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

0. Midterms and hw.
1. Paired data and studying with music example.
2. Simulation approach with paired data and baseball example.
3. Theory based approach for paired data and M&M example.
4. Multiple testing and publication bias.
5. Two variables and correlation.
6. Linear regression.

Read ch7 and 10.
Hw4 is due Tue and is 10.1.8, 10.3.14, 10.3.21, 10.4.11.
http://www.stat.ucla.edu/~frederic/13/sum17 .

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 scores are listed on the course website in 13midscores.xlsx.
They are out of 20.
The mean was 15.1 = 75.5%. Median = 75%.
SD = 16.5%.
The grading is the standard scale, i.e. 90-100 = A range, 80-89.9 = B range, etc.
I do reward improvement on the final.

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.

1. 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.

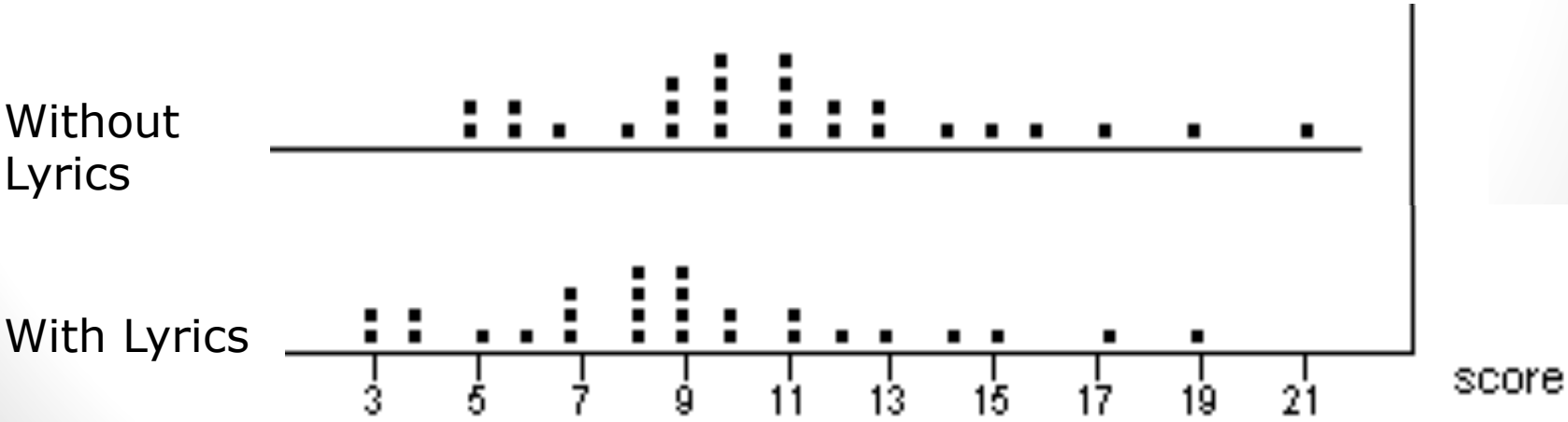
Experiment C — Paired design using matching

- 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.
- Can we make cause-and-effect conclusions in paired design?
- Should we still have random assignment?

Pairing and Random Assignment

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.

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

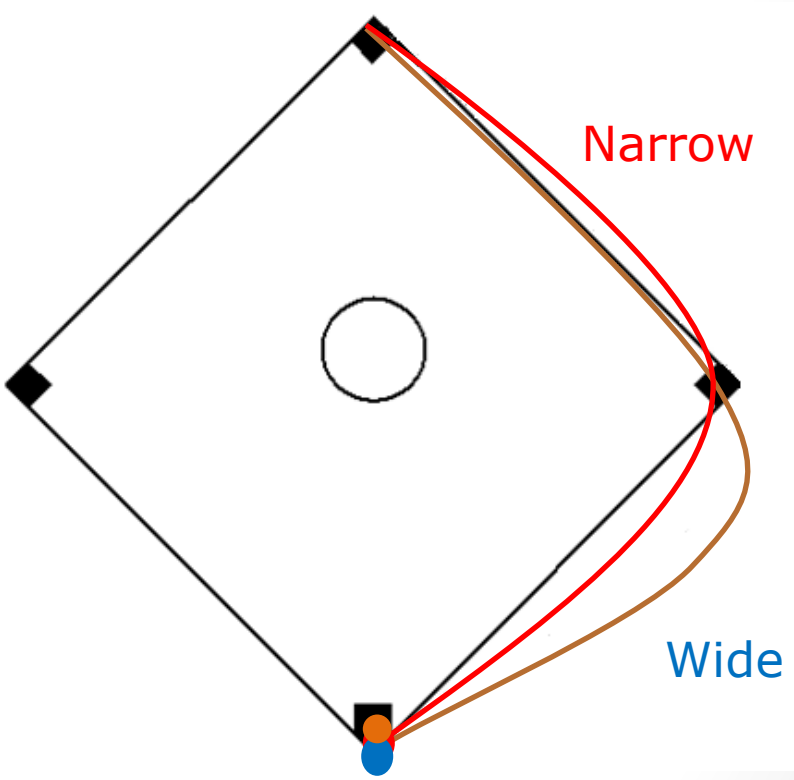
Section 7.2

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.**

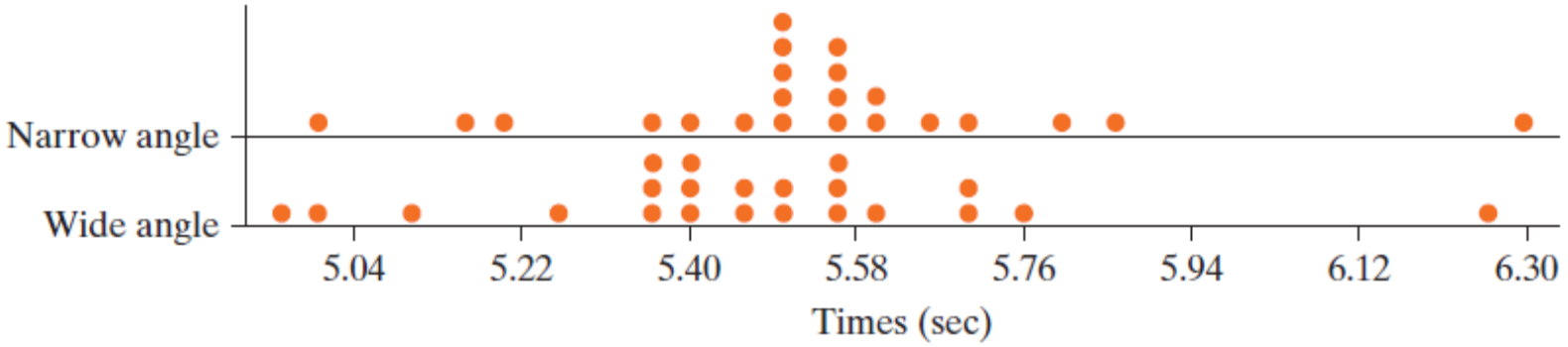
First Base

- 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

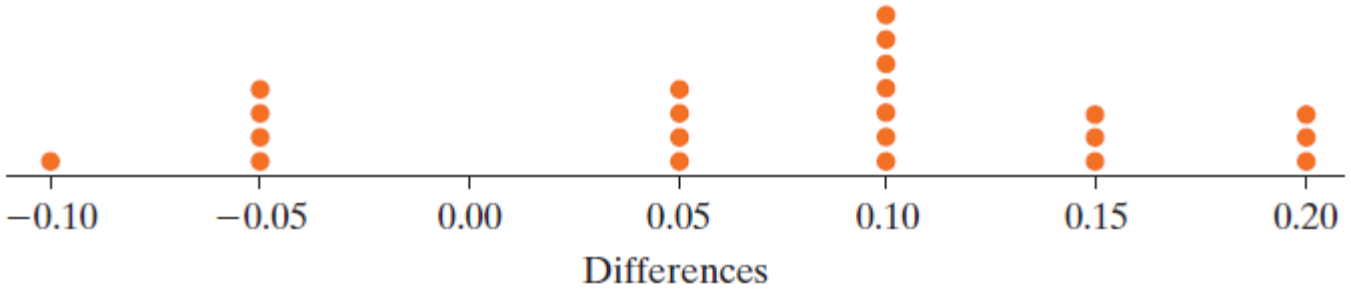
- It is difficult to detect a difference between the methods when these is so much variation.
-

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	…

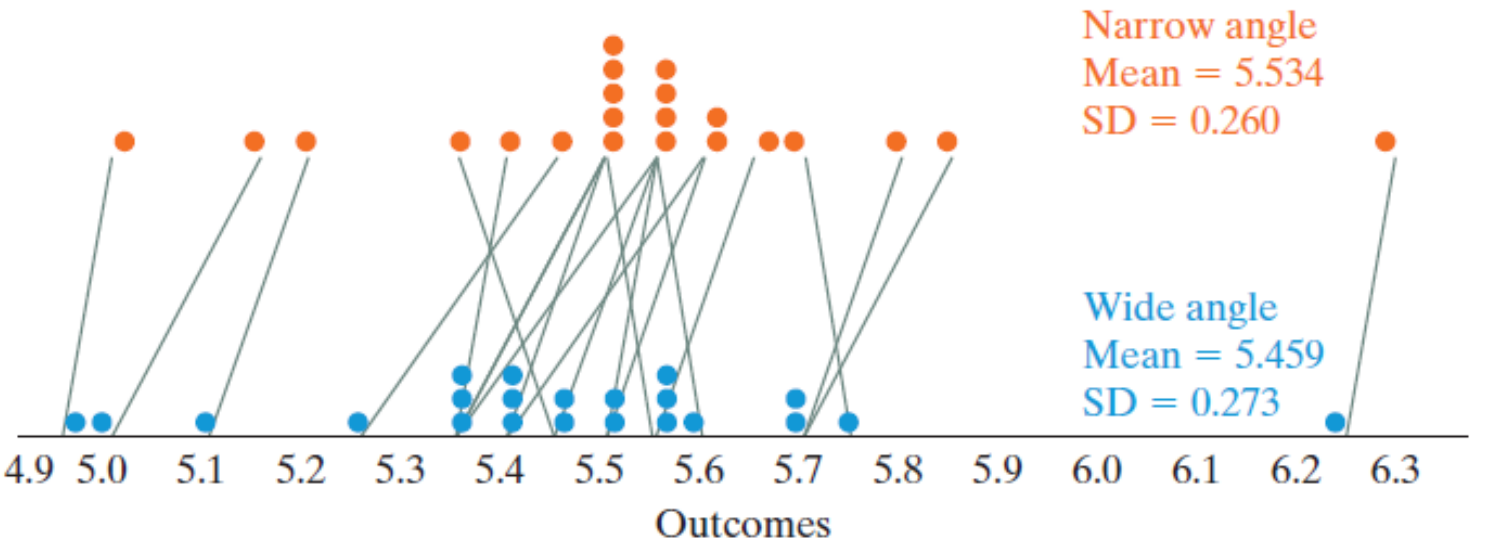


The Differences in Times

- 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

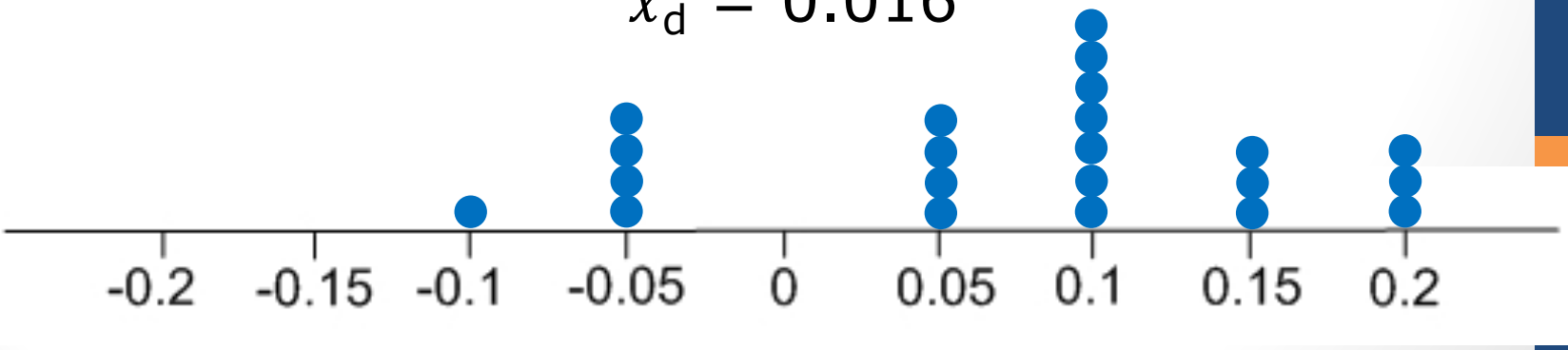
How can we use simulation-based methods to find an approximate p-value?

- 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

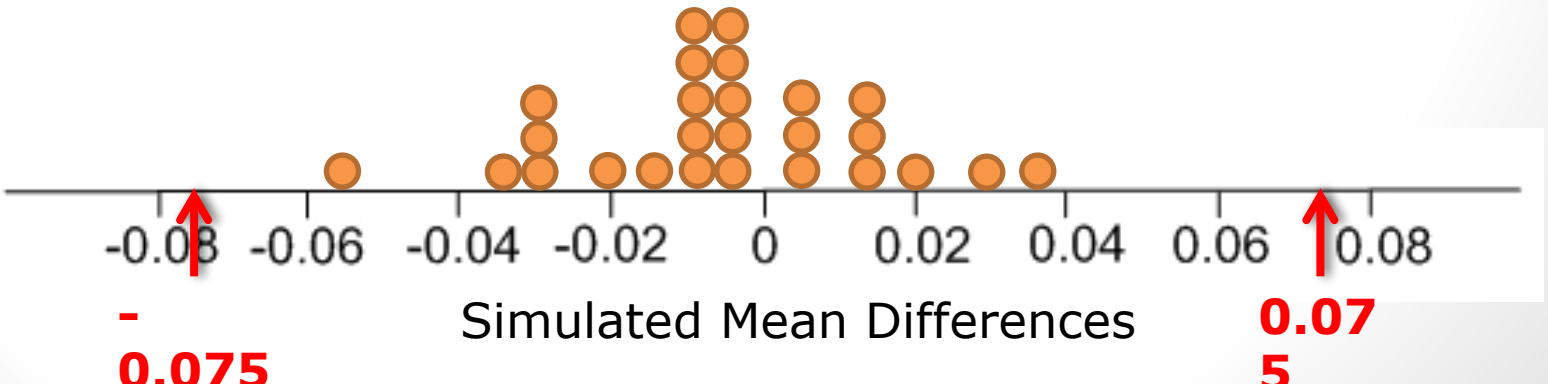
Subject	1	2	3	4	5	6	7	8	9	10	
narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
diff	0.05	-0.05	-0.10	0.10	0.15	0.05	0.05	0.15	0.15	-0.10	…

$\bar{x}_d = 0.016$



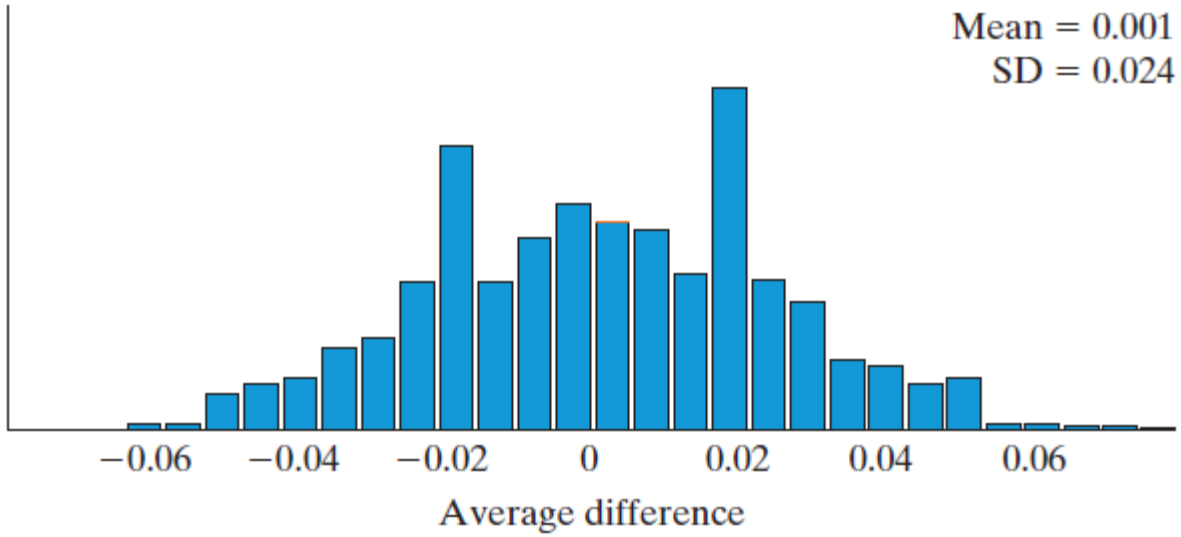
More Simulations

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



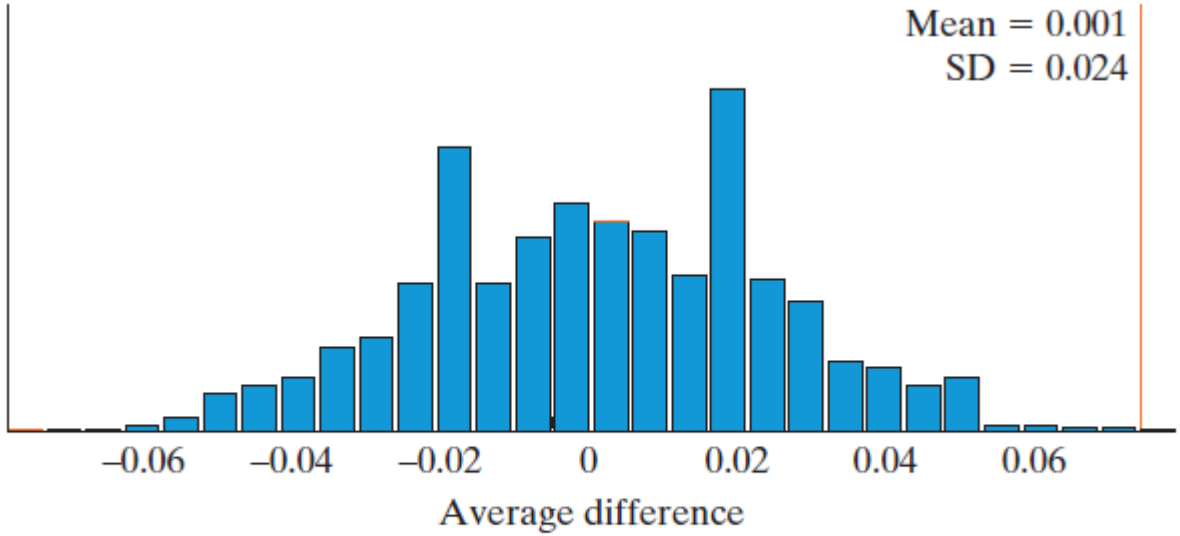
First Base

- 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 = SE.
  SD of time differences was 0.0883. SD of mean time diff.s = .024.
- Where is our observed statistic of 0.075?



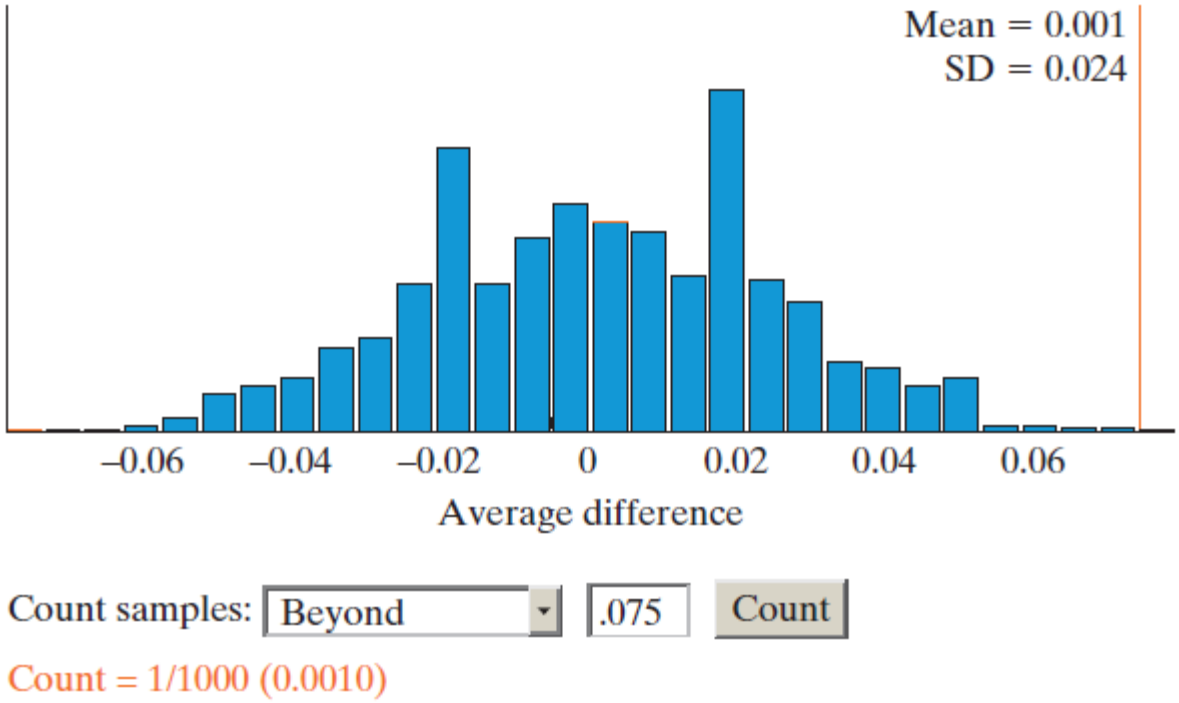
First Base

- 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.

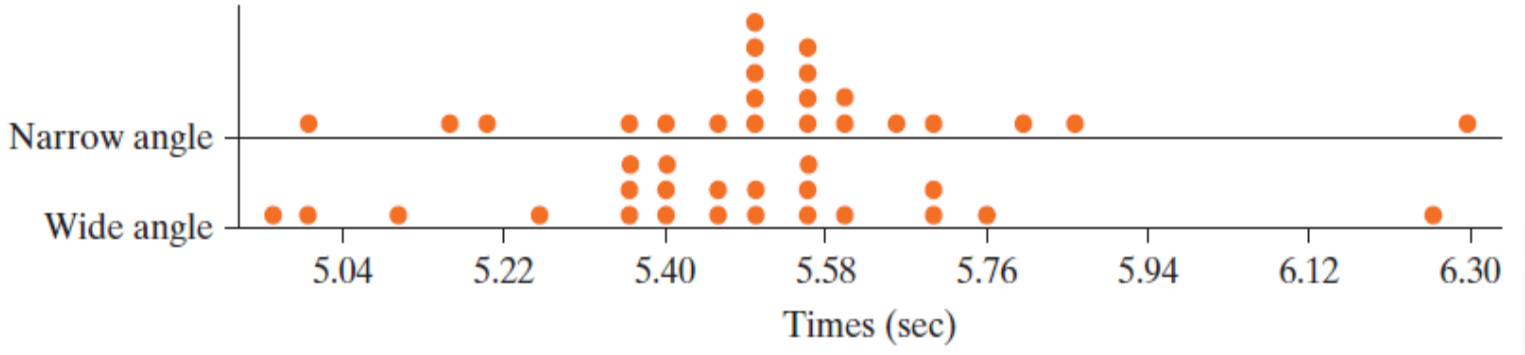
First Base

- 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.

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.

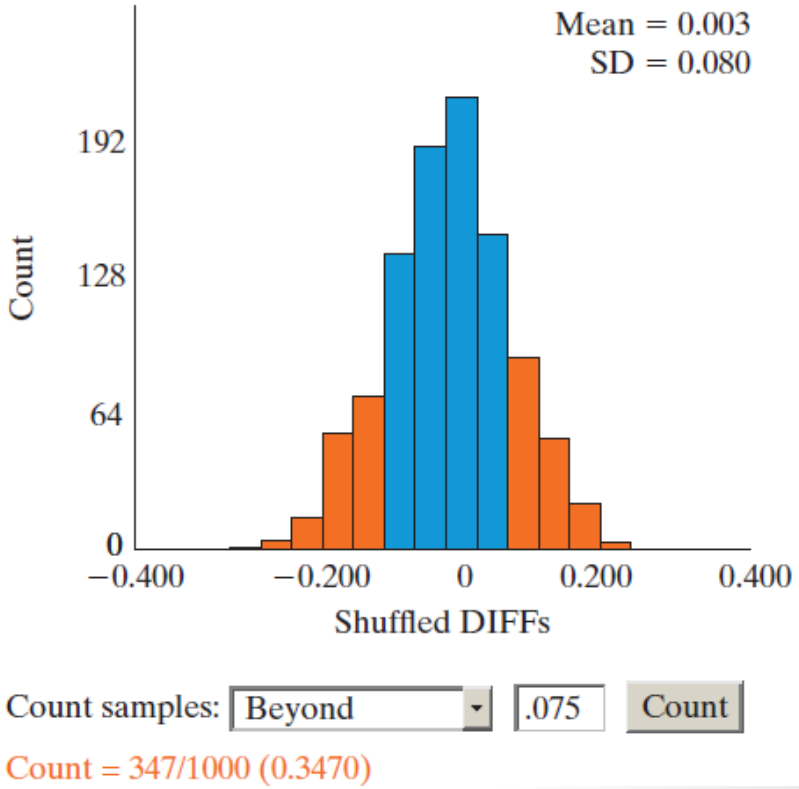


First Base

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?



3. 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:

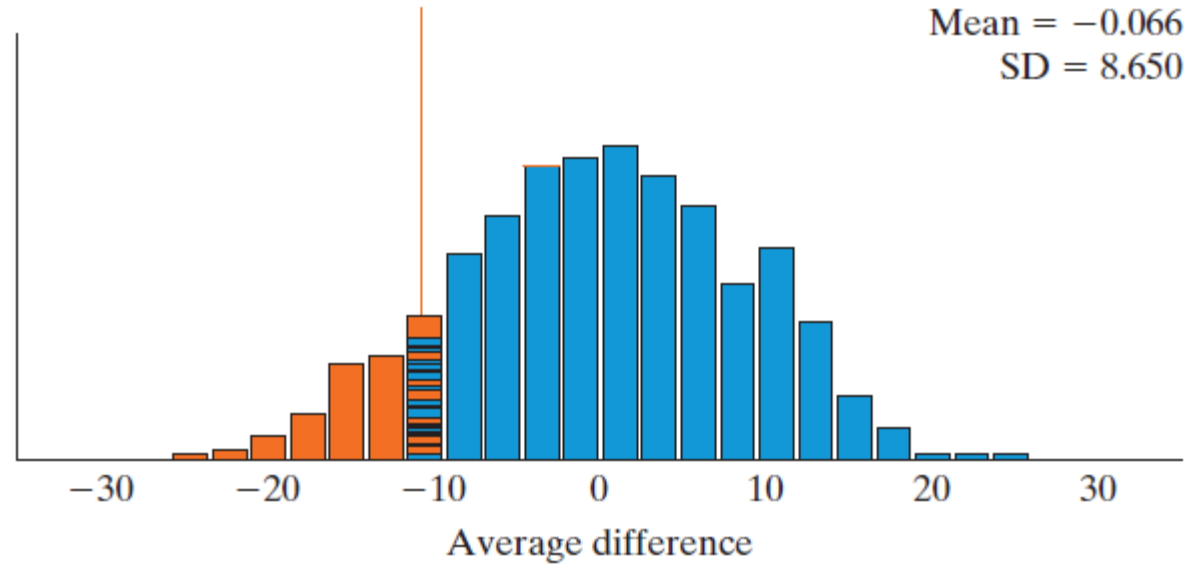
- H_0: $\mu_d = 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is 0.
- H_a: $\mu_d < 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is less than 0.

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	$\bar{x}_d = -10.88$	$s_d = 36.30$

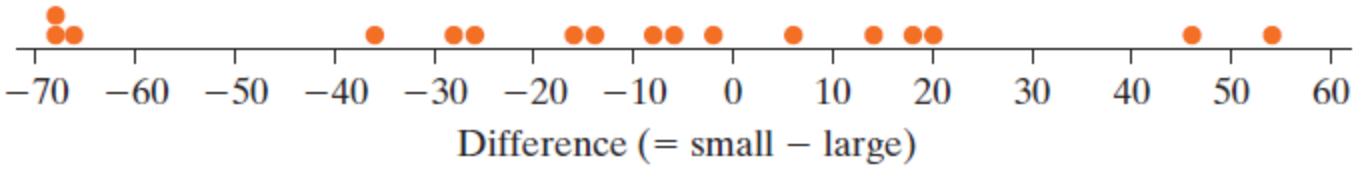
How Many M&Ms Would You Like?

- 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.
- This can all be predicted (along with the variability) using theory-based methods.
- Theory-based methods should be valid if the population distribution of differences is symmetric (we can guess at this by looking at the sample distribution of differences) or our sample size is at least 20.
- Our sample size was only 17, but this distribution of differences is fairly symmetric, so we will proceed with a theory-based test.



Theory-based test

- We can do theory-based methods with the applet we used last time or the theory-based applet.
- With the applet we used last time, we need to calculate the t-statistic:

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

- With the theory-based applet, we just need to enter the summary statistics and use a **test for a one mean.**
- 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)

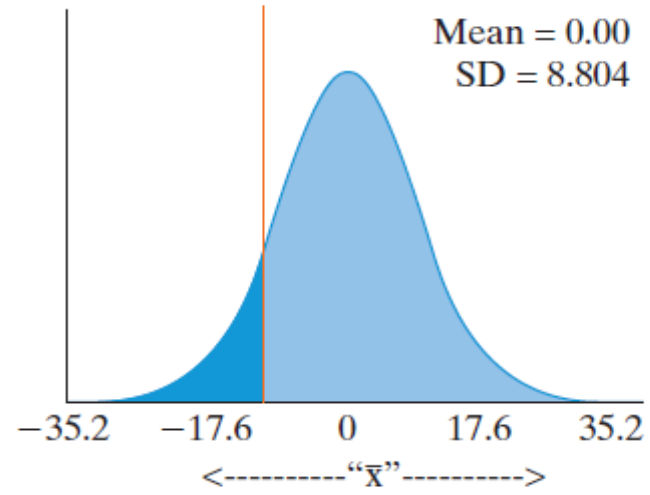
Theory-based inference

☑ 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 model gives slightly different results than simulation, 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.
  - Other ideas?

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.

4. 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, 5% of the time you are looking at unrelated things, 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.

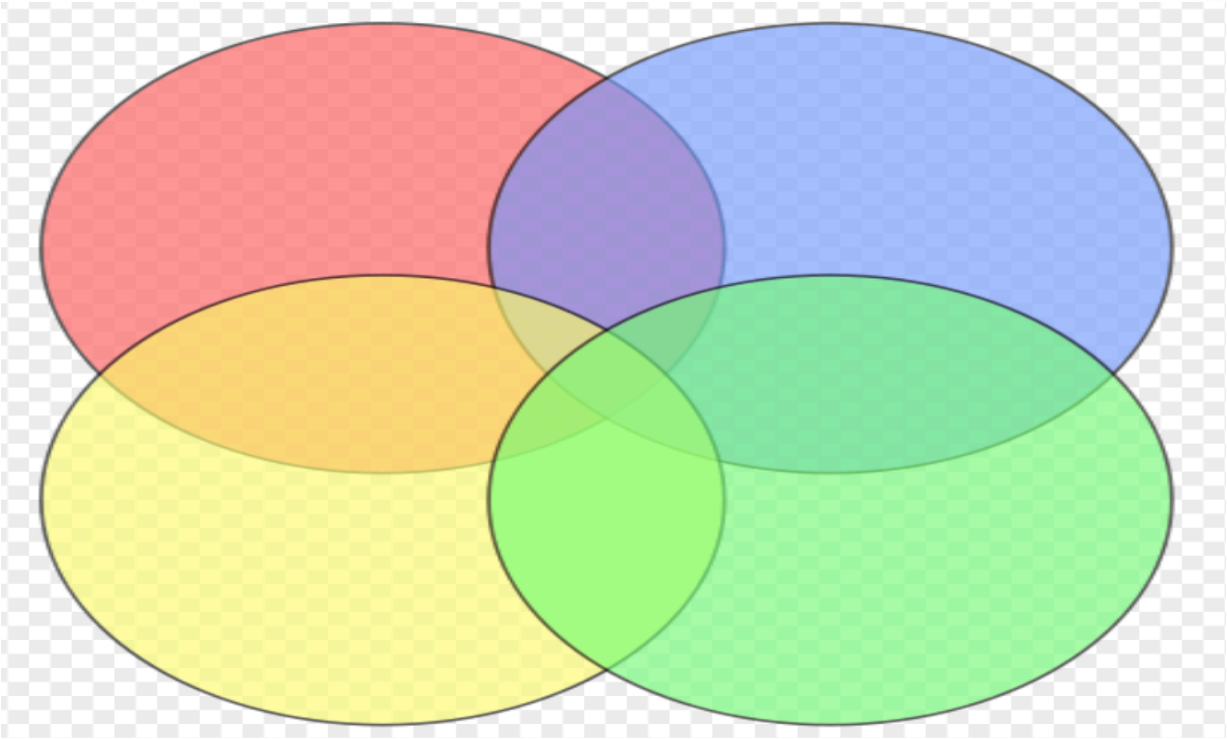
4. Multiple testing and publication bias.

* 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%.

* 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 *any* 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%.



P(Type I error on explanatory 1) = 5%/m.

P(Type I error on explanatory 2) = 5%/m.

P(Type 1 error on at least one explanatory) ≤

P(error on 1) + P(error on 2) + ... + P(error on m) = m x 5%/m.

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 testes 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.

5. Two quantitative variables.

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

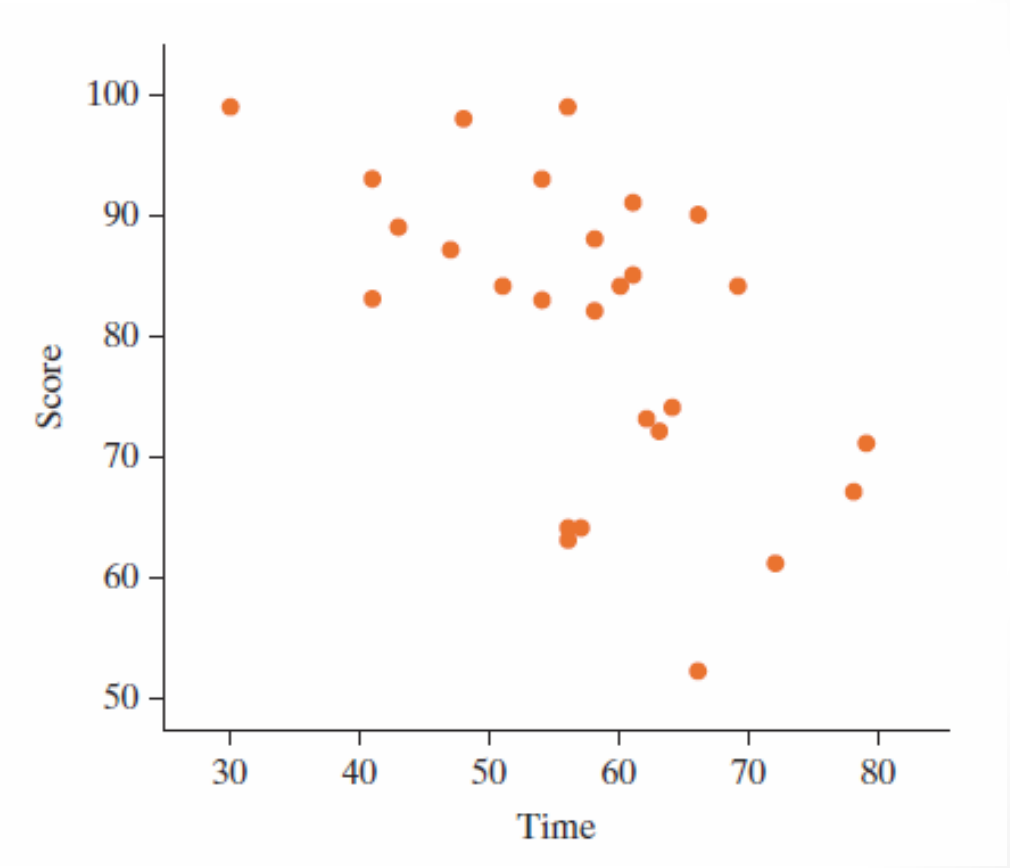
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

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.

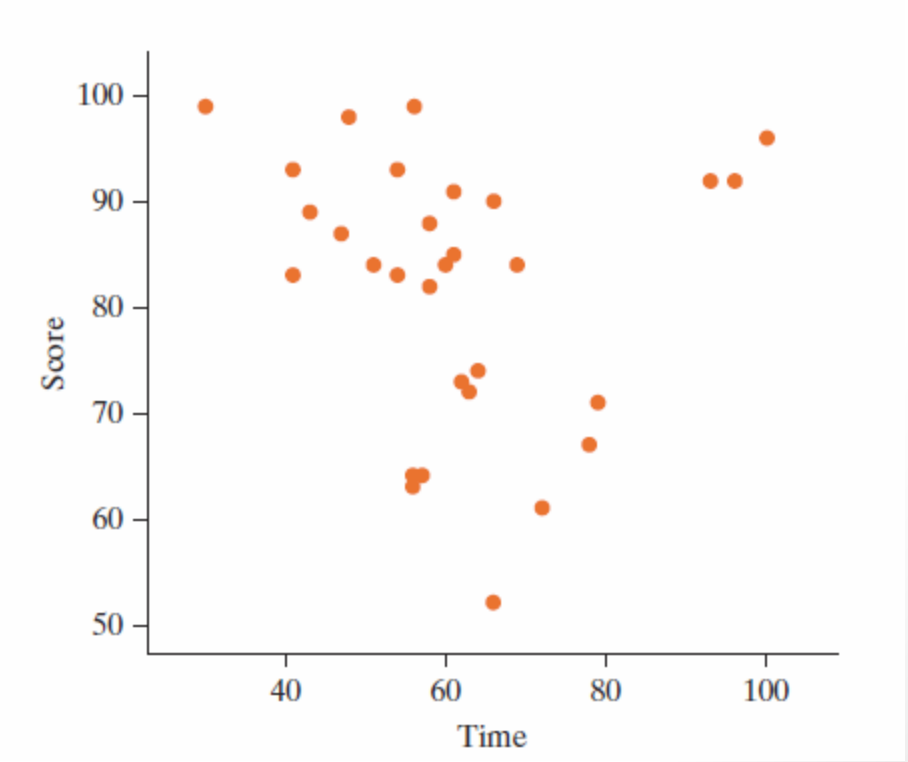
Correlation Guidelines

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

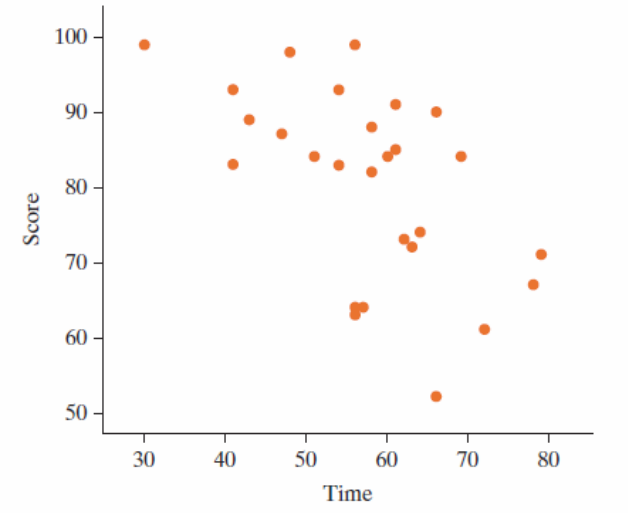
Actually the last three people to finish the test had scores of 93, 93, and 97.

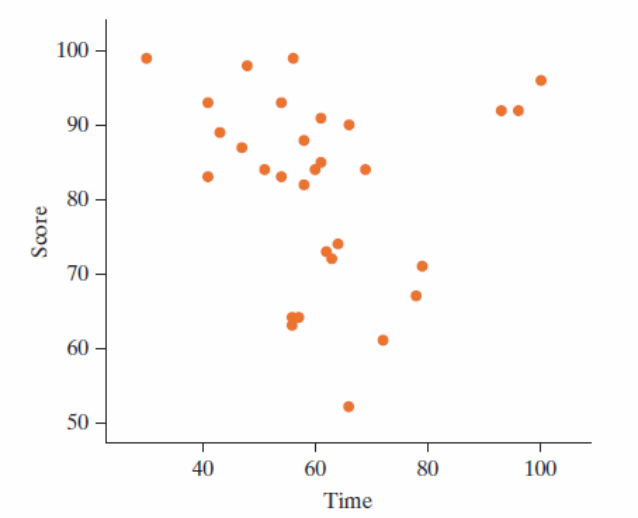
What does this do to the correlation?



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.

Inference for the Correlation Coefficient: Simulation-Based Approach

Section 10.2

We will look at a small sample example to see if body temperature is associated with heart rate.

Temperature and Heart Rate

Hypotheses

- Null: There is no association between heart rate and body temperature. ($\rho = 0$)
- Alternative: There is a positive linear association between heart rate and body temperature. ($\rho > 0$)

ρ = rho

Inference for Correlation with Simulation (Section 10.2)

1. Compute the observed statistic. (Correlation)
2. Scramble the response variable, compute the simulated statistic, and repeat this process many times.
3. Reject the null hypothesis if the observed statistic is in the tail of the null distribution.

Temperature and Heart Rate

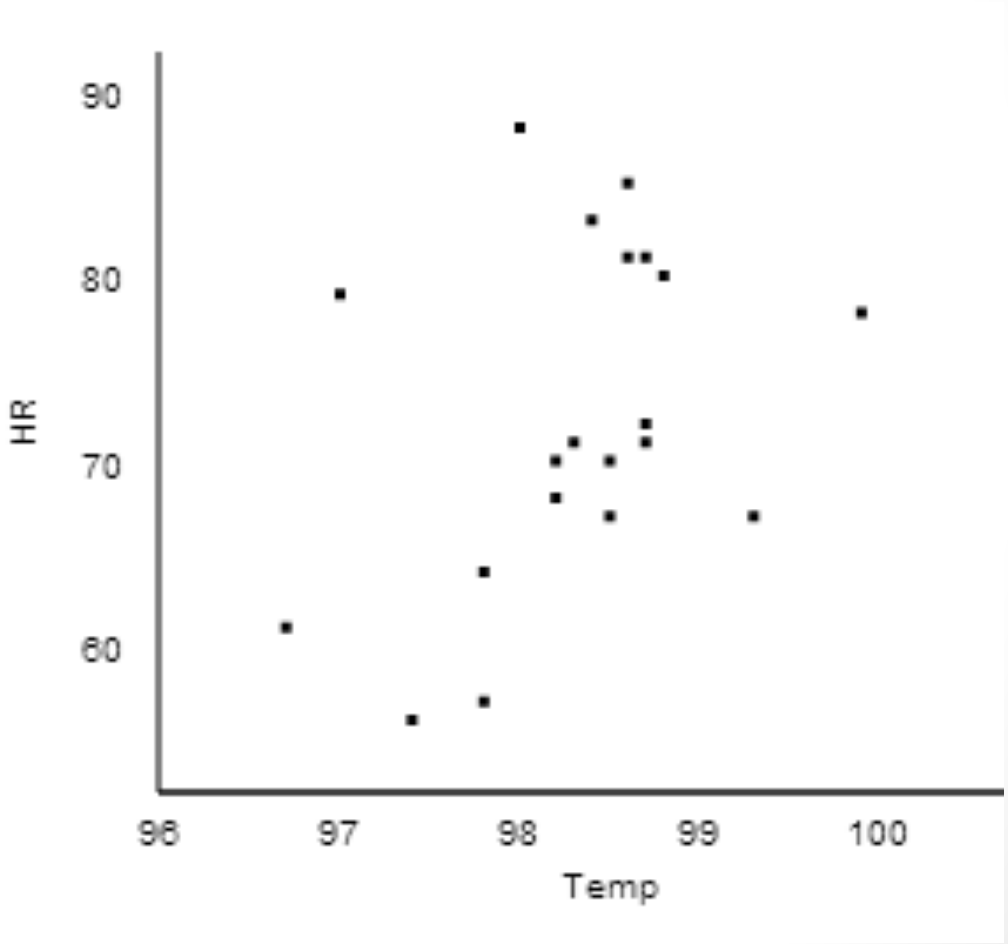
Collect the Data

Tmp	98.3	98.2	98.7	98.5	97.0	98.8	98.5	98.7	99.3	97.8
HR	72	69	72	71	80	81	68	82	68	65
Tmp	98.2	99.9	98.6	98.6	97.8	98.4	98.7	97.4	96.7	98.0
HR	71	79	86	82	58	84	73	57	62	89

Temperature and Heart Rate

Explore the Data

r = 0.378



Temperature and Heart Rate

- If there was no association between heart rate and body temperature, what is the probability we would get a correlation as high as 0.378 just by chance?

- If there is no association, we can break apart the temperatures and their corresponding heart rates. We will do this by shuffling one of the variables.

Shuffling Cards

- Let's remind ourselves what we did with cards to find our simulated statistics.
- With two proportions, we wrote the response on the cards, shuffled the cards and placed them into two piles corresponding to the two categories of the explanatory variable.
- With two means we did the same thing except this time the responses were numbers instead of words.

Dolphin Therapy

Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver

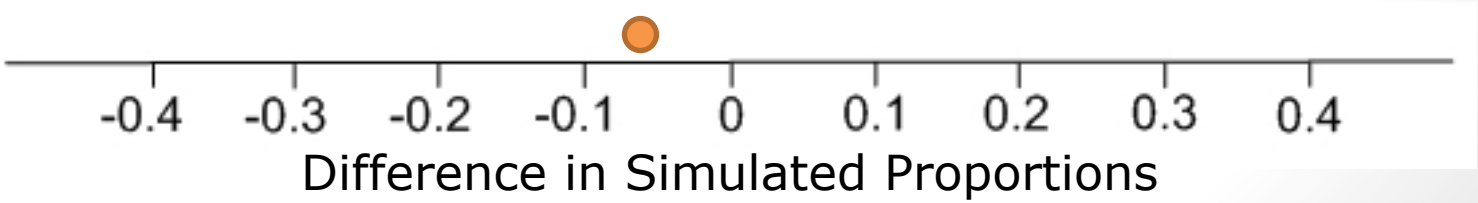
Control

Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

40.0% Improvers

0.400 – 0.467 = -0.067

26.7% Improvers



Music

25.2
14.5
-7.0
12.6
34.5

45.6
11.6
18.6
12.1
30.5

mean = 16.12

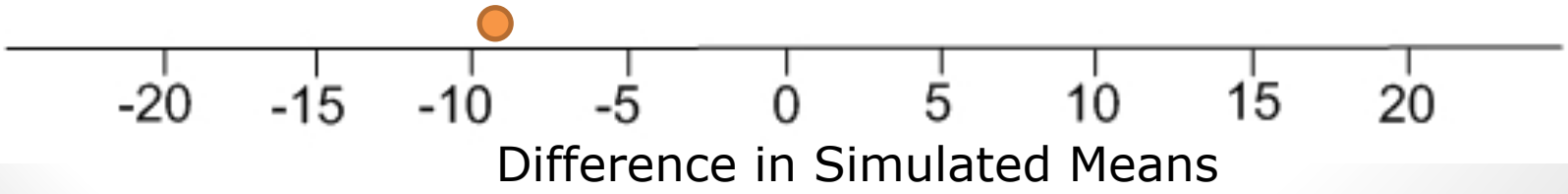
No music

-10.7	-10.7
4.5	9.6
2.2	2.4
21.3	21.8
-14.7	7.2

10.0

mean = 6.38

6.38 – 16.12 = -9.74



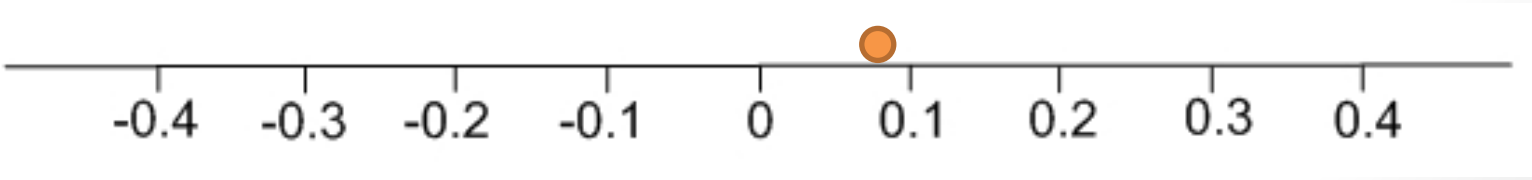
Shuffling Cards

- Now how will this shuffling be different when both the response and the explanatory variable are quantitative?
- We can’t put things in two piles anymore.
- We still shuffle values of the response variable, but this time place them next to two values of the explanatory variable.

Body Temperature and Heart Rate

98.3°	98.2°	97.7°	98.5°	97.0°	98.8°	98.5°	98.7°	99.3°	97.8°
72	69	72	71	80	81	68	82	68	65
98.2°	99.9°	98.6°	98.6°	97.8°	98.4°	98.7°	97.4°	96.7°	98.0°
71	79	86	82	58	84	73	57	62	89

r = 0.078

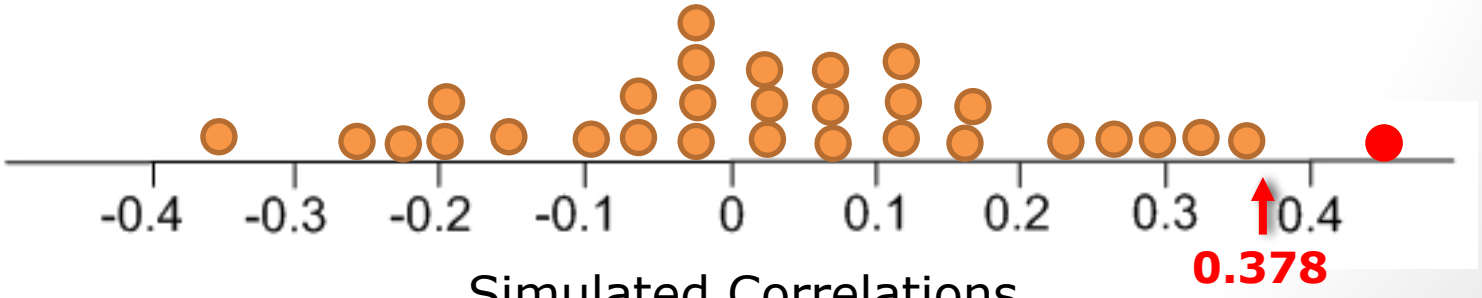


Simulated Correlations

More Simulations

0.097 0.054 0.062 0.259 0.034
-0.253 -0.345 0.339
0.314
0.059 0.200 0.447
0.167
0.067
0.232 -0.327 0.100 0.329

Only one simulated statistic out of 30 was as large or larger than our observed correlation of 0.378, hence our p-value for this null distribution is 1/30 ≈ 0.03.

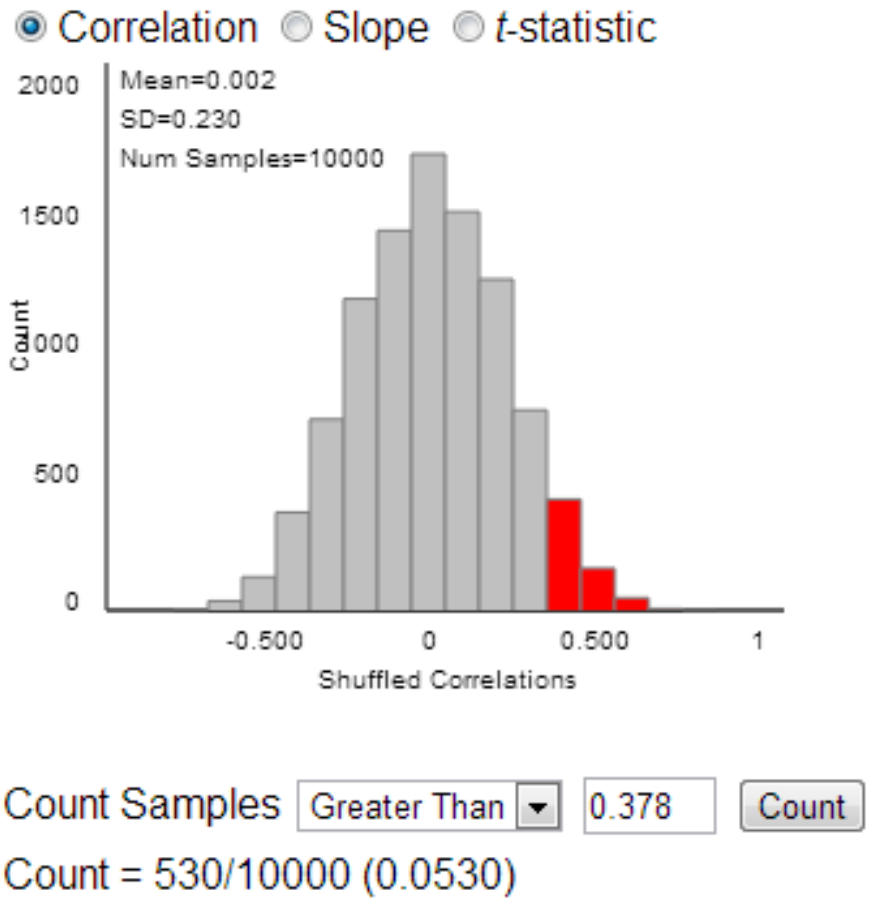


Temperature and Heart Rate

- We can look at the output of 1000 shuffles with a distribution of 1000 simulated correlations.

Temperature and Heart Rate

- Notice our null distribution is centered at 0 and somewhat symmetric.
- We found that 530/10000 times we had a simulated correlation greater than or equal to 0.378.



Temperature and Heart Rate

- With a p-value of 0.053 = 5.3%, we almost but do not quite have statistical significance. This is moderate evidence of a positive linear association between body temperature and heart rate. Perhaps a larger sample would give a smaller p-value.

6. Linear Regression

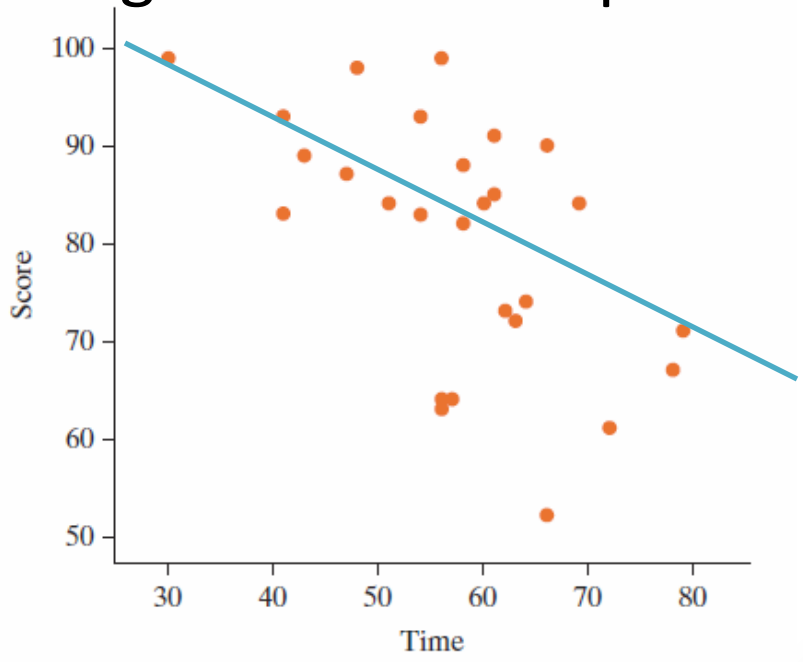
Section 10.3

Introduction

- If we decide an association is linear, it is helpful to develop a mathematical model of that association.
- Helps make predictions about the response variable.
- The ***least-squares regression line*** is the most common way of doing this.

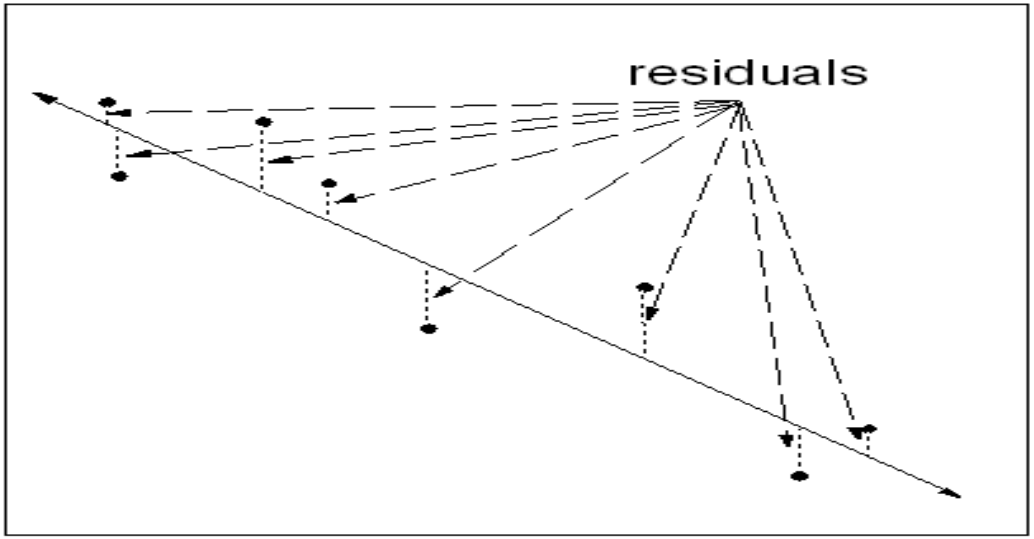
Introduction

- Unless the points are perfectly linearly alligned, there will not be a single line that goes through every point.
- We want a line that gets as close as possible to all the points.



Introduction

- We want a line that minimizes the vertical distances between the line and the points
  - These distances are called **residuals.**
  - The line we will find actually minimizes the sum of the squares of the residuals.
  - This is called a **least-squares regression line**.

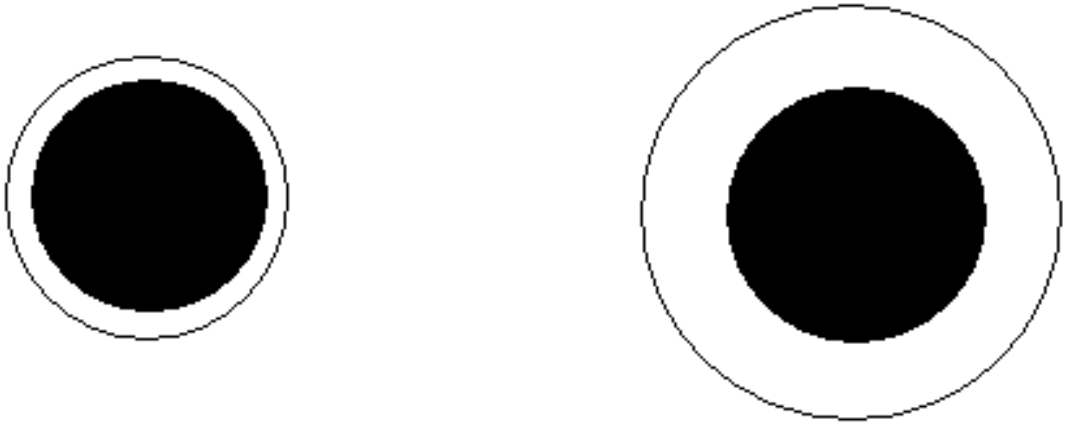


Are Dinner Plates Getting Larger?

Example 10.3

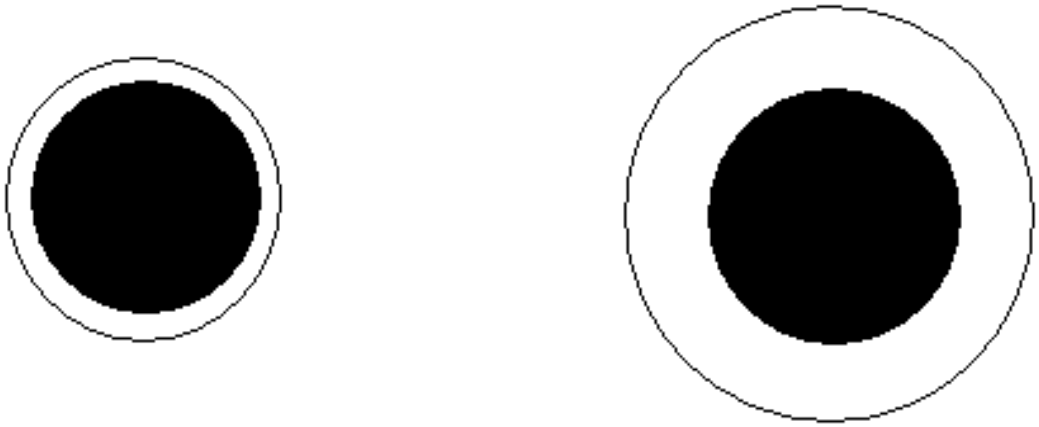
Growing Plates?

- There are many recent articles and TV reports about the obesity problem.
- One reason some have given is that the size of dinner plates are increasing.
- Are these black circles the same size, or is one larger than the other?



Growing Plates?

- They appear to be the same size for many, but the one on the right is about 20% larger than the left.



- This suggests that people will put more food on larger dinner plates without knowing it.
- There is name for this phenomenon: *Delboeuf illusion*

Growing Plates?

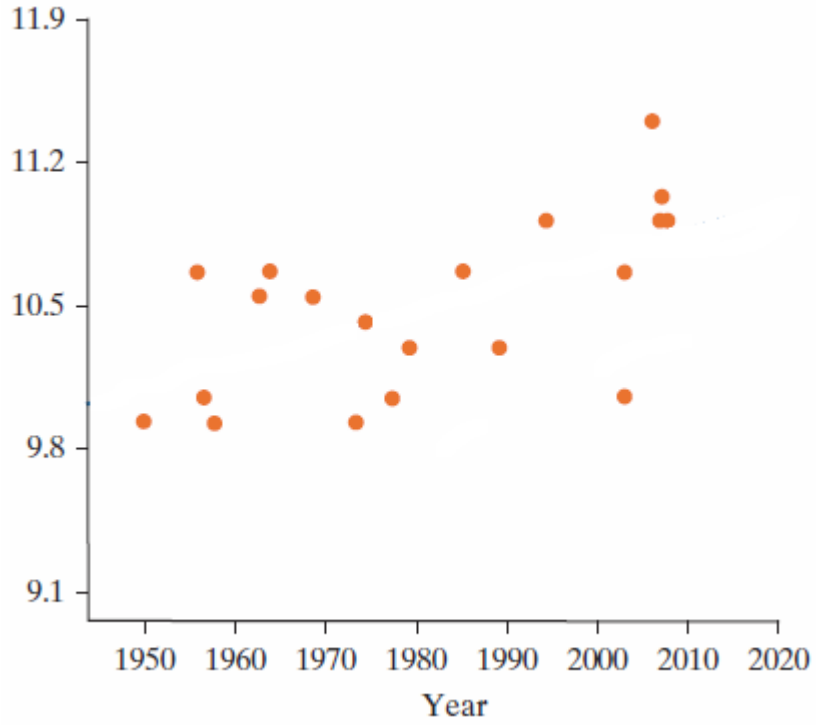
- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on ebay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

TABLE 10.1 Data for size (diameter, in inches) and year of manufacture for 20 American-made dinner plates

Year	1950	1956	1957	1958	1963	1964	1969	1974	1975	1978
Size	10	10.75	10.125	10	10.625	10.75	10.625	10	10.5	10.125
Year	1980	1986	1990	1995	2004	2004	2007	2008	2008	2009
Size	10.375	10.75	10.375	11	10.75	10.125	11.5	11	11.125	11

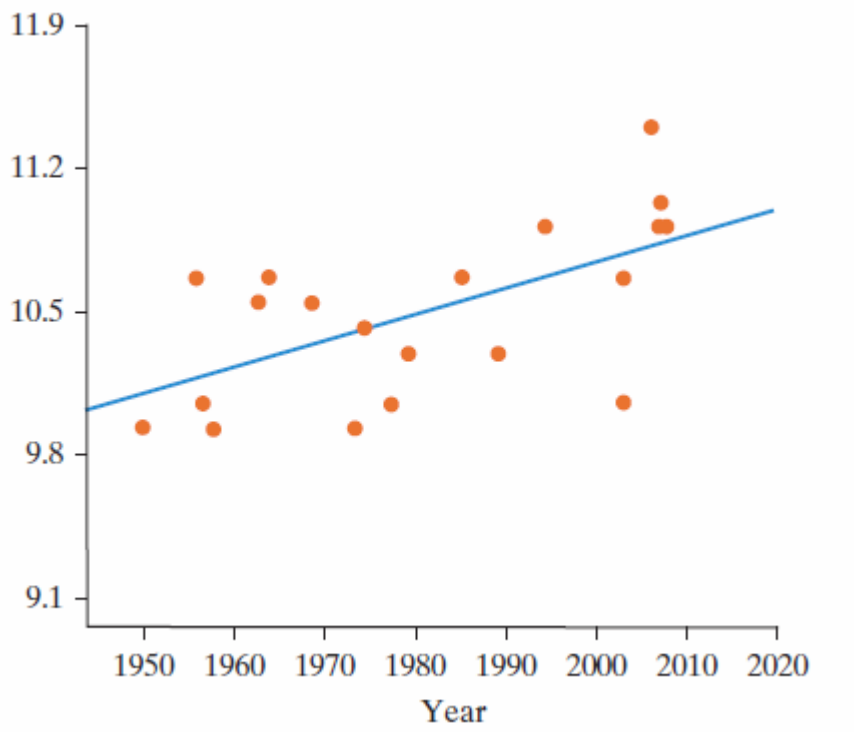
Growing Plates?

- Both year (explanatory variable) and diameter in inches (response variable) are quantitative.
- Each dot represents one plate in this scatterplot.
- Describe the association here.



Growing Plates?

- The association appears to be roughly linear
- The least squares regression line is added
- How can we describe this line?



Regression Line

The regression equation is $\hat{y} = a + bx$:

- *a* is the *y*-intercept
- *b* is the slope
- *x* is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

- For a specific value of *x*, the corresponding distance $y - \hat{y}$ (or actual – predicted) is a residual

Regression Line

- The least squares line for the dinner plate data is $\hat{y} = -14.8 + 0.0128x$
- Or $\widehat{\text{diameter}} = -14.8 + 0.0128(\text{year})$
- This allows us to predict plate diameter for a particular year.

Slope

$$\hat{y} = -14.8 + 0.0128x$$

- What is the predicted diameter for a plate manufactured in 2000?
  - -14.8 + 0.0128(2000) = 10.8 in.
- What is the predicted diameter for a plate manufactured in 2001?
  - -14.8 + 0.0128(2001) = 10.8128 in.
- How does this compare to our prediction for the year 2000?
  - 0.0128 larger
- Slope *b* = 0.0128 means that diameters are predicted to increase by 0.0128 inches per year on average

Slope

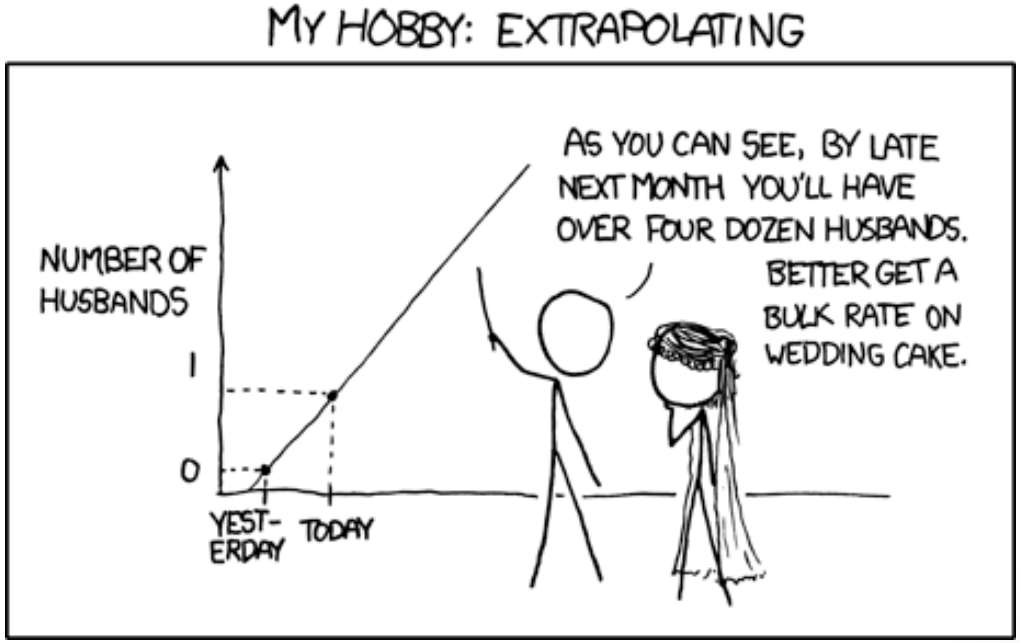
- Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
- Both the slope and the correlation coefficient for this study were positive.
  - The slope is 0.0128
  - The correlation is 0.604
- The slope and correlation coefficient will always have the same sign.

y-intercept

- The *y*-intercept is where the regression line crosses the *y*-axis or the predicted response when the explanatory variable equals 0.
- We had a *y*-intercept of -14.8 in the dinner plate equation. What does this tell us about our dinner plate example?
  - Dinner plates in year 0 were -14.8 inches.
- How can it be negative?
  - The equation works well within the range of values given for the explanatory variable, but fails outside that range.
- Our equation should only be used to predict the size of dinner plates from about 1950 to 2010.

Extrapolation

- Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is called ***extrapolation***.



Coefficient of Determination

- While the intercept and slope have meaning in the context of year and diameter, remember that the correlation does not. It is just 0.604.
- However, the square of the correlation (coefficient of determination or r^2) does have meaning.
- $r^2 = 0.604^2 = 0.365$ or 36.5%
- 36.5% of the variation in plate size (the response variable) can be explained by its linear association with the year (the explanatory variable).

Learning Objectives for Section 10.3

- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for a given value of the explanatory variable.
- Understand the concept of residual and find and interpret the residual for an observational unit given the raw data and the equation of the best fit (regression) line.
- Understand the relationship between residuals and strength of association and that the best-fit (regression) line this minimizes the sum of the squared residuals.

Learning Objectives for Section 10.3

- Find and interpret the coefficient of determination (r^2) as the squared correlation and as the percent of total variation in the response variable that is accounted for by the linear association with the explanatory variable.
- Understand that extrapolation is when a regression line is used to predict values outside of the range of observed values for the explanatory variable.
- Understand that when slope = 0 means no association, slope < 0 means negative association, slope > 0 means positive association, and that the sign of the slope will be the same as the sign of the correlation coefficient.
- Understand that influential points can substantially change the equation of the best-fit line.