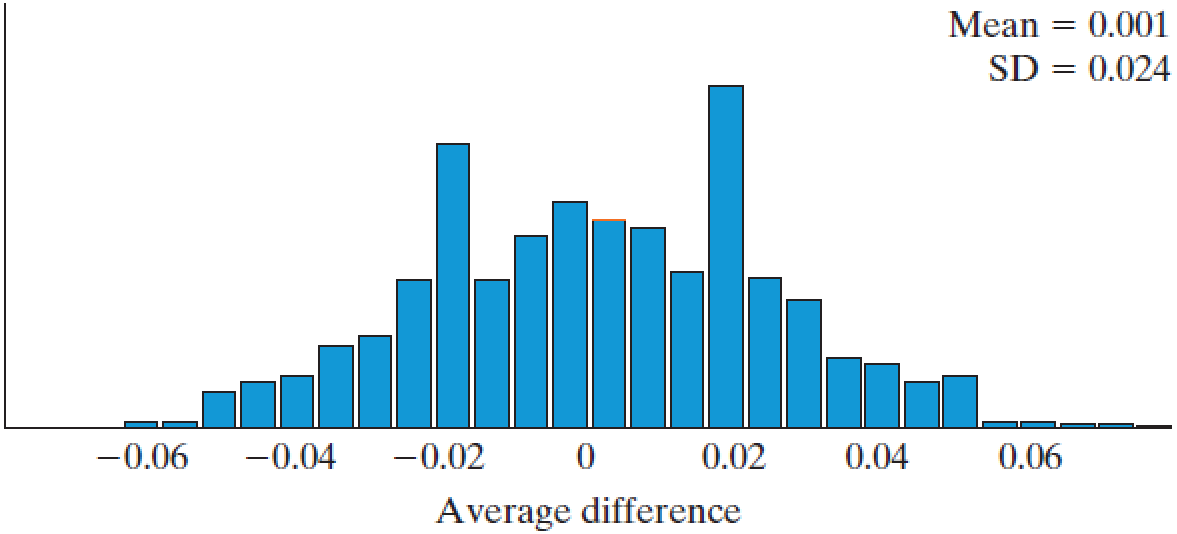
-0.002 -0.002 -0.025 0.066

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



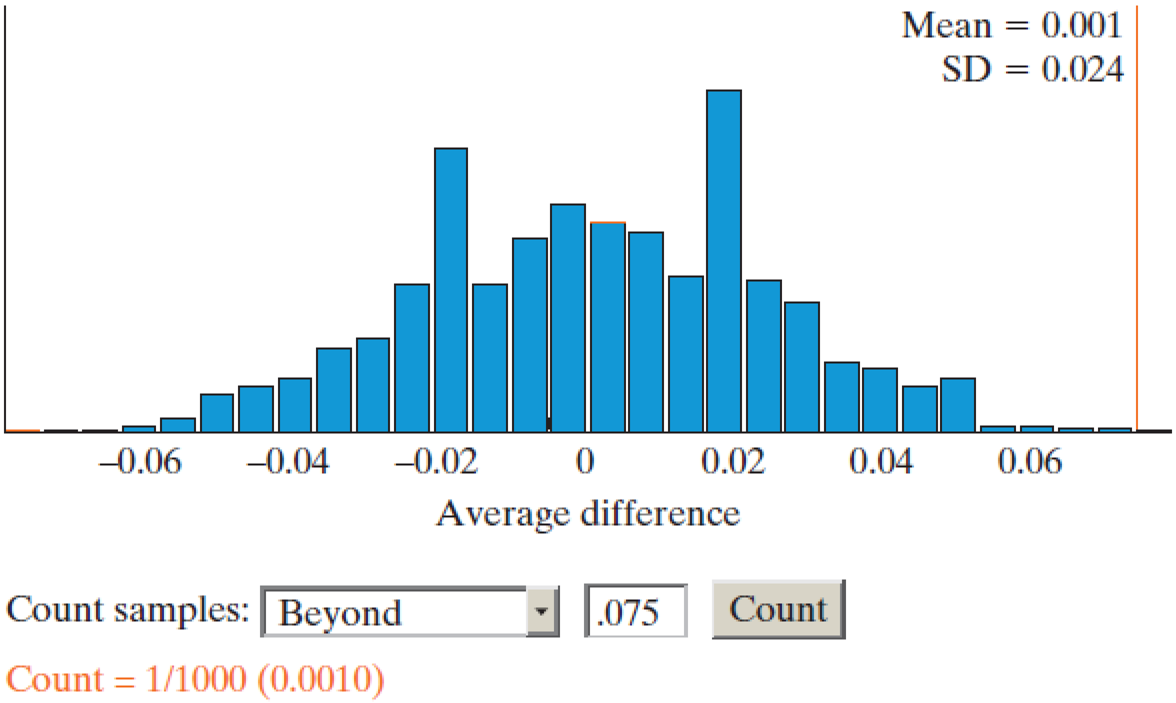
First Base

- 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

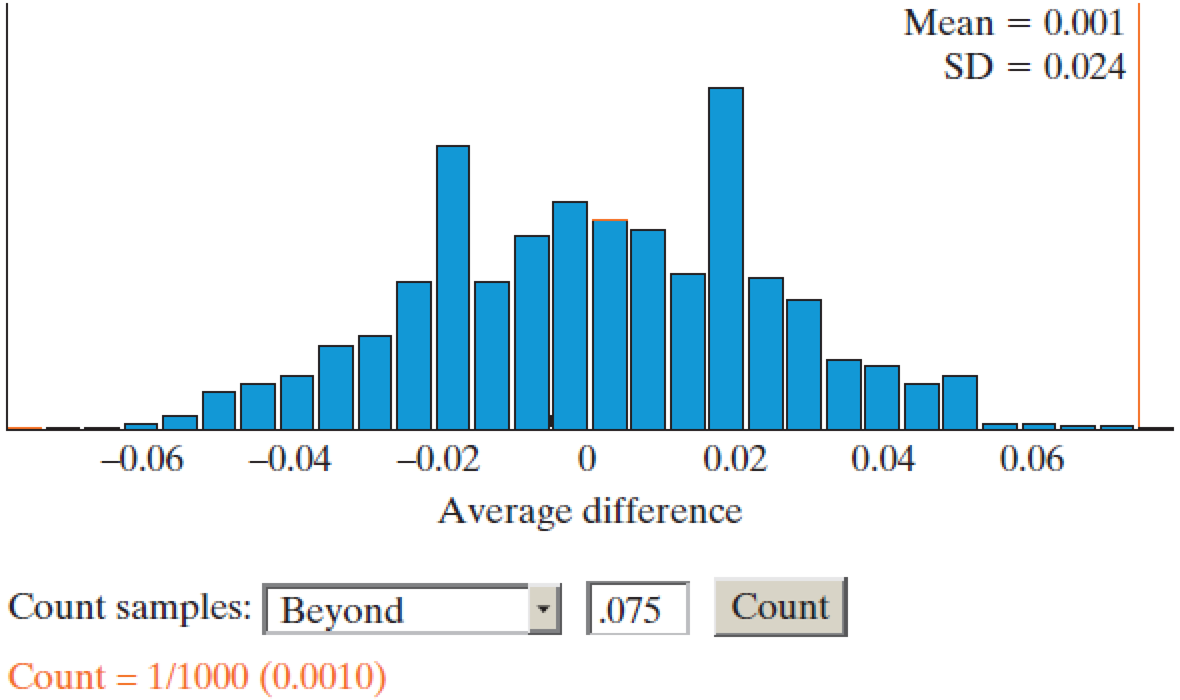


Count samples: Beyond .075 Count

Count = 1/1000 (0.0010)

First Base

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



Count samples: Beyond ▾ .075 Count

Count = 1/1000 (0.0010)

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

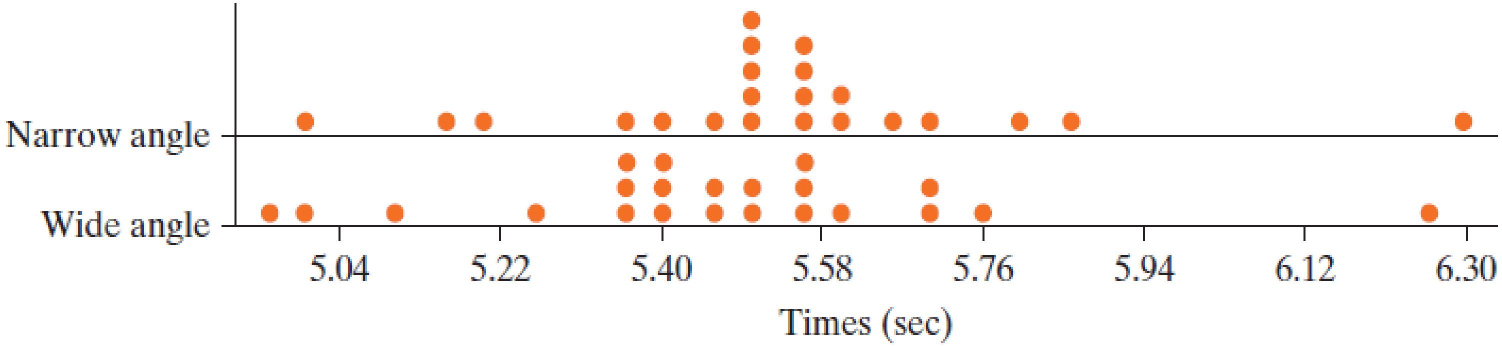
- Approximate a 95% confidence interval for μ_d:
  - 0.075 ± 1.96(0.024) seconds.
  - (0.028, 0.122) seconds.
- What does this mean?
  - We are 95% confident that, if we were to keep testing this indefinitely, the narrow angle route would take somewhere between 0.028 to 0.122 seconds longer on average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



of

d,

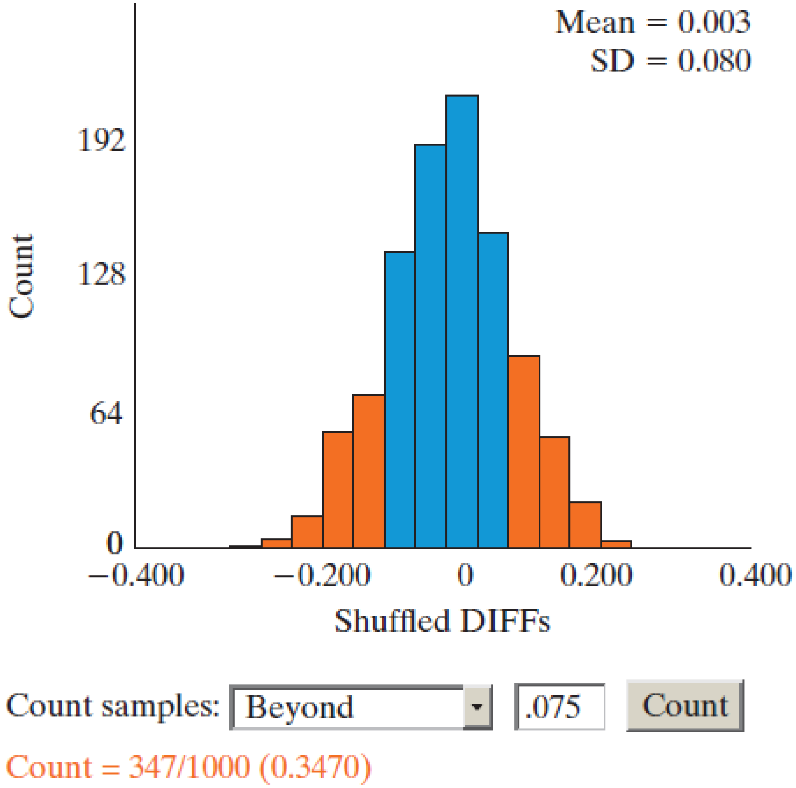
al

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?



Count samples: Beyond .075

Count = 347/1000 (0.3470)

2. 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.

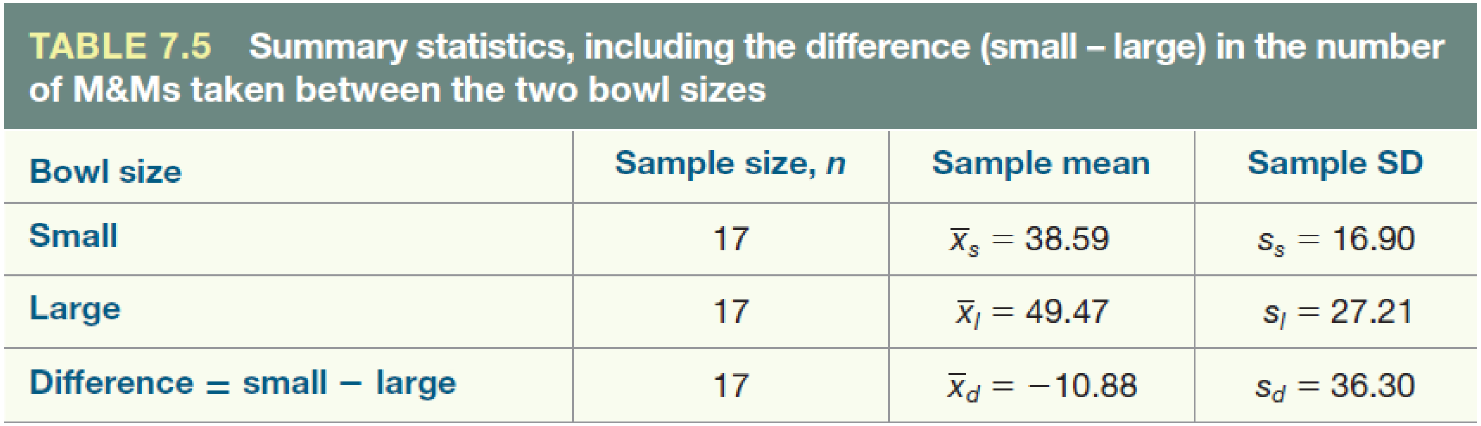
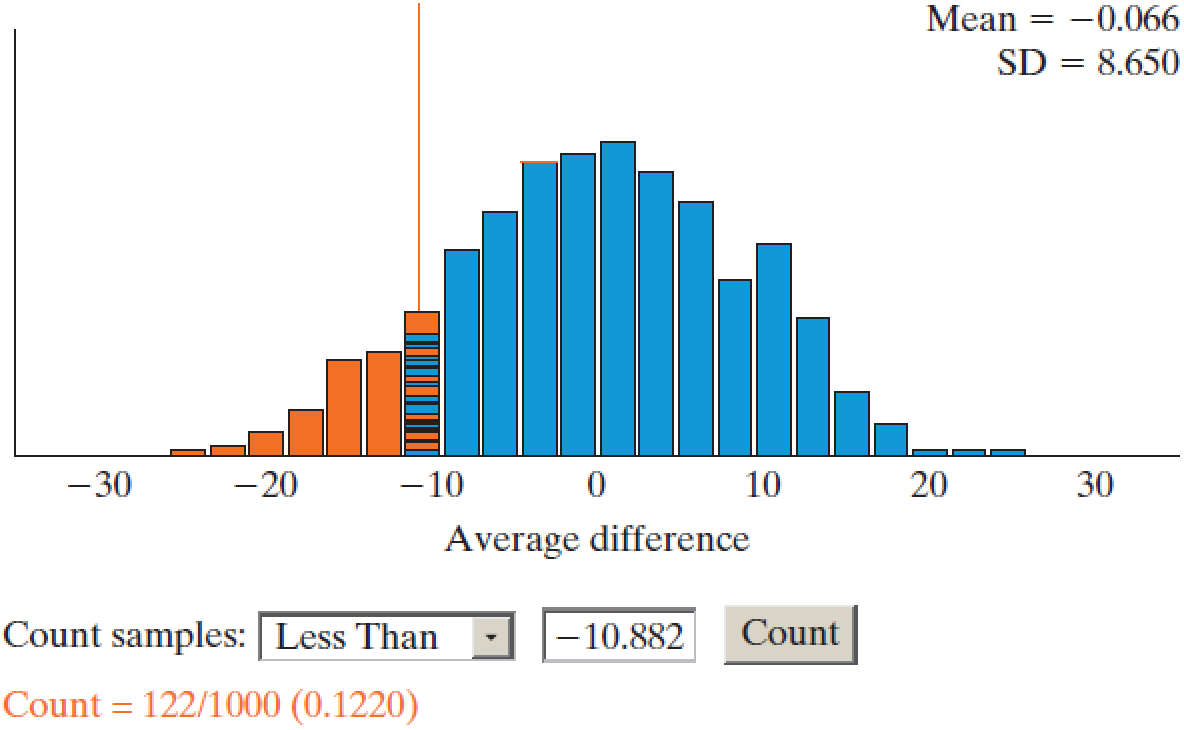


TABLE 7.5 Summary statistics, including the difference (small – large) in the number of M&Ms taken between the two bowl sizes

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.

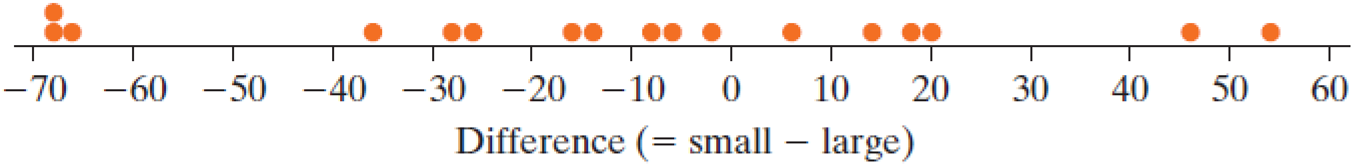


Count samples: Less Than −10.882

Count = 122/1000 (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.

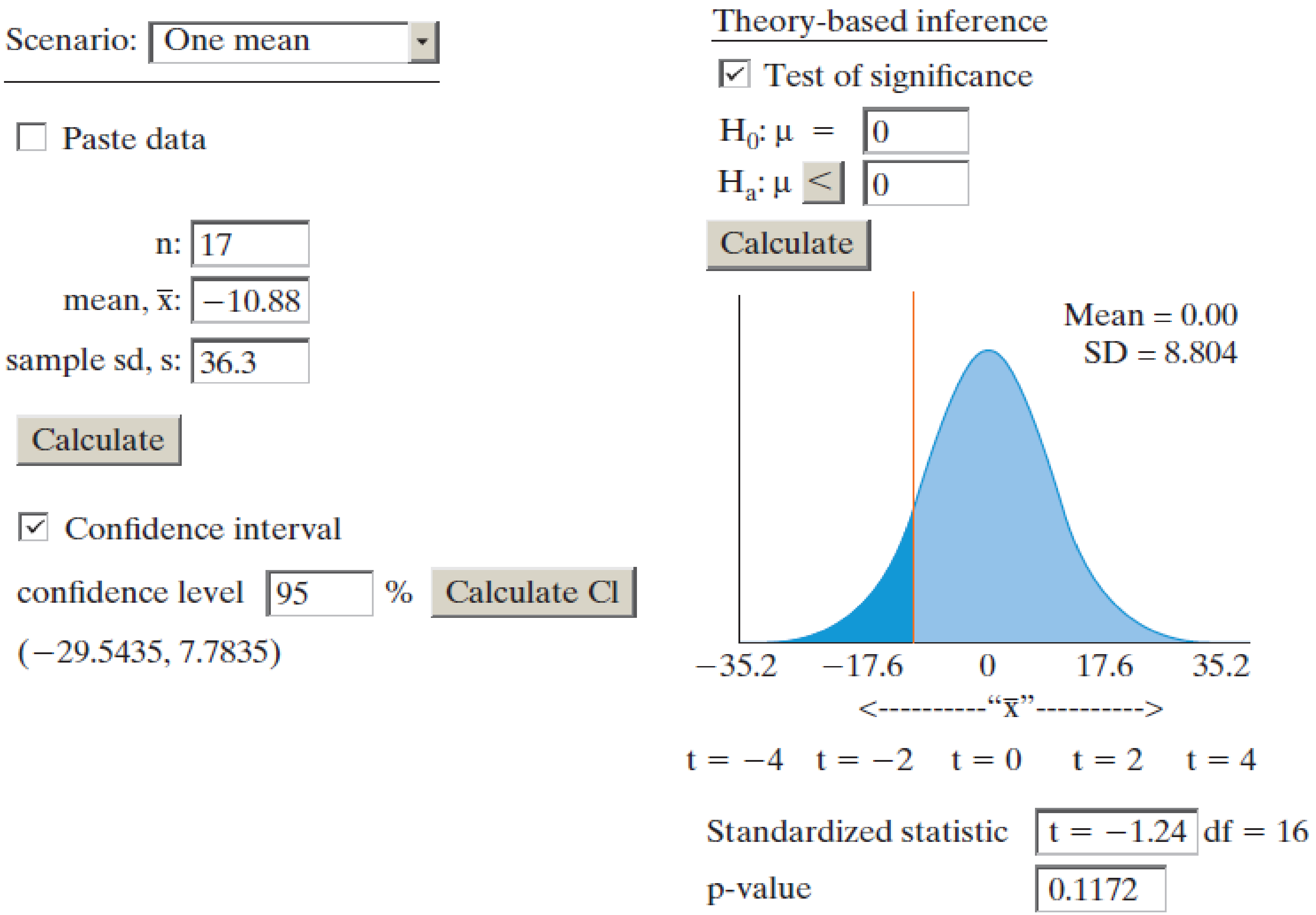


Theory-based test

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

- 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

☑ Confidence interval

confidence level 95 %

(−29.5435, 7.7835)

Theory-based inference

☑ Test of significance

H_0: μ = 0

H_a: μ < 0

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.

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.

3. Multiple testing and publication bias.

A p-value is the probability, assuming the null hypothesis of no relationship is true, that you will see a difference as extreme as, or more extreme than, you observed.

So, when you are looking at unrelated things, 5% of the time you will find a statistically significant relationship.

This underscores the need for followup confirmation studies. If testing many explanatory variables simultaneously, it can become very likely to find something significant even if nothing is actually related to the response variable.

Multiple testing and publication bias.

* For example, if the significance level is 5%, then for 100 tests where all null hypotheses are true, the expected number of incorrect rejections (Type I errors) is 5. If the tests are independent, the probability of at least one Type I error would be 99.4%. P(no Type I errors) = $.95^{100}$ = 0.6%.

* To address this problem, scientists sometimes change the significance level so that, under the null hypothesis that none of the explanatory variables is related to the response variable, the probability of rejecting at least one of them is 5%.

* One way is to use Bonferroni's correction: with *m* explanatory variables, use significance level 5%/m.

P(at least 1 Type I error) will be ≤ m (5%/m) = 5%.

Multiple testing and publication bias.

Imagine a scenario where a drug is tested many times to see if it reduces the incidence of some response variable. If the drug is tested 100 times by 100 different researchers, the results will be stat. sig. about 5 times.

If only the stat. sig. results are published, then the published record will be very misleading.

Multiple testing and publication bias.

A drug called Reboxetine made by Pfizer was approved as a treatment for depression in Europe and the UK in 2001, based on positive trials.

A meta-analysis in 2010 found that it was not only ineffective but also potentially harmful. The report found that 74% of the data on patients who took part in the trials of Reboxetine were not published because the findings were negative. Published data about reboxetine overestimated its benefits and underestimated its harm.

A subsequent 2011 analysis indicated Reboxetine might be effective for severe depression though.

4. Two quantitative variables, scatterplots and correlation.

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

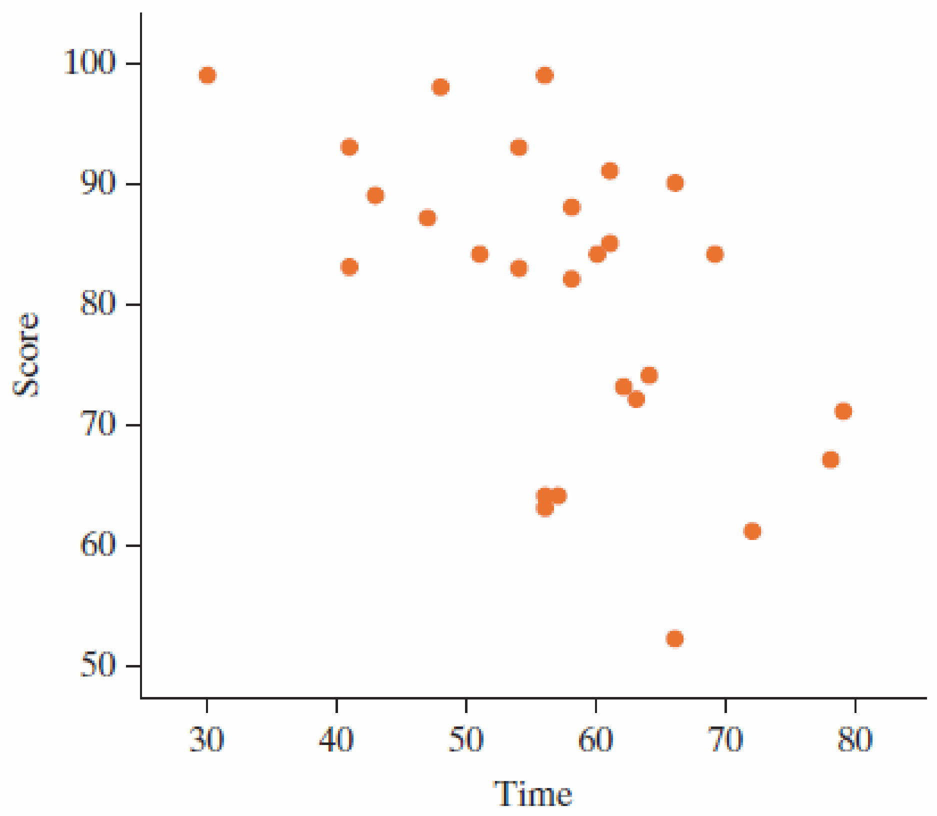
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

Time	30	41	41	43	47	48	51	54	54	56	56	56	57	58
Score	100	84	94	90	88	99	85	84	94	100	65	64	65	89
Time	58	60	61	61	62	63	64	66	66	69	72	78	79	
Score	83	85	86	92	74	73	75	53	91	85	62	68	72	

Scatterplot

Put explanatory variable on the horizontal axis.

Put response variable on the vertical axis.



Describing Scatterplots

- When we describe data in a scatterplot, we describe the
  - Direction (positive or negative)
  - Form (linear or not)
  - Strength (strong-moderate-weak, we will let correlation help us decide)
  - Unusual Observations
- How would you describe the time and test scatterplot?

Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
- Correlation is a number between -1 and 1.
- With positive correlation one variable increases, on average, as the other increases.
- With negative correlation one variable decreases, on average, as the other increases.
- The closer it is to either -1 or 1 the closer the points fit to a line.
- The correlation for the test data is -0.56.

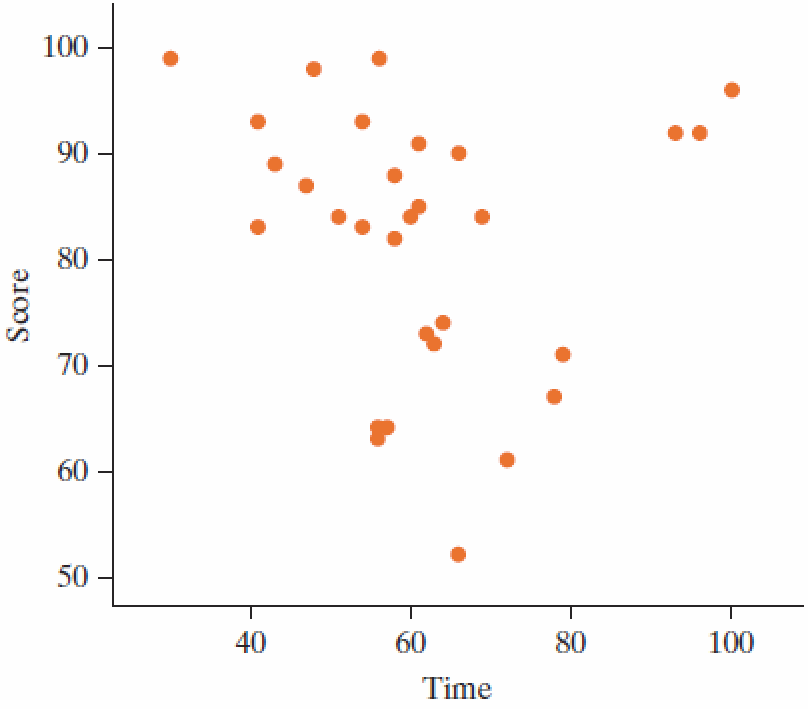
Correlation Guidelines

Correlation Value	Strength of Association	What this means
0.7 to 1.0	Strong	The points will appear to be nearly a straight line
0.3 to 0.7	Moderate	When looking at the graph the increasing/decreasing pattern will be clear, but there is considerable scatter.
0.1 to 0.3	Weak	With some effort you will be able to see a slightly increasing/decreasing pattern
0 to 0.1	None	No discernible increasing/decreasing pattern
Same Strength Results with Negative Correlations		

Back to the test data

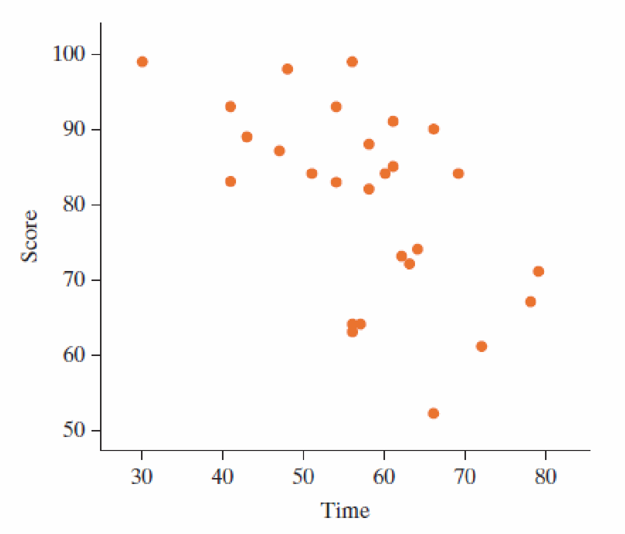
Actually the last three people to finish the test had scores of 93, 93, and 97.

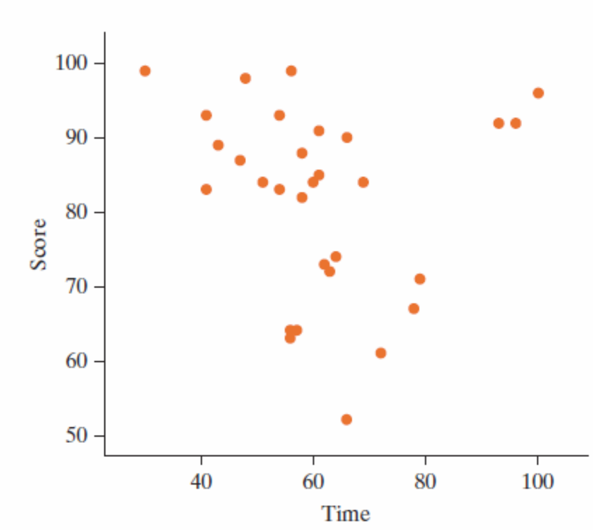
What does this do to the correlation?



Influential Observations

- The correlation changed from -0.56 (a fairly moderate negative correlation) to -0.12 (a weak negative correlation).
- Points that are far to the left or right and not in the overall direction of the scatterplot can greatly change the correlation. (influential observations)





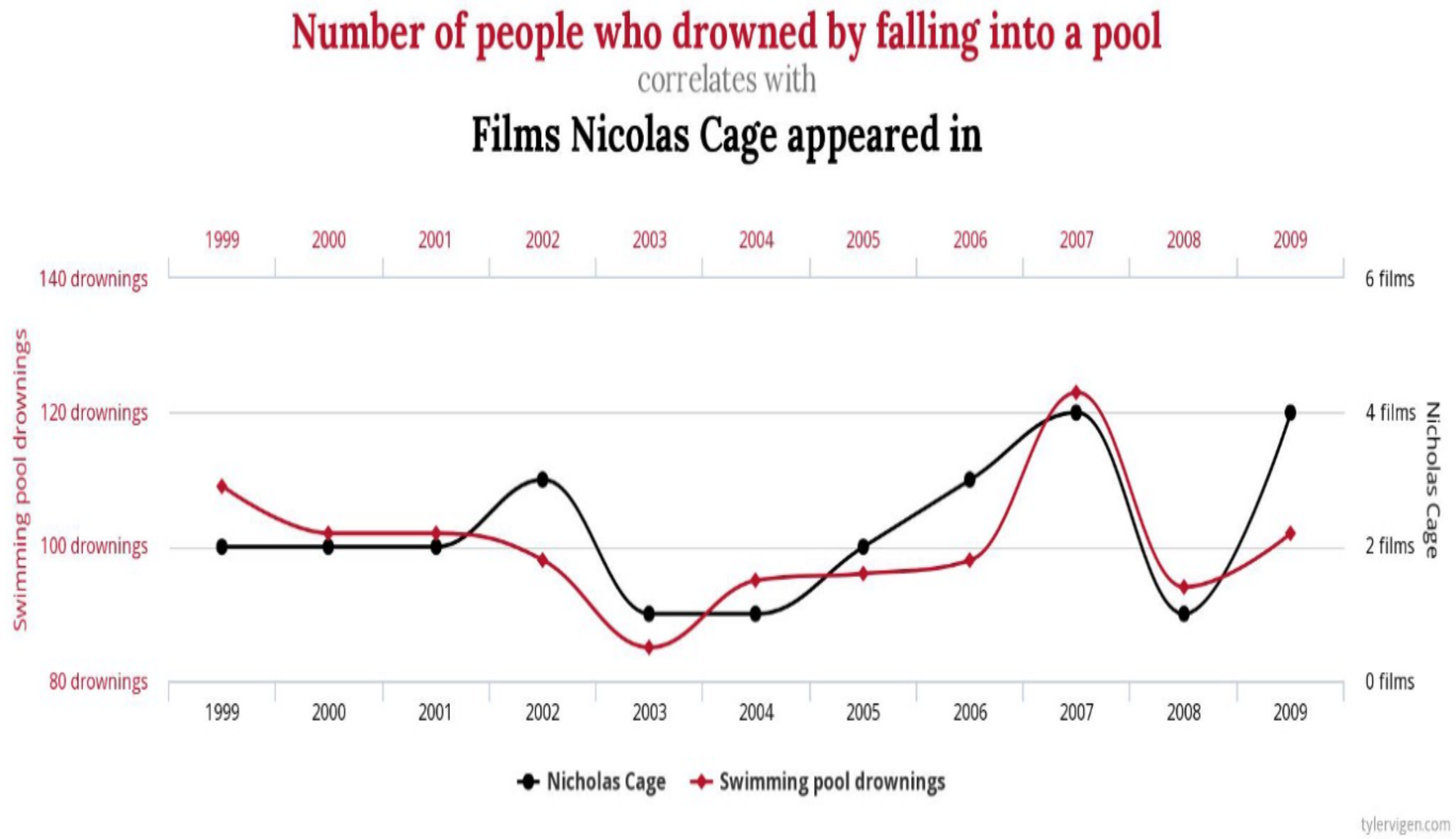
Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
  - $-1 \leq r \leq 1$
  - Correlation makes no distinction between explanatory and response variables.
  - Correlation has no units.
  - Correlation is not resistant to outliers. It is sensitive.

Learning Objectives for Section 10.1

- Summarize the characteristics of a scatterplot by describing its direction, form, strength and whether there are any unusual observations.
- Recognize that the correlation coefficient is appropriate only for summarizing the strength and direction of a scatterplot that has linear form.
- Recognize that a scatterplot is the appropriate graph for displaying the relationship between two quantitative variables and create a scatterplot from raw data.
- Recognize that a correlation coefficient of 0 means there is no linear association between the two variables and that a correlation coefficient of -1 or 1 means that the scatterplot is exactly a straight line.
- Understand that the correlation coefficient is influenced by extreme observations.

Note that correlation ≠ causation.



Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

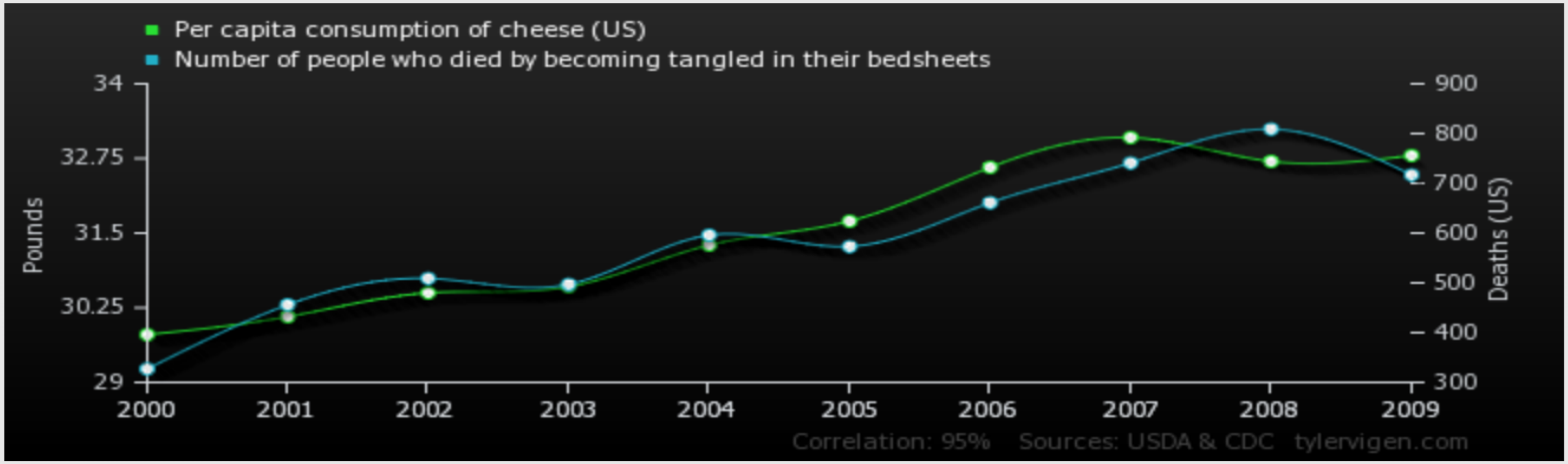
from: http://tylervigen.com

Note that correlation ≠ causation.

Per capita consumption of cheese (US)

correlates with

Number of people who died by becoming tangled in their bedsheets

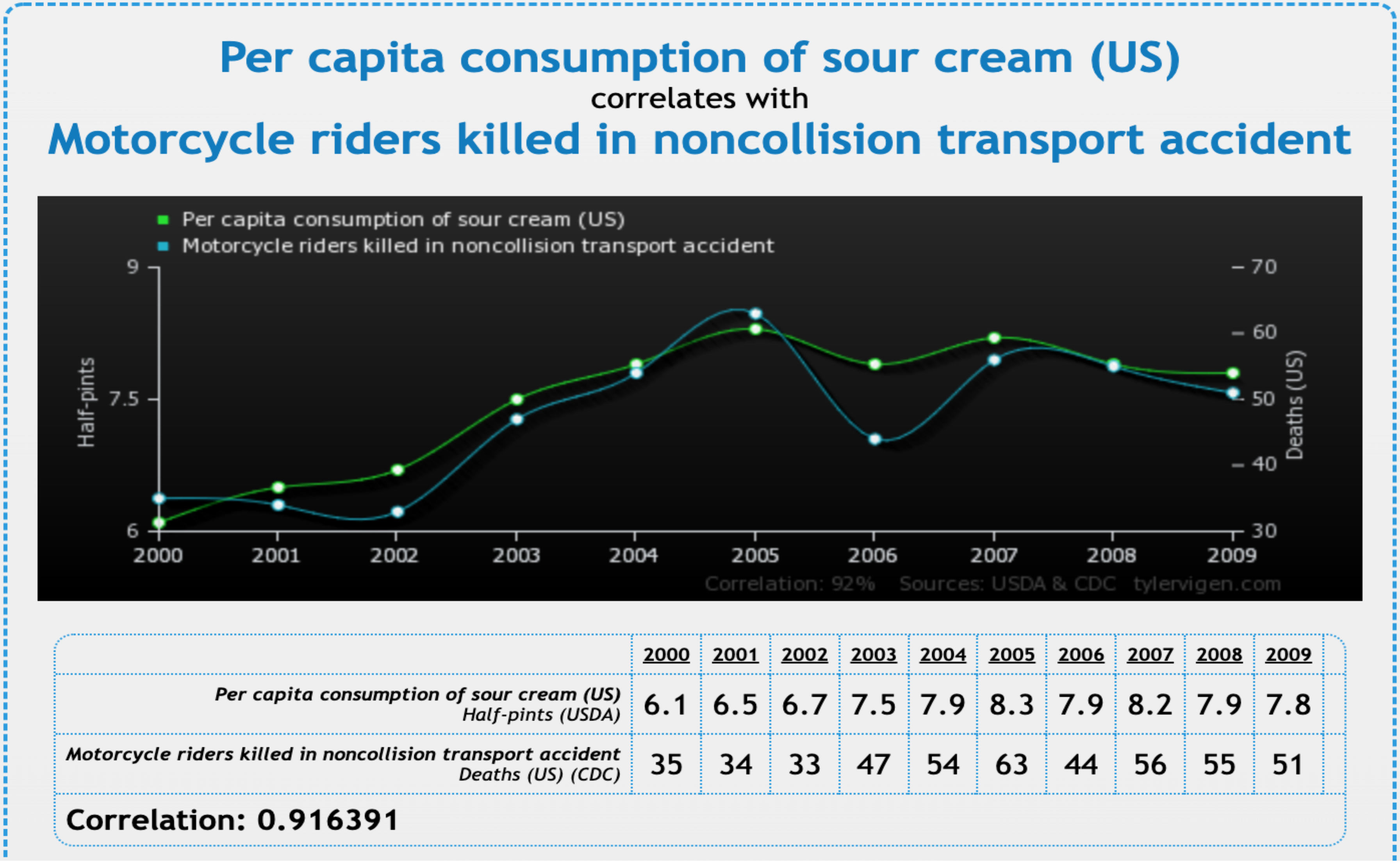


Correlation: 95%   Sources: USDA & CDC   tylervigen.com

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

Note that correlation ≠ causation.



Per capita consumption of sour cream (US)

correlates with

Motorcycle riders killed in noncollision transport accident

🟩 Per capita consumption of sour cream (US)
🟦 Motorcycle riders killed in noncollision transport accident
Half-pints
9
7.5
6
Deaths (US)
70
60
50
40
30
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009

Correlation: 92% Sources: USDA & CDC tylervigen.com

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of sour cream (US) *Half-pints (USDA)*	6.1	6.5	6.7	7.5	7.9	8.3	7.9	8.2	7.9	7.8
Motorcycle riders killed in noncollision transport accident *Deaths (US) (CDC)*	35	34	33	47	54	63	44	56	55	51

Correlation: 0.916391