

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

1. Collect hw2.
2. Blinding.
3. Portacaval shunt example.
4. Coverage, adherer bias and clofibrate example.
5. More about confounding factors.
6. Confounding and lefties example.
7. Comparing two proportions using numerical and visual summaries, good or bad year example.
8. Comparing 2 proportions with CIs + testing using simulation, dolphin example.
9. Comparing 2 props. with theory-based testing, smoking and gender example.
10. Five number summary, IQR, and geysers.
11. Comparing two means with simulations and bicycling to work example.

Read ch5 and 6. The midterm will be on ch 1-6.

<http://www.stat.ucla.edu/~frederic/13/sum18> .

Bring a PENCIL and CALCULATOR and any books or notes you want to the midterm and final.

HW3 4.CE.10, 5.3.28, 6.1.17, and 6.3.14.

4.CE.10 starts out "Studies have shown that children in the U.S. who have been spanked have a significantly lower IQ score on average...."

5.3.28 starts out "Recall the data from the Physicians' Health Study: Of the 11,034 physicians who took the placebo"

6.1.17 starts out "The graph below displays the distribution of word lengths"

6.3.14 starts out "In an article titled 'Unilateral Nostril Breathing Influences Lateralized Cognitive Performance' that appeared" 1

Some good hw problems from the book are

1.2.18, 1.2.19, 1.2.20, 1.3.17, 1.5.18, 2.1.38, 2.2.6,
2.2.24, 2.3.3, 2.3.25, 3.2.11, 3.2.12, 3.3.8, 3.3.19,
3.3.22, 3.5.23, 4.1.14, 4.1.18, 5.2.2, 5.2.10, 5.2.24,
5.3.11, 5.3.21, 5.3.24, 6.2.23, 6.3.1, 6.3.12, 6.3.22,
6.3.23.

1. Hand in hw2.

2. Blinding.

Even in experiments, the treatment and control groups can be different in ways other than the explanatory variable. This is especially true when the response variable is somewhat subjective.

Pain is an example. One study found that 1/4 of patients suffering from post-operative pain, when given a placebo (just a pill of sugar and water) claimed they experienced "significant prompt pain relief".

2. Blinding.

People might not be able to judge their own levels of pain very well, and may be influenced by the belief that they have taken an effective treatment.

Thus in an experiment with such a response variable, researchers should ensure the subject does not know whether he or she received the treatment or the control. This is called blinding.

In a *double-blind* experiment, neither the subject nor the researcher recording the response variable knows the level of the explanatory variable for each subject, i.e. treatment or control.

3. Portacaval shunt example.

The following example shows the importance of doing a randomized controlled experiment.

The portacaval shunt is a medical procedure aimed at curbing bleeding to death in patients with cirrhosis of the liver.

The following table summarizes 51 studies on the portacaval shunt. The poorly designed studies were very enthusiastic about the surgery, while the carefully designed studies prove that the surgery is largely ineffective.

Design	Degree of enthusiasm		
	High	Moderate	None
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

3. Portacaval shunt example.

Why did the poorly designed studies come to the wrong conclusion?

A likely explanation is that in the studies where patients were not randomly assigned to the treatment or control group, by and large the healthier patients were given the surgery.

This alone could explain why the treatment group outlived the control group in these studies.

Design	Degree of enthusiasm		
	High	Moderate	None
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

4. Coverage, adherer bias and Clofibrate.

Surveys are observational.

- Coverage is a common issue. Coverage is the extent to which the people you sampled from represent the overall population. A survey at a fancy research hospital in a wealthy neighborhood may yield patients with higher incomes, higher education, etc.
- Non-response bias is another common problem. Poor coverage means the people getting the survey do not represent the general population. Non-response bias means that out of the people you gave the survey to, the people actually filling it out and submitting it are different from the people who did not.
- Same exact issues in web surveys.

Coverage, adherer bias, and Clofibrate example.

Non-response bias is similar to adherer bias, in experiments.

A drug called clofibrate was tested on 3,892 middle-aged men with heart trouble. It was supposed to prevent heart attacks.

1,103 assigned at random to take clofibrate,

2,789 to placebo (lactose) group.

Subjects were followed for 5 years.

Is this an experiment or an observational study?

Clofibrate	patients who died during followup
adherers	15%
non-adherers	25%
total	20%

Coverage, adherer bias, and Clofibrate example.

Non-response bias is similar to adherer bias, in experiments.

A drug called clofibrate was tested on 3,892 middle-aged men with heart trouble. It was supposed to prevent heart attacks.

1,103 assigned at random to take clofibrate,

2,789 to placebo (lactose) group.

Subjects were followed for 5 years.

Is this an experiment or an observational study?

It is an experiment. Does Clofibrate work?

Clofibrate	patients who died during followup
------------	-----------------------------------

adherers	15%
----------	------------

non-adherers	25%
--------------	------------

total	20%
-------	------------

Clofibrate patients who died during followup

adherers **15%**

non-adherers **25%**

total 20%

Placebo

adherers 15%

nonadherers 28%

total 21%

Those who took clofibrate did much better than those who didn't keep taking clofibrate. Does this mean clofibrate works?

Clofibrate patients who died during followup

adherers 15%

non-adherers 25%

total 20%

Placebo

adherers **15%**

nonadherers **28%**

total 21%

Those who adhered to placebo also did much better than those who stopped adhering.

Clofibrate	patients who died during followup
adherers	15%
non-adherers	25%
total	20%

Placebo	
adherers	15%
nonadherers	28%
total	21%

All in all there was little difference between the two groups.

Clofibrate	patients who died during followup
adherers	15%
non-adherers	25%
total	20%

Placebo	
adherers	15%
nonadherers	28%
total	21%

Adherers did better than non-adherers, not because of clofibrate, but because they were healthier in general. Why?

Clofibrate	patients who died during followup
adherers	15%
non-adherers	25%
total	20%

Placebo	
adherers	15%
nonadherers	28%
total	21%

Adherers did better than non-adherers, not because of clofibrate, but because they were healthier in general. Why?

- adherers are the type to engage in healthier behavior.
- sick patients are less likely to adhere.

5. More about confounding factors.

- By a confounding factor, we mean an alternative explanation that could explain the apparent relationship between the two variables, even if they are not causally related. Typically this is done by finding another difference between the treatment and control group. For instance, different studies have examined smokers and non-smokers and have found that smokers have higher rates of liver cancer. One explanation would be that smoking causes liver cancer. But is there any other, alternative explanation?
- One alternative would be that the smokers tend to drink more alcohol, and it is the alcohol, not the smoking, that causes liver cancer.

5. More about confounding factors.

- Another plausible explanation is that the smokers are probably older on average than the non-smokers, and older people are more at risk for all sorts of cancer than younger people.
- Another might be that smokers engage in other unhealthy activities more than non-smokers.
- Note that if one said that “smoking makes you want to drink alcohol which causes liver cancer,” that would not be a valid confounding factor, since in that explanation, smoking effective is causally related to liver cancer risk.

6. Lefties example.

- A confounding factor must be plausibly linked to both the explanatory and response variables. So for instance saying “perhaps a higher proportion of the smokers are men” would not be a very convincing confounding factor, unless you have some reason to think gender is strongly linked to liver cancer.
- Another example: left-handedness and age at death. Psychologists Diane Halpern and Stanley Coren looked at 1,000 death records of those who died in Southern California in the late 1980s and early 1990s and contacted relatives to see if the deceased were righthanded or lefthanded. They found that the average ages at death of the lefthanded was 66, and for the righthanded it was 75. Their results were published in prestigious scientific journals, Nature and the New England Journal of Medicine.

6. Lefties example.

All sorts of causal conclusions were made about how this shows that the stress of being lefthanded in our righthanded world leads to premature death.

The New York Times

U.S.

WORLD

U.S.

N.Y. / REGION

BUSINESS

TECHNOLOGY


SCIENCE

HEALTH


SPORTS

OPINION

POLITICS EDUCATION TEXAS




0% APR up to 72 months
\$1,000 Special Bonus Cash*
ON A 2016 SMART FORTWO COUPE
[GET OFFER](#)



* DISCLAIMER


Being Left-Handed May Be Dangerous To Life, Study Says

Reuters
Published: April 4, 1991

BOSTON, April 3— Left-handed people tend to live significantly shorter lives than right-handers, perhaps because they face more perils in a world dominated by the right-handed, according to new research.

 FACEBOOK

 TWITTER

 GOOGLE+

6. Lefties example.

- Is this an observational study or an experiment?

6. Lefties example.

- Is this an observational study or an experiment?

It is an observational study.

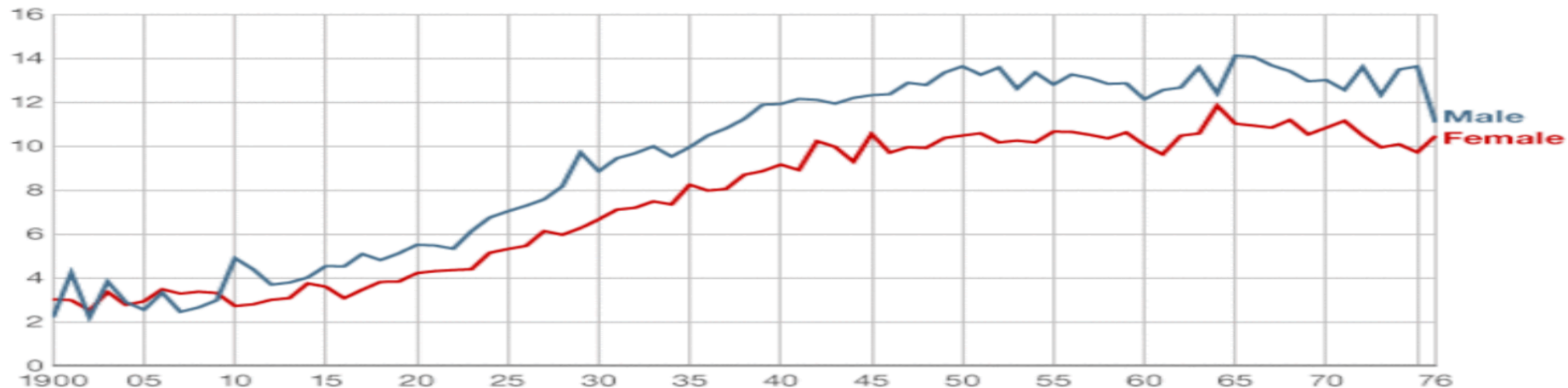
- Are there plausible confounding factors you can think of?

6. Lefties example.

- A confounding factor is the age of the two populations in general. Lefties in the 1980s were on average younger than righties. Many old lefties were converted to righties at infancy, in the early 20th century, but this practice has subsided. Thus in the 1980s and 1990s, there were relatively few old lefties but many young lefties in the overall population. This alone explains the discrepancy.

Left handedness 1900-1976

% of population



Source: Chris McManus Right Hand, Left Hand

Unit 2. Comparing Two Groups

- In Unit 1, we learned the basic process of statistical inference using tests and confidence intervals. We did all this by focusing on a single proportion.
- In Unit 2, we will take these ideas and extend them to comparing two groups. We will compare two proportions, two independent means, and paired data.

7. Comparing two proportions using numerical and visual summaries, and the good or bad year example.

Section 5.1

Example 5.1:

Positive and Negative Perceptions

- Consider these two questions:
 - Are you having a good year?
 - Are you having a bad year?
- Do people answer each question in such a way that would indicate the same answer? (e.g. Yes for the first one and No for the second.)

Positive and Negative Perceptions

- Researchers questioned 30 students (randomly giving them one of the two questions).
- They then recorded if a positive or negative response was given.
- They wanted to see if the wording of the question influenced the answers.

Positive and negative perceptions

- Observational units
 - The 30 students
- Variables
 - Question wording (good year or bad year)
 - Perception of their year (positive or negative)
- Which is the explanatory variable and which is the response variable?
- Is this an observational study or experiment?

Raw Data in a Spreadsheet

Individual	Type of Question	Response
1	Good Year	Positive
2	Good Year	Negative
3	Bad Year	Positive
4	Good Year	Positive
5	Good Year	Negative
6	Bad Year	Positive
7	Good Year	Positive
8	Good Year	Positive
9	Good Year	Positive
10	Bad Year	Negative
11	Good Year	Negative
12	Bad Year	Negative
13	Good Year	Positive
14	Bad Year	Negative
15	Good Year	Positive

Individual	Type of Question	Response
16	Good Year	Positive
17	Bad Year	Positive
18	Good Year	Positive
19	Good Year	Positive
20	Good Year	Positive
21	Bad Year	Negative
22	Good Year	Positive
23	Bad Year	Negative
24	Good Year	Positive
25	Bad Year	Negative
26	Good Year	Positive
27	Bad Year	Negative
28	Good Year	Positive
29	Bad Year	Positive
30	Bad Year	Negative

Two-Way Tables

- A **two-way table** organizes data
 - Summarizes *two* categorical variables
 - Also called contingency table
- Are students more likely to give a positive response if they were given the good year question?

	Good Year	Bad Year	Total
Positive response	15	4	19
Negative response	3	8	11
Total	18	12	30

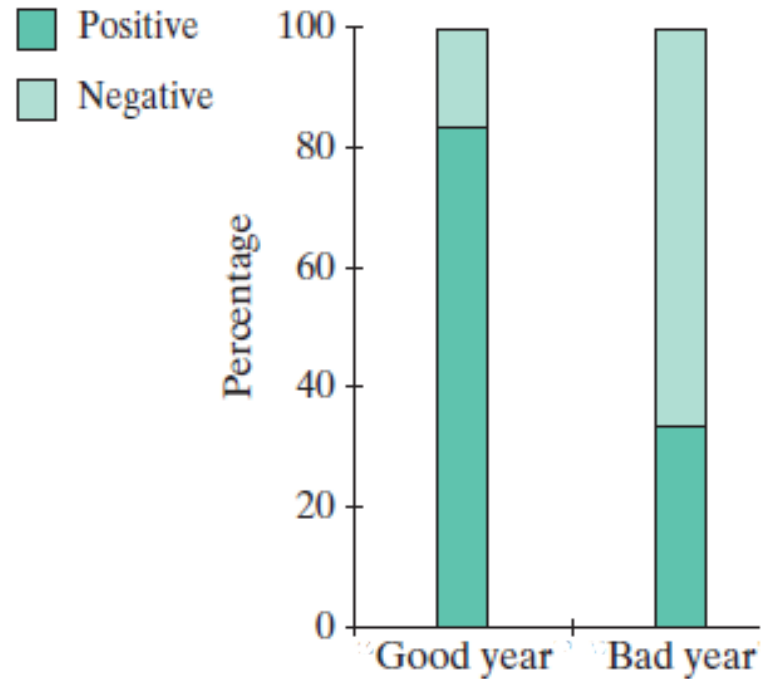
Two-Way Tables

- Conditional proportions will help us better determine if there is an association between the question asked and the type of response.
- We can see that the subjects with the positive question were ***more likely*** to respond positively.

	Good Year	Bad Year	Total
Positive response	$15/18 \approx 0.83$	$4/12 \approx 0.33$	19
Negative response	3	8	11
Total	18	12	30

Segmented Bar Graphs

- We can also use segmented bar graphs to see this association between the "good year" question and a positive response.



Statistic

- The statistic we will mainly use to summarize this table is the difference in proportions of positive responses is $0.83 - 0.33 = 0.50$.

	Good Year	Bad Year	Total
Positive response	15 (83%)	4 (33%)	19
Negative response	3	8	11
Total	18	12	30

Another Statistic

- Another statistic that is often used, called **relative risk**, is the ratio of the proportions: $0.83 / 0.33 = 2.5$.
- We can say that those who were given the good year question were 2.5 times as likely to give a positive response.

	Good Year	Bad Year	Total
Positive response	15 (83%)	4 (33%)	19
Negative response	3	8	11
Total	18	12	30

8. Comparing two proportions with CIs and testing using simulation, dolphin example.

Section 5.2

Swimming with Dolphins

Example 5.2

Swimming with Dolphins

Is swimming with dolphins therapeutic for patients suffering from clinical depression?

- Researchers Antonioli and Reveley (2005), in British Medical Journal, recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression
- Discontinued antidepressants and psychotherapy 4 weeks prior to and throughout the experiment
- 30 subjects went to an island near Honduras where they were randomly assigned to two treatment groups

Swimming with Dolphins

- Both groups engaged in one hour of swimming and snorkeling each day
- One group swam in the presence of dolphins and the other group did not
- Participants in both groups had identical conditions except for the dolphins
- After two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study
- The response variable is whether or not the subject achieved substantial reduction in depression

Swimming with Dolphins

Null hypothesis: Dolphins do not help.

- Swimming with dolphins is not associated with substantial improvement in depression

Alternative hypothesis: Dolphins help.

- Swimming with dolphins **increases** the probability of substantial improvement in depression symptoms

Swimming with Dolphins

- The parameter is the (long-run) difference between the probability of improving when receiving dolphin therapy and the prob. of improving with the control ($\pi_{\text{dolphins}} - \pi_{\text{control}}$)
- So we can write our hypotheses as:

$$\mathbf{H}_0: \pi_{\text{dolphins}} - \pi_{\text{control}} = 0.$$

$$\mathbf{H}_a: \pi_{\text{dolphins}} - \pi_{\text{control}} > 0.$$

or

$$\mathbf{H}_0: \pi_{\text{dolphins}} = \pi_{\text{control}}$$

$$\mathbf{H}_a: \pi_{\text{dolphins}} > \pi_{\text{control}}$$

(Note: we are not saying our parameters equal any certain number.)

Swimming with Dolphins

Results:

	Dolphin group	Control group	Total
Improved	10 (66.7%)	3 (20%)	13
Did Not Improve	5	12	17
Total	15	15	30

The difference in proportions of improvers is:

$$\hat{p}_d - \hat{p}_c = 0.667 - 0.20 = \mathbf{0.467}.$$

Swimming with Dolphins

- There are two possible explanations for an observed difference of 0.467.
 - A tendency to be more likely to improve with dolphins (alternative hypothesis)
 - The 13 subjects were going to show improvement with or without dolphins and random chance assigned more improvers to the dolphins (null hypothesis)

Swimming with Dolphins

- If the null hypothesis is true (no association between dolphin therapy and improvement) we would have 13 improvers and 17 non-improvers regardless of the group to which they were assigned.
- Hence the assignment doesn't matter and we can just randomly assign the subjects' results to the two groups to see what would happen under a true null hypothesis.

Swimming with Dolphins

- We can simulate this with cards
 - 13 cards to represent the improvers
 - 17 cards represent the non-improvers
- Shuffle the cards
 - put 15 in one pile (dolphin therapy)
 - put 15 in another (control group)

Swimming with Dolphins

- Compute the proportion of improvers in the Dolphin Therapy group
- Compute the proportion of improvers in the Control group
- The difference in these two proportions is what could just as well have happened under the assumption there is no association between swimming with dolphins and substantial improvement in depression.

Dolphin Therapy

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

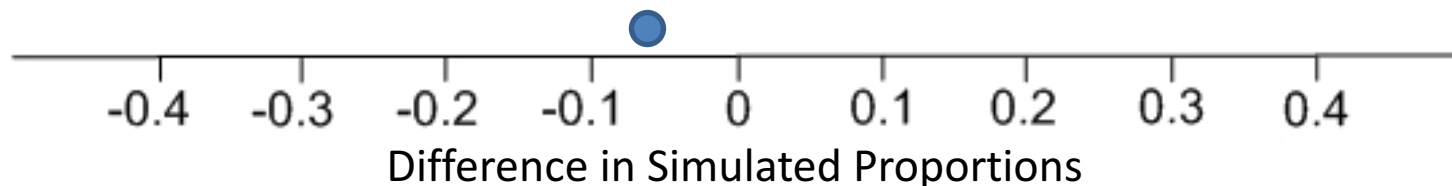
60.0% Improvers

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

$$0.400 - 0.467 = -0.067$$



Dolphin Therapy

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

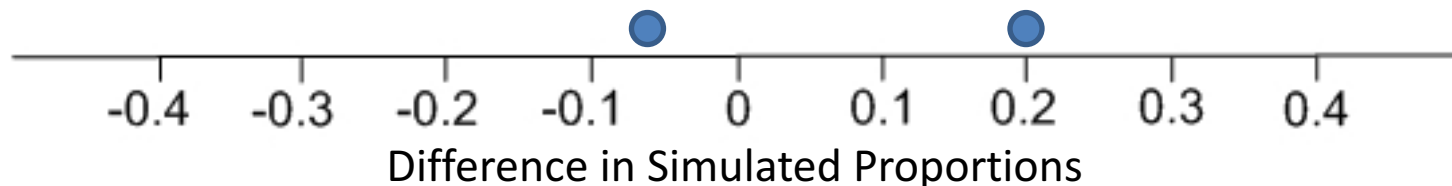
33.3% Improvers

Control

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

33.3% Improvers

$$0.533 - 0.333 = 0.200$$



Dolphin Therapy

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

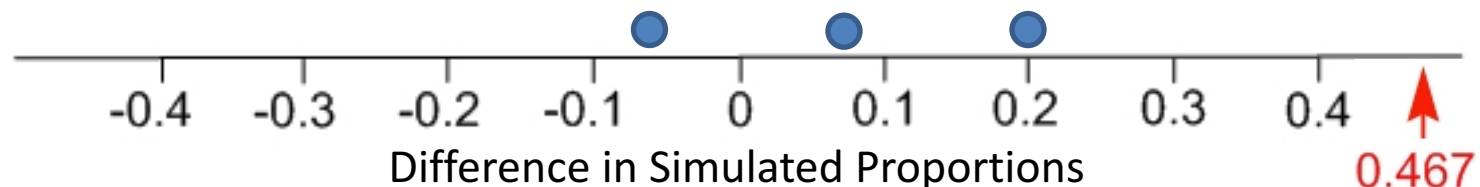
46.7% Improvers

Control

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

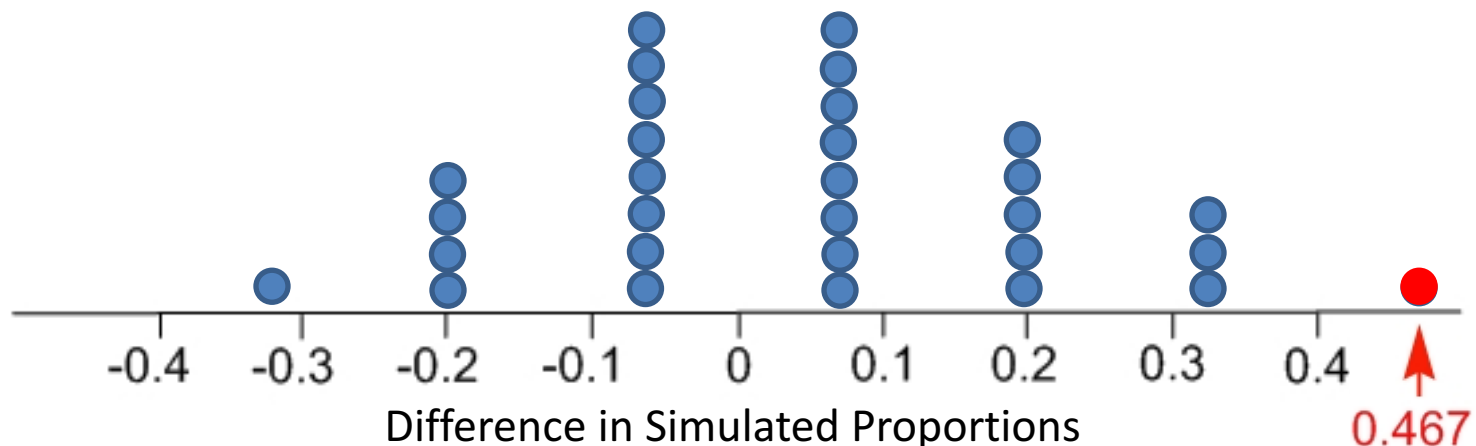
40.0% Improvers

$$0.467 - 0.400 = 0.067$$



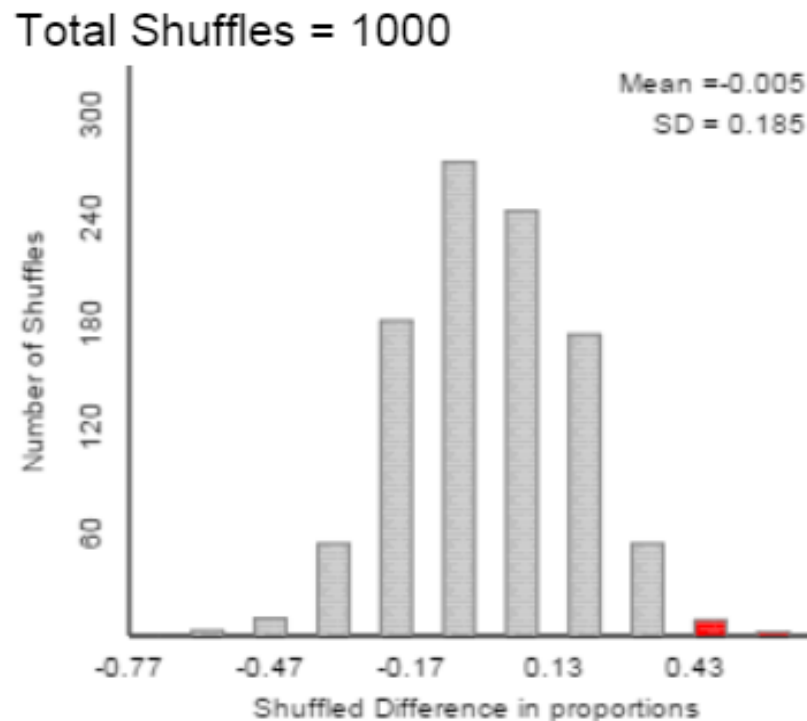
More Simulations

Only one simulated statistics out of 30 was as large or larger than our observed difference in proportions of 0.467, hence our p-value for this null distribution is $1/30 \approx 0.03$.



Swimming with Dolphins

- We did 1000 repetitions to develop a null distribution.



Swimming with Dolphins

- 13 out of 1000 results had a difference of 0.467 or higher (p-value = 0.013).
- 0.467 is $\frac{0.467 - 0}{0.185} \approx 2.52$ SE above zero.
- Using either the p-value or standardized statistic, we have strong evidence against the null and can conclude that the improvement due to swimming with dolphins was statistically significant.

Swimming with Dolphins

- A 95% confidence interval for the difference in the probability using the standard error from the simulations is $0.467 \pm 1.96(0.185) = 0.467 \pm 0.363$, or $(.104, .830)$.
- We are 95% confident that when allowed to swim with dolphins, the probability of improving is between 0.104 and 0.830 higher than when no dolphins are present.
- How does this interval back up our conclusion from the test of significance?

Swimming with Dolphins

- Can we say that the presence of dolphins *caused* this improvement?
 - Since this was a randomized experiment, and assuming everything was identical between the groups, we have strong evidence that dolphins were the cause
- Can we generalize to a larger population?
 - Maybe mild to moderately depressed 18-65 year old patients willing to volunteer for this study
 - We have no evidence that random selection was used to find the 30 subjects. "Outpatients, recruited through announcements on the internet, radio, newspapers, and hospitals."

9. Comparing two proportions: Theory-Based Approach, and smoking and gender example.

Section 5.3

Introduction

- Just as with a single proportion, we can often predict results of a simulation using a theory-based approach.
- The theory-based approach also gives a simpler way to generate a confidence intervals.
- The main new mathematical fact to use is the SE for the difference between two proportions is

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

Parents' Smoking Status and their Babies' Gender

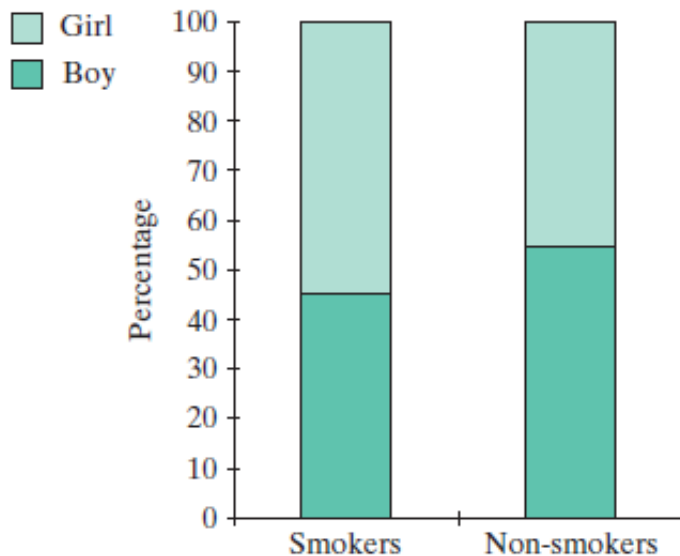
Example 5.3

Smoking and Gender

- How does parents' behavior affect the gender of their children?
- Fukuda et al. (2002) found the following in Japan.
 - Out of 565 births where both parents smoked more than a pack a day, 255 were boys. This is 45.1% boys.
 - Out of 3602 births where both parents did not smoke, 1975 were boys. This 54.8% boys.
 - In total, out of 4170 births, 2230 were boys, which is 53.5%.
- Other studies have shown a reduced male to female birth ratio where high concentrations of other environmental chemicals are present (e.g. industrial pollution, pesticides)

Smoking and Gender

- A segmented bar graph and 2-way table
- Let's compare the proportions to see if the difference is statistically significantly.



	Both Smoked	Neither Smoked
Boy	255 (45.1%)	1,975 (54.8%)
Girl	310	1,627
Total	565	3,602

Smoking and Gender

Null Hypothesis:

- There **is no association** between smoking status of parents and sex of child.
- The probability of having a boy **is the same** for parents who smoke and don't smoke.
- $\pi_{\text{smoking}} - \pi_{\text{nonsmoking}} = 0$

Smoking and Gender

Alternative Hypothesis:

- There **is an association** between smoking status of parents and sex of child.
- The probability of having a boy **is not the same** for parents who smoke and don't smoke
- $\pi_{\text{smoking}} - \pi_{\text{nonsmoking}} \neq 0$

Smoking and Gender

- What are the observational units in the study?
- What are the variables in this study?
- Which variable should be considered the explanatory variable and which the response variable?
- What is the parameter of interest?
- Can you draw cause-and-effect conclusions for this study?

Smoking and Gender

Using the 3S Strategy to assess the strength

1. Statistic:

- The proportion of boys born to nonsmokers minus the proportion of boys born to smokers is $0.548 - 0.451 = 0.097$.

Smoking and Gender

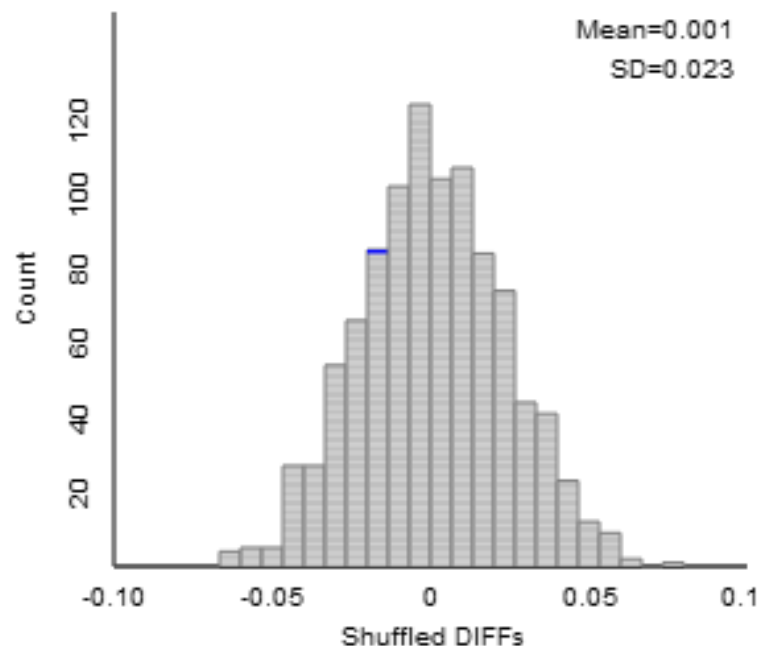
2. Simulate:

- Many repetitions of shuffling the 2230 boys and 1937 girls to the 565 smoking and 3602 nonsmoking parents
- Calculate the difference in proportions of boys between the groups for each repetition.
- Shuffling simulates the null hypothesis of no association

Smoking and Gender

3. Strength of evidence:

- Nothing as extreme as our observed statistic (≥ 0.097 or ≤ -0.097) occurred in 5000 repetitions,
- How many SEs is 0.097 above the mean?
 $Z = 0.097/0.023 = 4.22$
using simulations. What about using the theory-based approach?

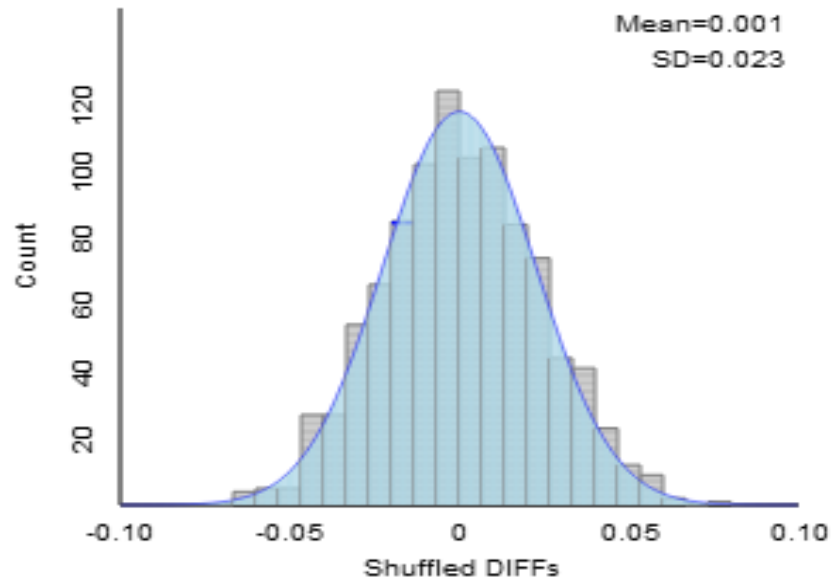


Count Samples

Count = 0/1000 (0.0000)

Smoking and Gender

- Notice the null distribution is centered at zero and is bell-shaped.
- This can be approximated by the normal distribution.



Formulas

- The theory-based approach yields $z = 4.30$.

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

- Here $z = \frac{.548 - .451}{\sqrt{.535 (1 - .535) \left(\frac{1}{3602} + \frac{1}{565} \right)}} = 4.30$.
- p-value is $2 * (1 - \text{pnorm}(4.30)) = 0.00171\%$.

Smoking and Gender

- Fukuda et al. (2002) found the following in Japan.
 - Out of 3602 births where both parents did not smoke, 1975 were boys. This is 54.8% boys.
 - Out of 565 births where both parents smoked more than a pack a day, 255 were boys. This is 45.1% boys.
 - In total, out of 4170 births, 2230 were boys, which is 53.5% boys.

Formulas

- How do we find the margin of error for the difference in proportions?

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

- The multiplier is from the normal distribution and is dependent upon the confidence level.
 - 1.645 for 90% confidence
 - 1.96 for 95% confidence
 - 2.576 for 99% confidence
- We can write the confidence interval in the form:
 - statistic \pm margin of error.

Smoking and Gender

- Our statistic is the observed sample difference in proportions, 0.097.
- Plugging in $1.96 \times \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} = 0.044$, we get 0.097 ± 0.044 as our 95% CI.
- We could also write this interval as (0.053, 0.141).
- We are 95% confident that the probability of a boy baby where neither family smokes minus the probability of a boy baby where both parents smoke is between 0.053 and 0.141.

A clarification on the formulas

- The margin of error for the difference in proportions is

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

In testing, the null hypothesis is no difference between the two groups, so we used the 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. But in

CIs, we use the formula $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$ because we

are not assuming $\hat{p}_1 = \hat{p}_2$ with CIs.

Smoking and Gender

- How would the interval change if the confidence level was 99%?
- The SE = $\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} = .0224$.
- Previously, for a 95% CI, it was $0.097 \pm 1.96 \times .0224 = 0.097 \pm 0.044$.
- For a 99% CI, it is $0.097 \pm 2.576 \times .0224 = 0.097 \pm 0.058$.

Smoking and Gender

- Written as the statistic \pm margin of error, the 99% CI for the difference between the two proportions is

$$0.097 \pm 0.058.$$

- Margin of error
 - 0.058 for the 99% confidence interval
 - 0.044 for the 95% confidence interval

Smoking and Gender

- How would the 95% confidence interval change if we were estimating

$$\pi_{\text{smoker}} - \pi_{\text{nonsmoker}}$$

instead of

$$\pi_{\text{nonsmoker}} - \pi_{\text{smoker}} ?$$

Smoking and Gender

- $(-0.141, -0.053)$ or -0.097 ± 0.044
instead of
- $(0.053, 0.141)$ or 0.097 ± 0.044 .
- The negative signs indicate the probability of a boy born to smoking parents is lower than that for nonsmoking parents.

Smoking and Gender

Validity Conditions of Theory-Based

- Same as with a single proportion.
- Should have at least 10 observations in each of the cells of the 2 x 2 table.

	Smoking Parents	Non-smoking Parents	Total
Male	255	1975	2230
Female	310	1627	1937
Total	565	3602	4167

Smoking and Gender

- The strong significant result in this study yielded quite a bit of press when it came out.
- Soon other studies came out which found no relationship between smoking and gender (Parazinni et al. 2004, Obel et al. 2003).
- James (2004) argued that confounding variables like social factors, diet, environmental exposure or stress were the reason for the association between smoking and gender of the baby. These are all confounded since it was an observational study. Different studies could easily have had different levels of these confounding factors.

10. Five number summary, IQR, and geysers.

6.1: Comparing Two Groups: Quantitative Response

6.2: Comparing Two Means: Simulation-Based Approach

6.3: Comparing Two Means: Theory-Based Approach

Exploring Quantitative Data

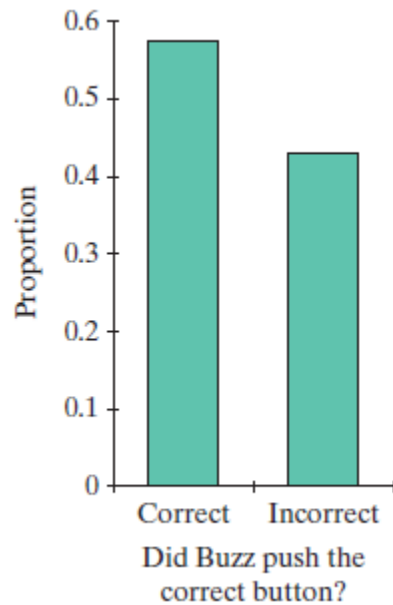
Section 6.1

Quantitative vs. Categorical Variables

- Categorical
 - Values for which arithmetic does not make sense.
 - Gender, ethnicity, eye color...
- Quantitative
 - You can add or subtract the values, etc.
 - Age, height, weight, distance, time...

Graphs for a Single Variable

Categorical



Bar Graph

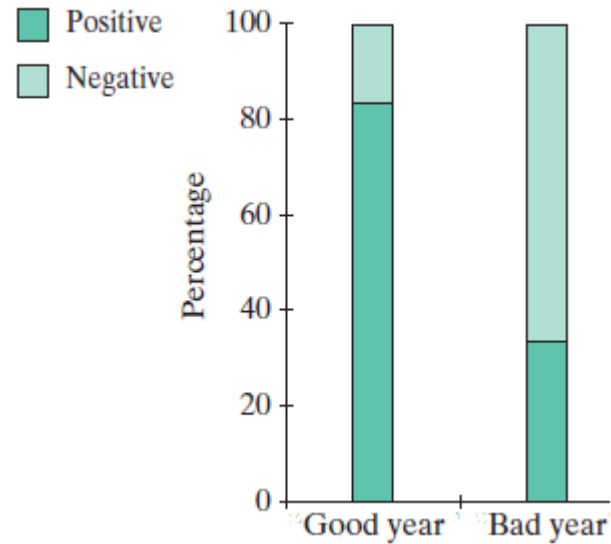
Quantitative



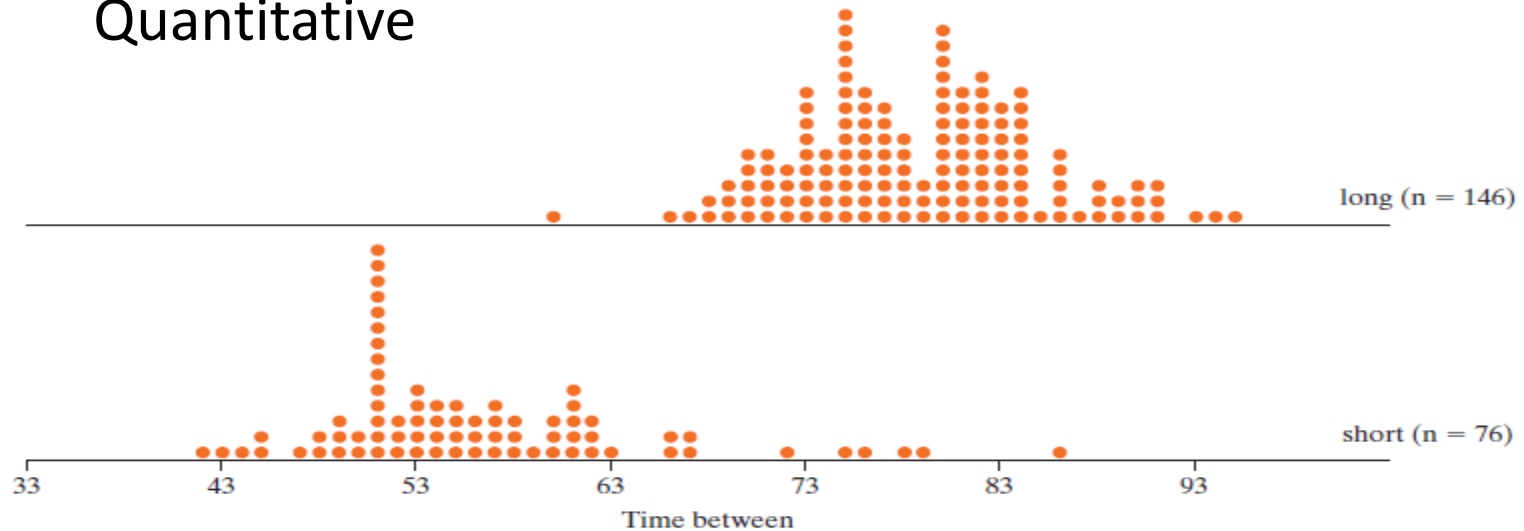
Dot Plot

Comparing Two Groups Graphically

Categorical



Quantitative



Notation Check

Statistics

- \bar{x} Sample mean
- \hat{p} Sample proportion.

Parameters

- μ Population mean
- π Population proportion or probability.

Statistics summarize a sample and parameters summarize a population

Quartiles

- Suppose 25% of the observations lie below a certain value x . Then x is called the ***lower quartile*** (or 25th percentile).
- Similarly, if 25% of the observations are greater than x , then x is called the ***upper quartile*** (or 75th percentile).
- The lower quartile can be calculated by finding the median, and then determining the median of the values below the overall median. Similarly the upper quartile is $\text{median}\{x_i : x_i > \text{overall median}\}$.

IQR and Five-Number Summary

- The difference between the quartiles is called the ***inter-quartile range*** (IQR), another measure of variability along with standard deviation.
- The ***five-number summary*** for the distribution of a quantitative variable consists of the minimum, lower quartile, median, upper quartile, and maximum.
- Technically the IQR is not the interval (25th percentile, 75th percentile), but the difference 75th percentile – 25th .
- Different software use different conventions, but we will use the convention that, if there is a range of possible quantiles, you take the middle of that range.
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
- $M = 7.5$, 25th percentile = 5, 75th percentile = 10.5. IQR = 5.5.

IQR and Five-Number Summary

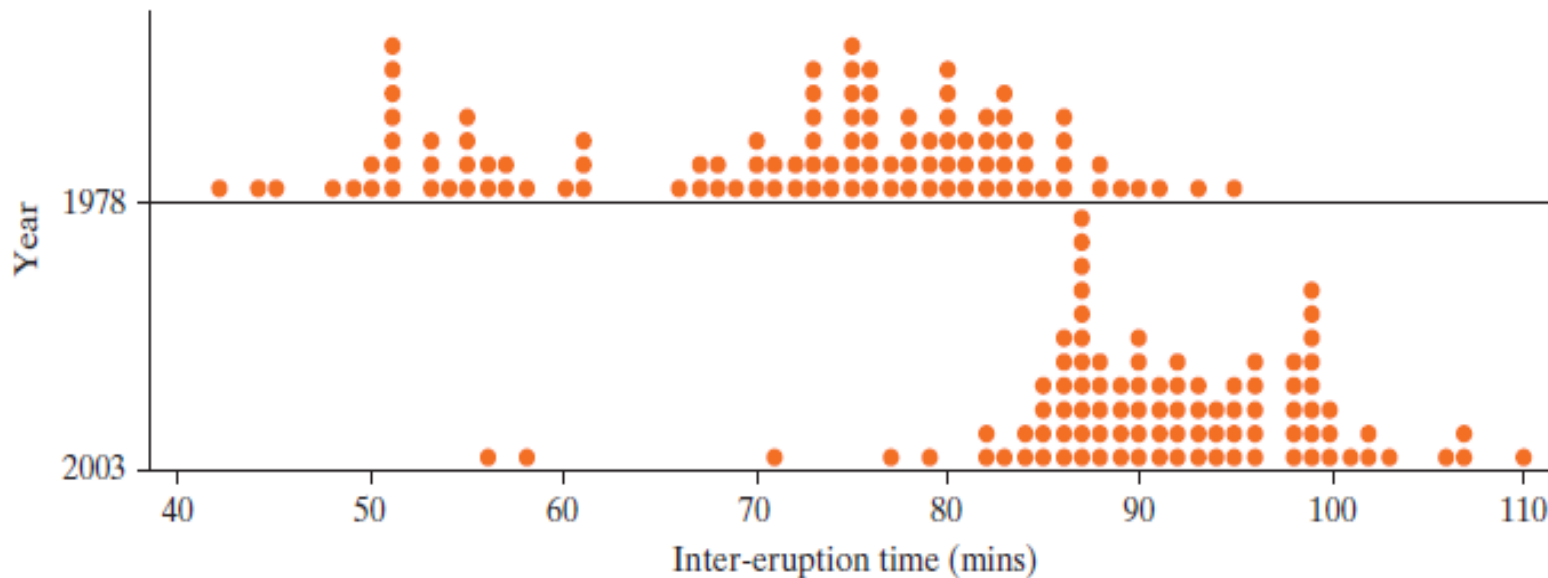
- For medians and quartiles, we will use the convention, if there is a range of possibilities, take the middle of the range.
 - In R, this is `type = 2`. `type = 1` means take the minimum.
 - `x = c(1, 3, 7, 7, 8, 9, 12, 14)`
 - `quantile(x,.25, type=2) ## 5.5`
 - `IQR(x,type=2) ## 5.5`
 - `IQR(x,type=1) ## 6`. Can you see why?
-
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
 - $M = 7.5$, 25th percentile = 5, 75th percentile = 10.5. IQR = 5.5.

Geyser Eruptions

Example 6.1

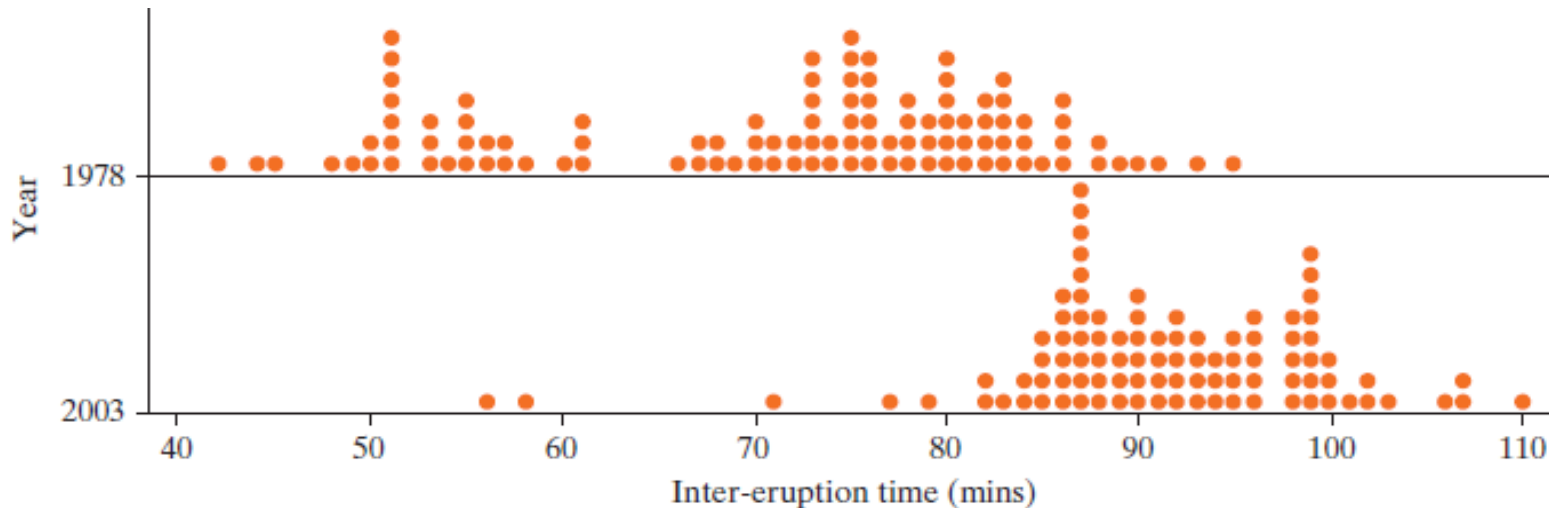
Old Faithful Inter-Eruption Times

- How do the five-number summary and IQR differ for inter-eruption times between 1978 and 2003?



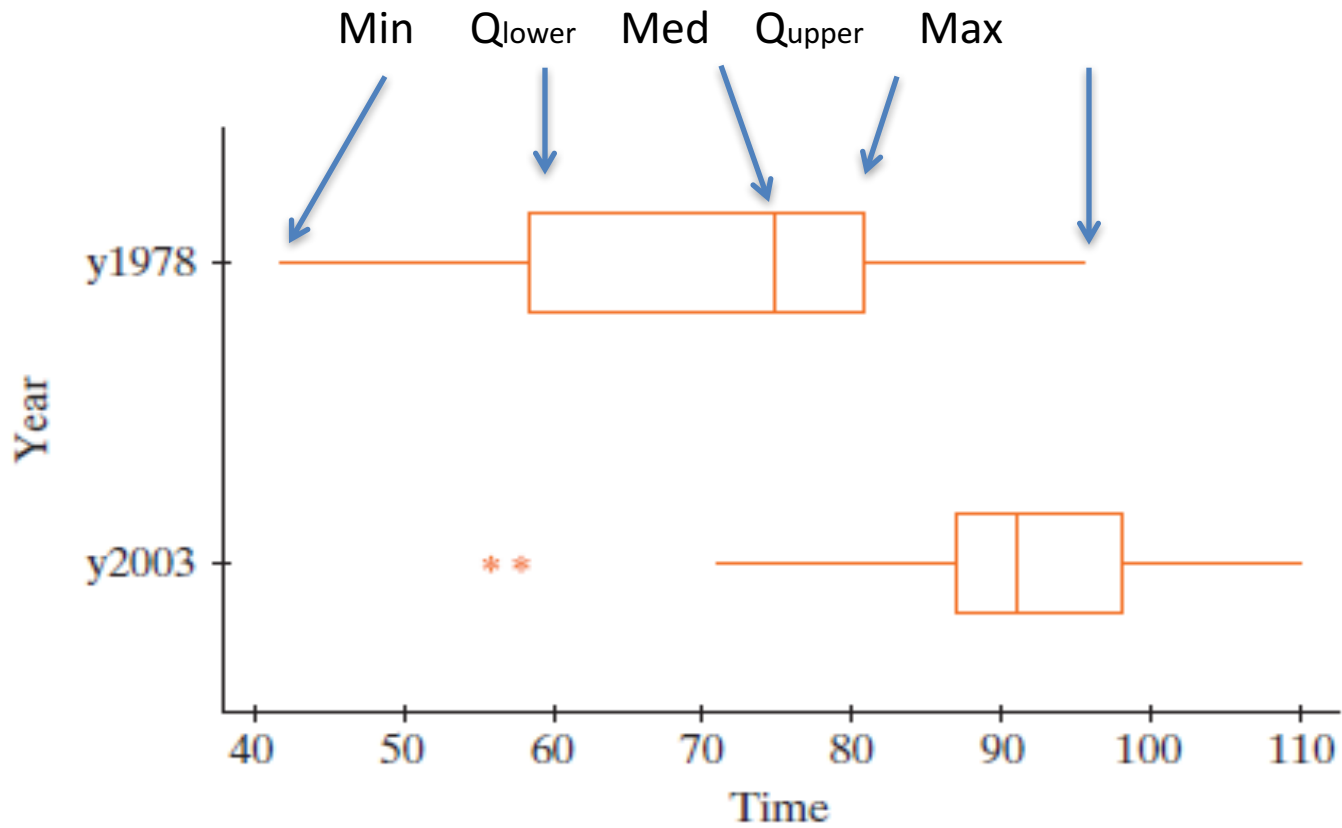
Old Faithful Inter-Eruption Times

	Minimum	Lower quartile	Median	Upper quartile	Maximum
1978 times	42	58	75	81	95
2003 times	56	87	91	98	110



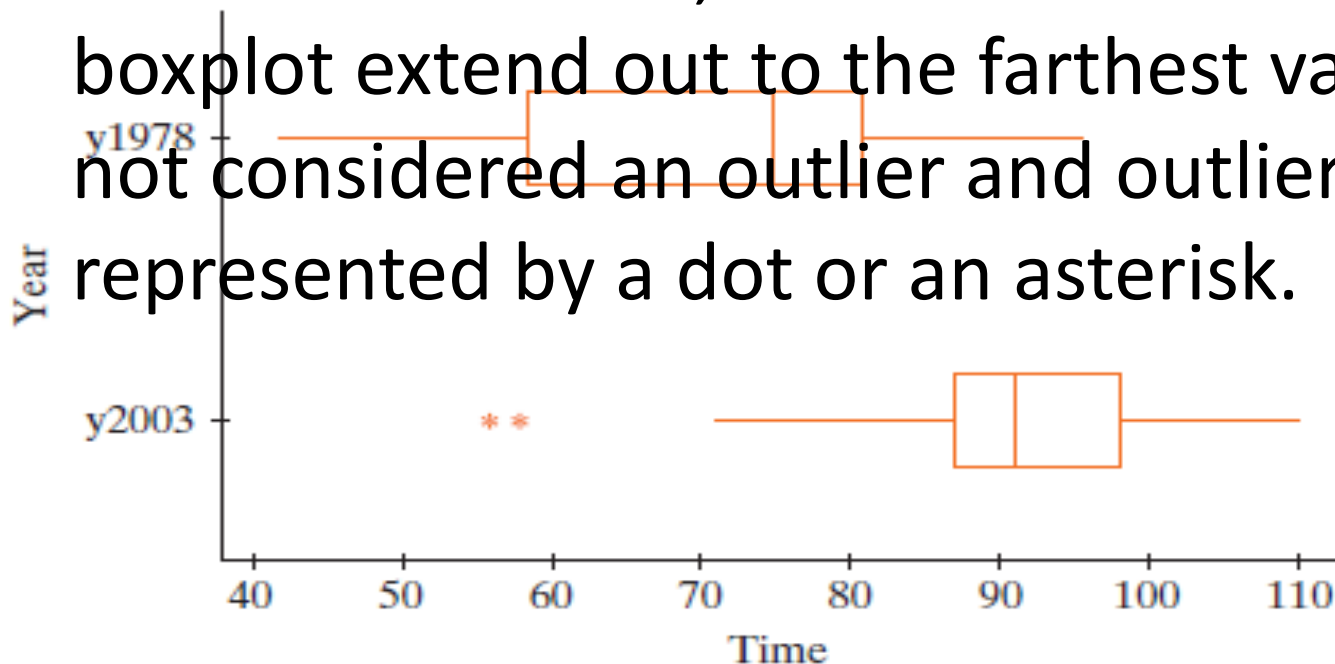
- $1978 \text{ IQR} = 81 - 58 = 23$
- $2003 \text{ IQR} = 98 - 87 = 11$

Boxplots



Boxplots (Outliers)

- A data value that is more than $1.5 \times \text{IQR}$ above the upper quartile or below the lower quartile is considered an outlier.
- When these occur, the whiskers on a boxplot extend out to the farthest value not considered an outlier and outliers are represented by a dot or an asterisk.

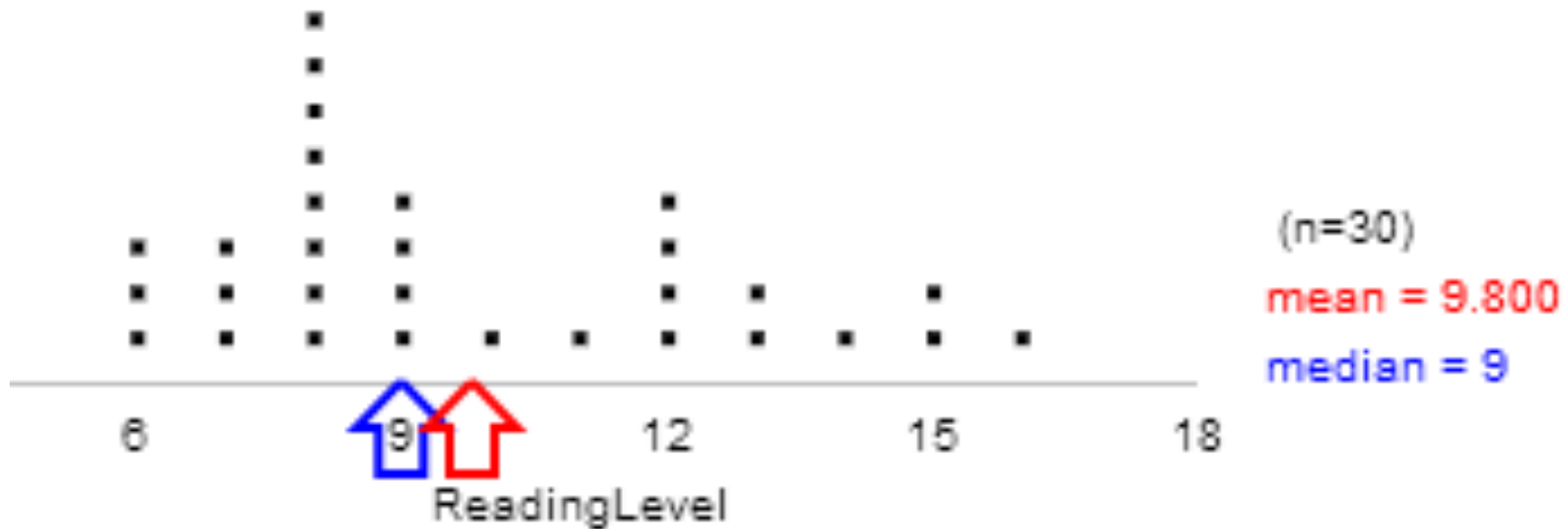


Cancer Pamphlet Reading Levels

- Short et al. (1995) compared reading levels of cancer patients and readability levels of cancer pamphlets. What is the:
 - Median reading level?
 - Mean reading level?
- Are the data skewed one way or the other?

Pamphlets' readability levels	6	7	8	9	10	11	12	13	14	15	16	Total
Count (number of pamphlets)	3	3	8	4	1	1	4	2	1	2	1	30

- Skewed a bit to the right
- Mean to the right of median



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

Section 6.2

Comparison with proportions.

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

Bicycling to Work

Example 6.2

Bicycling to Work

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

Bicycling to Work

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

Bicycling to Work

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

Bicycling to Work

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

Bicycling to Work

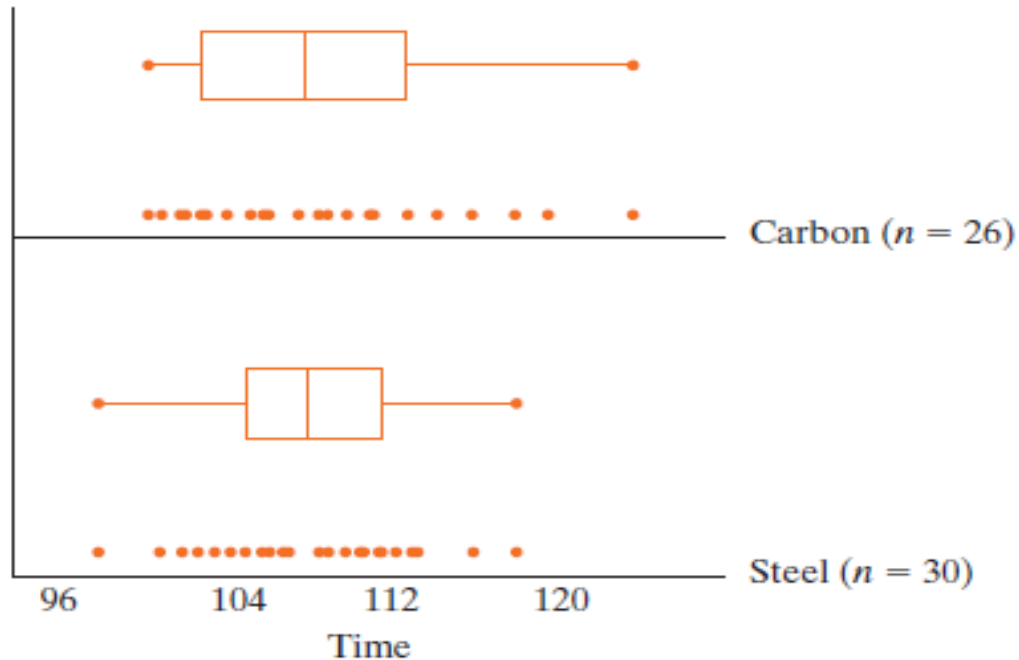
- μ is the population mean. It is a parameter.
- Using the symbols μ_{carbon} and μ_{steel} , we can restate the hypotheses.
- **H_0 :** $\mu_{\text{carbon}} = \mu_{\text{steel}}$
- **H_a :** $\mu_{\text{carbon}} \neq \mu_{\text{steel}}$.

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 average and variability for commute time was higher for the carbon frame bike
- Does this indicate a tendency?
- 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

- Is it *possible* to get a difference of 0.53 minutes if commute time isn't affected by the bike used?
- The same type of question was asked in Chapter 5 for categorical response variables.
- The same answer. Yes it's possible, how likely though?

Bicycling to Work

- The 3S Strategy

Statistic:

- Choose a statistic:
- The observed difference in average commute times

$$\begin{aligned}\bar{x}_{\text{carbon}} - \bar{x}_{\text{steel}} &= 108.34 - 107.81 \\ &= 0.53 \text{ minutes}\end{aligned}$$

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
- Let's see what this looks like

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				

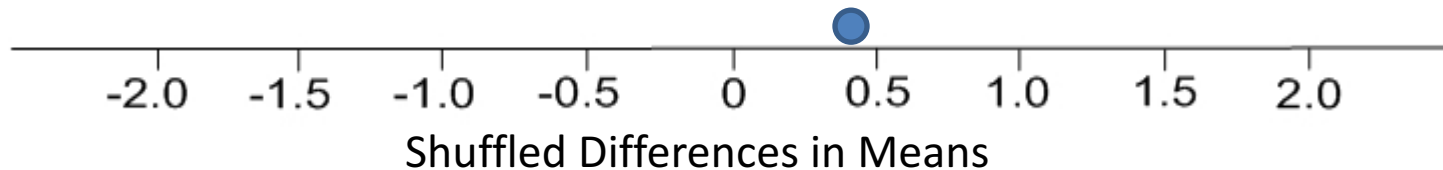
mean = 108.27

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

$$108.27 - 107.87 = 0.40$$



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				

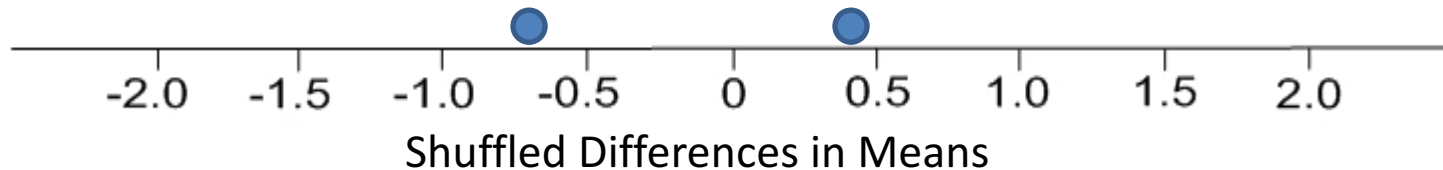
mean = 108.37

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 = 108.87

$$107.69 - 108.37 = -0.68$$



Carbon Frame

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

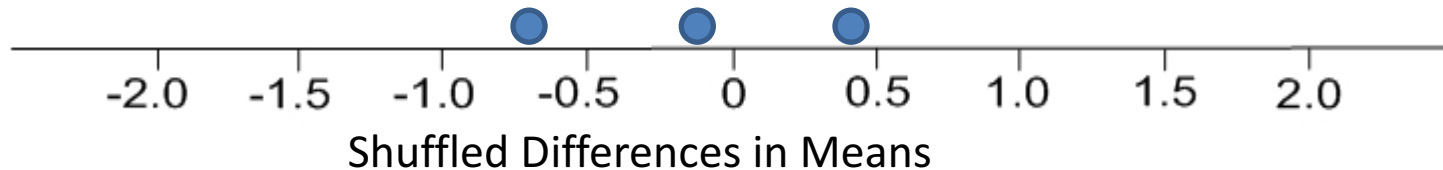
mean = 107.97

Steel Frame

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

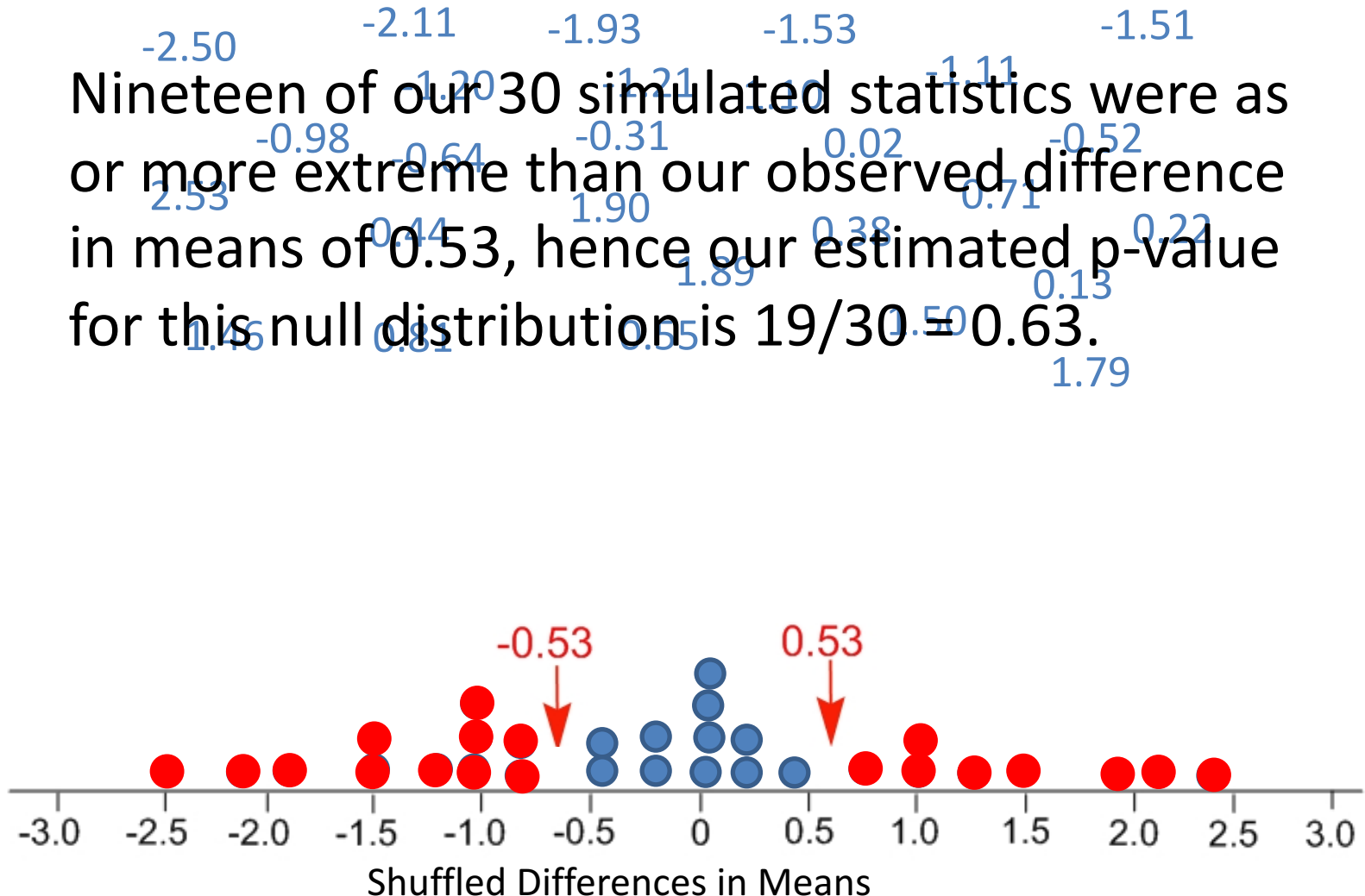
mean = 108.13

$$107.97 - 108.13 = -0.16$$



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 lighter bike to 26 days and the heavier to 30 days, a difference as extreme as 0.53 minutes or more would occur in about 72% of the repetitions.
- Therefore, we do not have strong evidence that the commute times for the two bikes will differ in the long run. The difference observed by Dr. Groves is not statistically significant.

Bicycling to Work

- Have we proven that the bike Groves chooses is not associated with commute time? (Can we conclude the null?)
 - No, a large p-value is not “strong evidence that the null hypothesis is true.”
 - It suggests that the null hypothesis is plausible
 - There could be a small long-term difference. But there also could be no 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 $\pm 1.96 \times \text{SE}$ for the difference between the two means
- From simulations, the SE = standard deviation of the differences = 1.47.
- $0.53 \pm 1.96(1.47) = 0.53 \pm 2.88$
- -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

Scope of conclusions

- Can we generalize our conclusion to a larger population?
- Two Key questions:
 - Was the sample randomly obtained and representative of the overall population of interest?
 - Was this an experiment? Were the observational units randomly assigned to treatments?

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.