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

1. 5 number summary, IQR, boxplots, and Geysers.
2. Comparing two means with simulations, and the bicycling example.
3. Comparing two means with theory based t test, and breastfeeding and intelligence example.
4. Paired data, studying with music, running bases,
5. When to use which formula.

Read through ch7.

# 1. 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

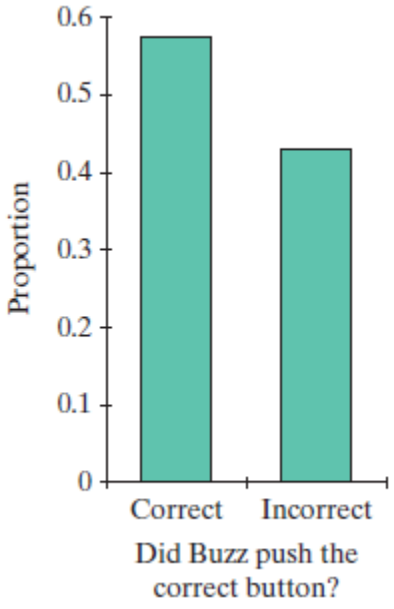
*Section 6.1*

# 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…

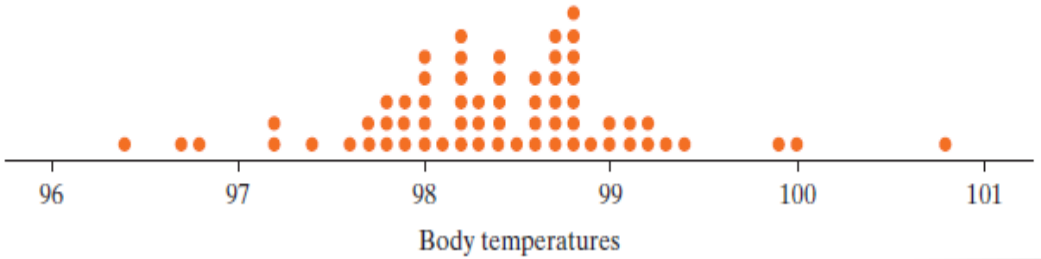
# Graphs for a Single Variable

Categorical




Bar Graph

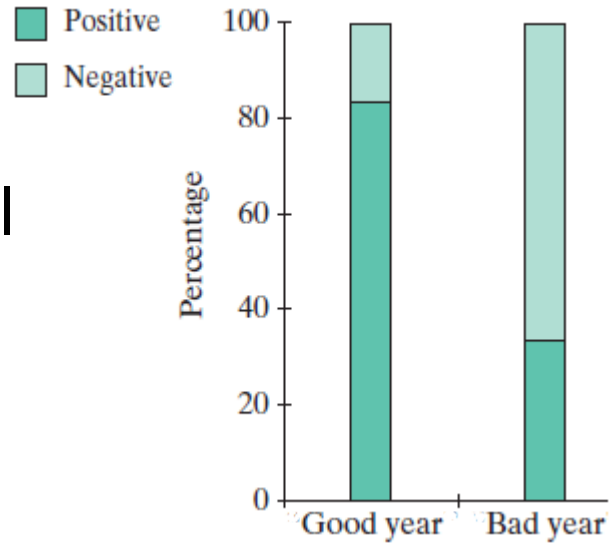
Quantitative



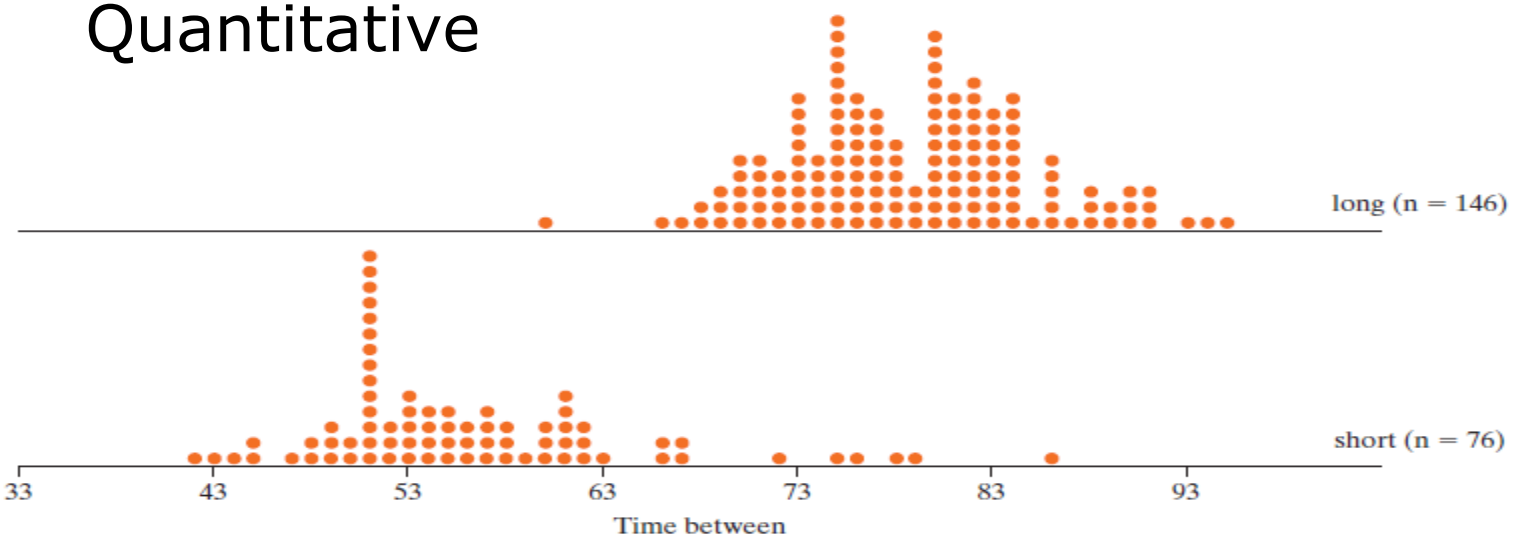

Dot Plot

# Comparing Two Groups Graphically

Categorical




Quantitative



# Notation Check

**Statistics**

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

**Parameters**

- $\mu$ Population mean
- $\pi$ 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 $25^{th}$ percentile).
- Similarly, if 25% of the observations are greater than x, then x is called the ***upper quartile*** (or $75^{th}$ 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 median$\{x_i : x_i >$ 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, $75^{th}$ percentile), but the difference $75^{th}$ percentile – $25^{th}$ .
- 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, $25^{th}$ percentile = 5, $75^{th}$ 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.
- IQR(x,type=2) ## 5.5.
- IQR(x,type=1) ## 6. Can you see why?

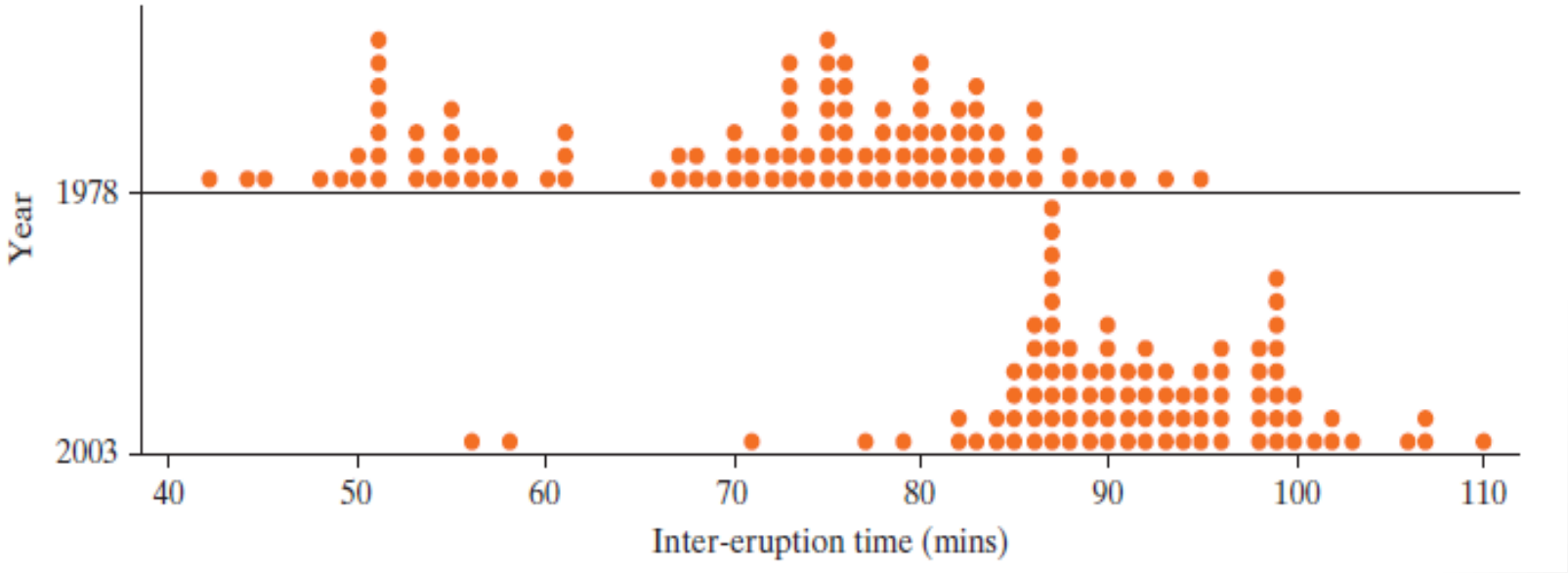
- For example, suppose data are 1, 3, 7, 7, 8, 9, 12, 14.
- M = 7.5, $25^{th}$ percentile = 5, $75^{th}$ percentile = 10.5. IQR = 5.5.

# Geyser Eruptions

Example 6.1

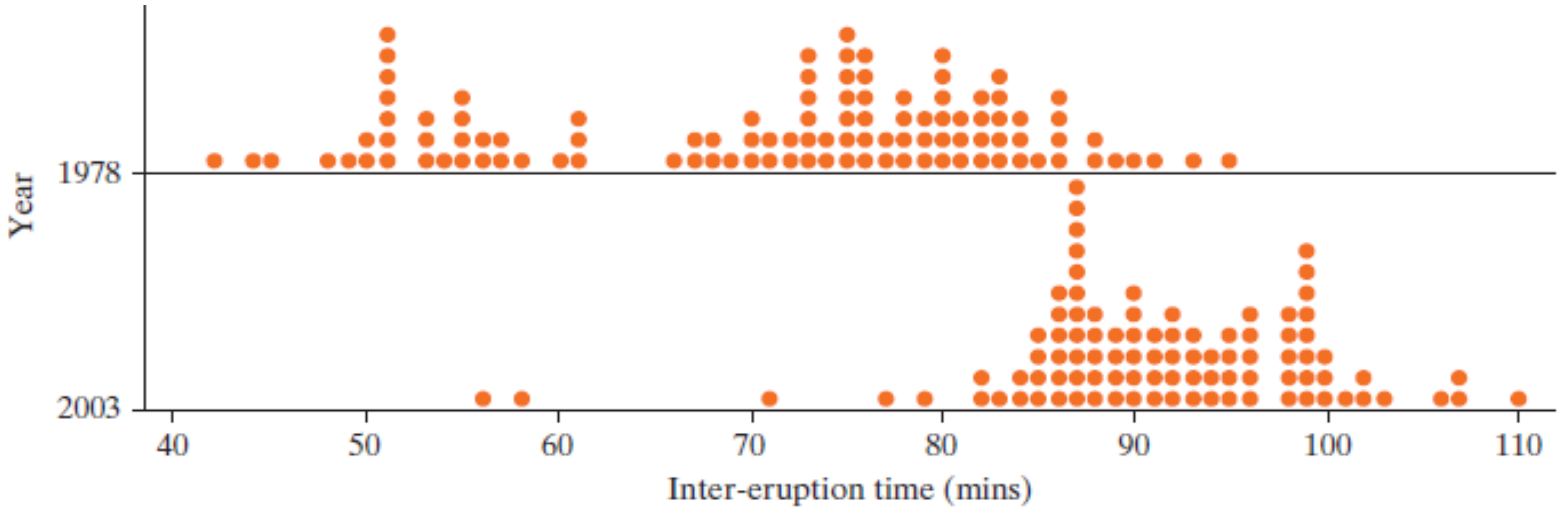
# Old Faithful Inter-Eruption Times

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



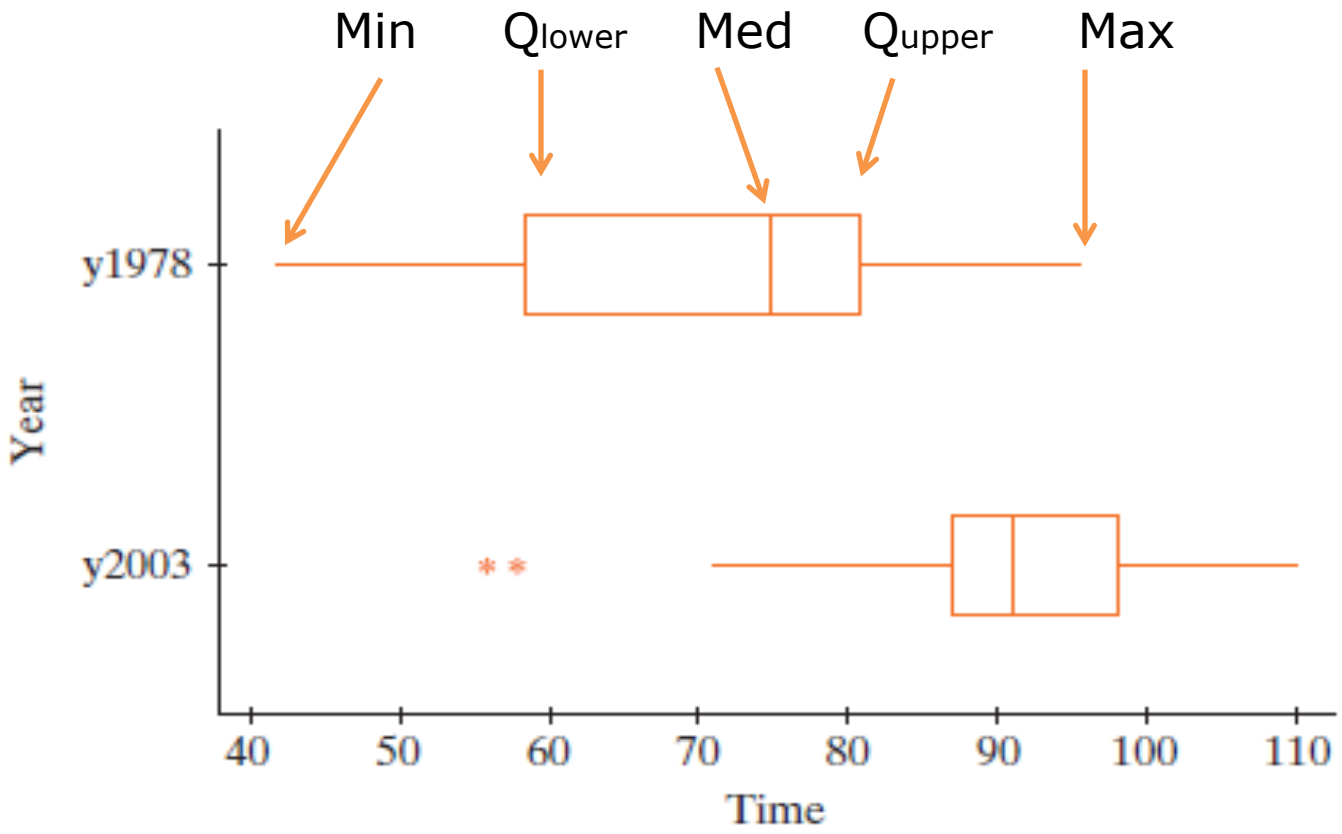
# Old Faithful Inter-Eruption Times

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



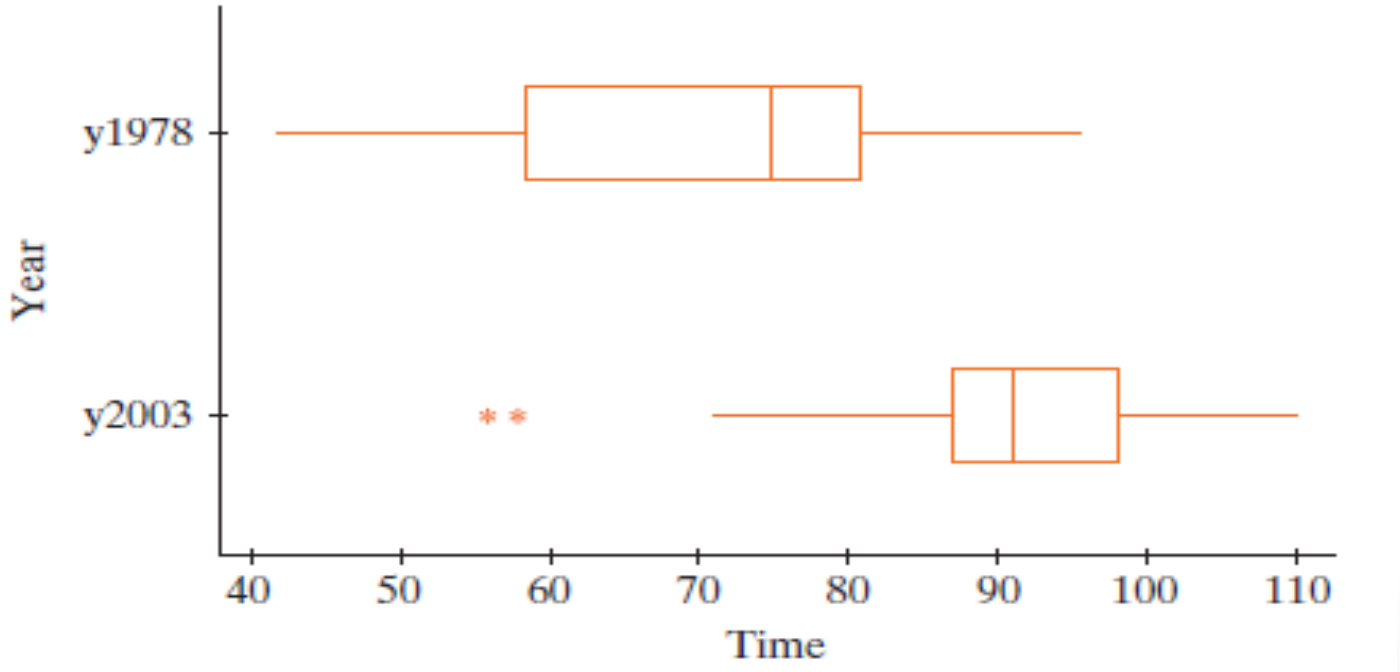

- 1978 IQR = 81 – 58 = 23
- 2003 IQR = 98 – 87 = 11

# Boxplots


Min
Qlower
Med
Qupper
Max
y1978
y2003
Year
40
50
60
70
80
90
100
110
Time

# Boxplots (Outliers)

- A data value that is more than 1.5 × 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.

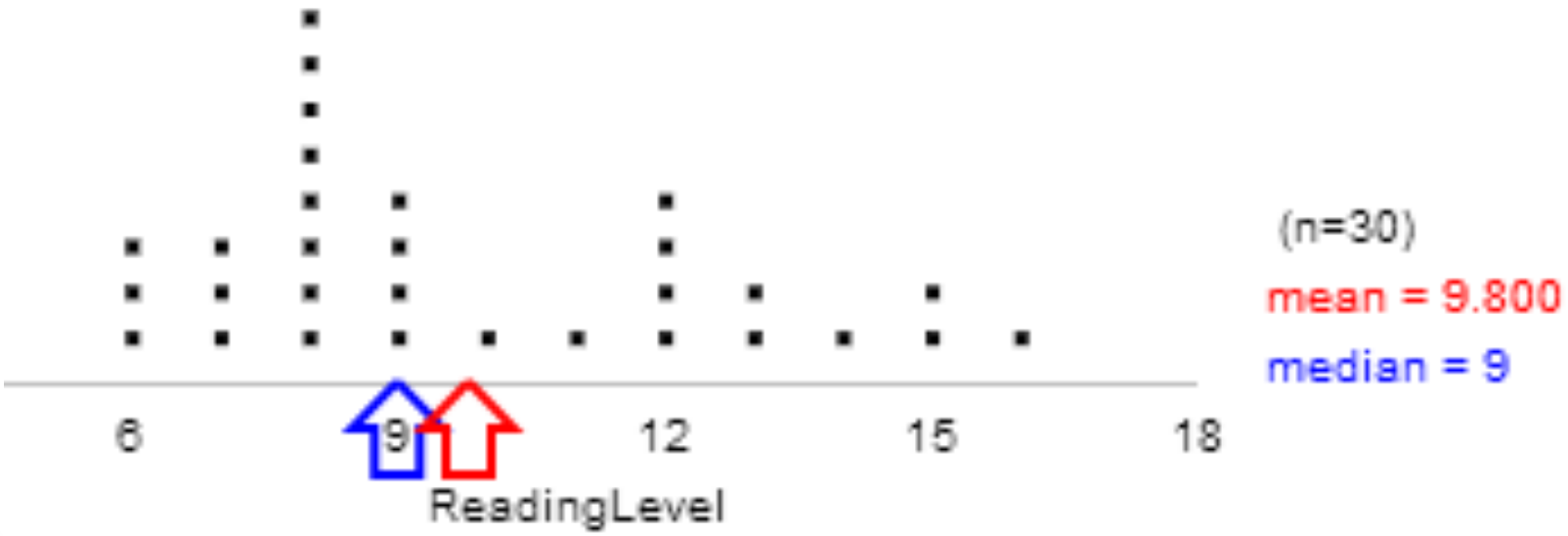


# Cancer Pamphlet Reading Levels

- Short et al. (1995) compared reading levels of cancel 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 > median



# 2. 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).

# 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:
  - $\mu_{carbon}$ = Long term average commute time with carbon framed bike
  - $\mu_{steel}$ = Long term average commute time with steel framed bike.

# Bicycling to Work

- $\mu$ is the population mean. It is a parameter.
- Using the symbols $\mu_{carbon}$ and $\mu_{steel}$, we can restate the hypotheses.

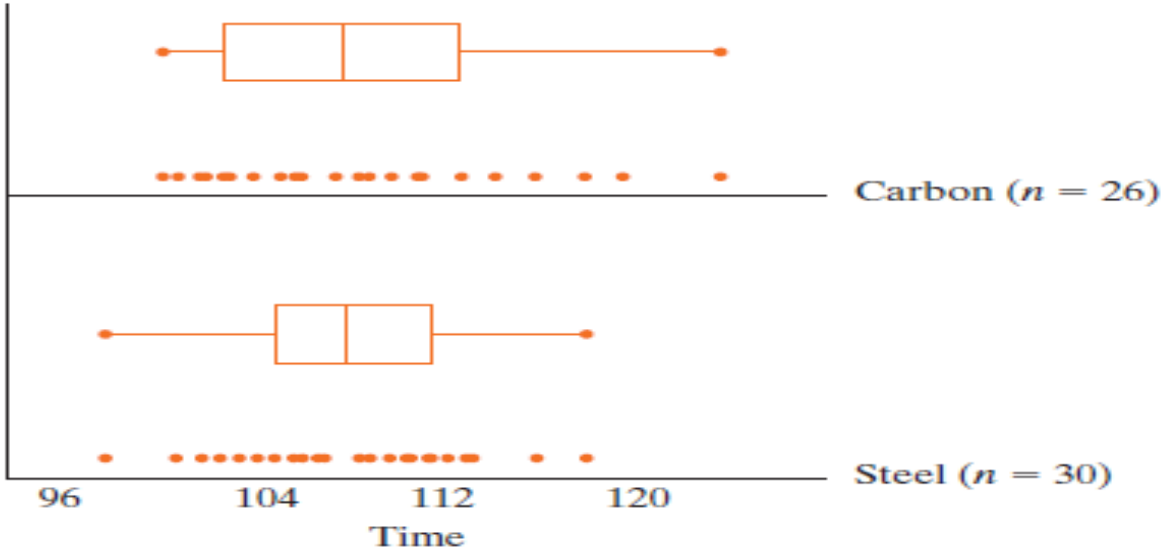
- $\mathbf{H_0}$: $\mu_{carbon} = \mu_{steel}$
- $\mathbf{H_a}$: $\mu_{carbon} \neq \mu_{steel}$ .

# Bicycling to Work

Remember:

- The hypotheses are about the longterm association between commute time and bike used, not just his 56 trips.
- Hypotheses are always about populations or processes, not the sample data.

# Bicycling to Work




| | Sample size | Sample mean | Sample SD |
|---|---|---|---|
| **Carbon frame** | 26 | 108.34 min | 6.25 min |
| **Steel frame** | 30 | 107.81 min | 4.89 min |

The observed difference in average commute times

$\bar{x}_{carbon} - \bar{x}_{steel} = 108.34 - 107.81$

$= 0.53$ minutes

# Bicycling to Work

**Simulation:**

- We can imagine simulating this study with index cards.
  - Write all 56 times on 56 cards.
- Shuffle all 56 cards and randomly redistribute into two stacks:
  - One with 26 cards (representing the times for the carbon-frame bike)
  - Another 30 cards (representing the times for the steel-frame bike)

# Bicycling to Work

**Simulation (continued):**

- Shuffling assumes the null hypothesis of no relationship 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 | | | | |

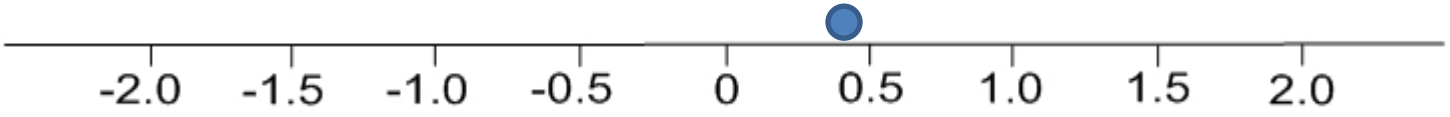
mean = 108.27

# 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




Shuffled Differences in Means

# Carbon Frame

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

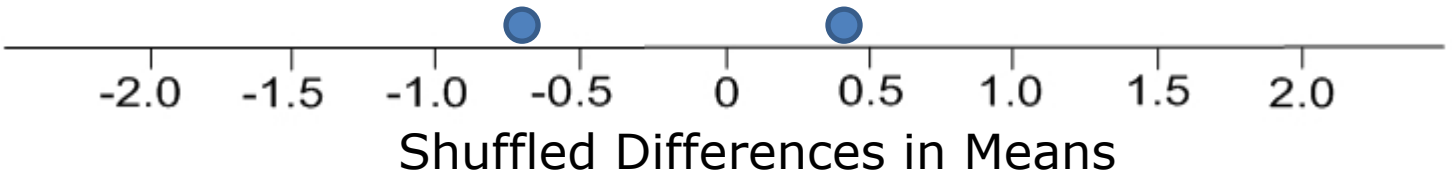
mean = 107.69

# 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 = 108.37

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 | | | | |

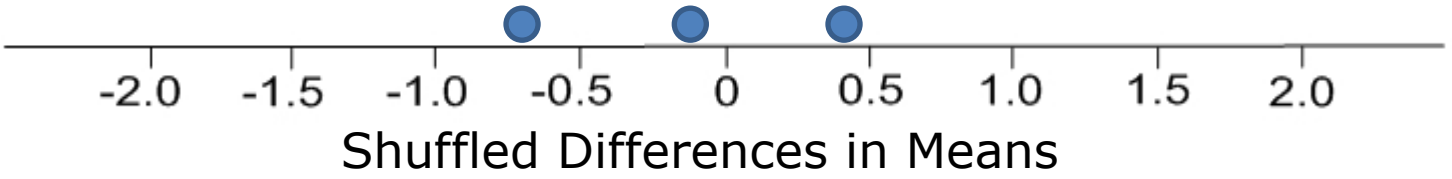
mean = 107.97

# 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 |

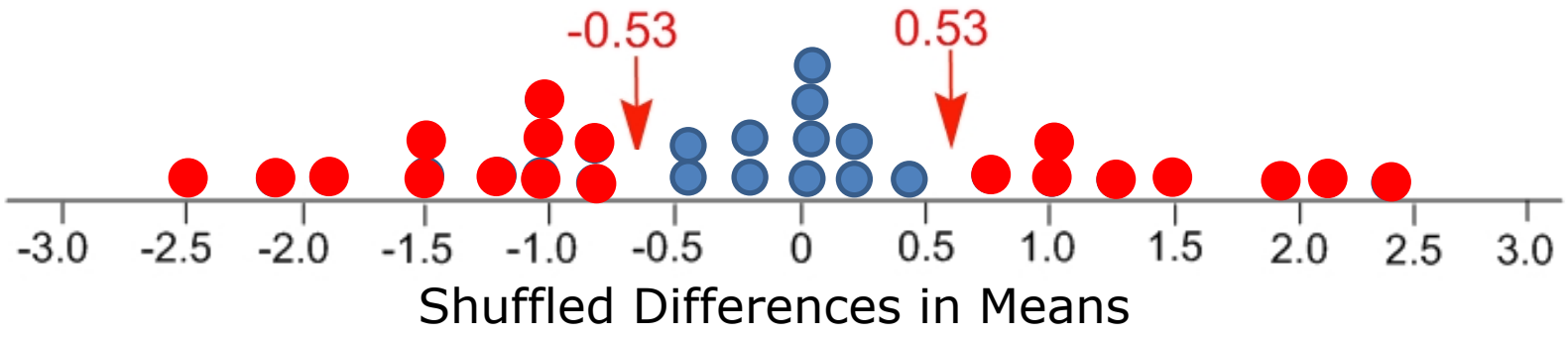
mean = 108.13

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 is true?)
  - No, a large p-value is not “strong evidence that the null hypothesis is true.”
  - The null hypothesis is consistent with the data.
  - 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 ± 1.96×SE for the difference between the two means
- From simulations, the SE = standard deviation of the differences = 1.47. (The theory-based formula is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = 1.52.)
- Using 1.47, we get 0.53 ± 1.96(1.47)= 0.53 ± 2.88
- -2.35 to 3.41.
- What does this mean?

# Bicycling to Work

- We are 95% confident that the true longterm difference (carbon – steel) in average commuting times is between -2.41 and 3.47 minutes.
  The carbon framed bike is between 2.41 minutes faster and 3.47 minutes slower than the steel framed bike.
- Note the interval contains 0.

# 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.

# Breastfeeding and Intelligence

*Example 6.3*

# Breastfeeding and Intelligence

- A 1999 study in *Pediatrics* examined if children who were breastfed during infancy differed from bottle-fed.
- 323 children recruited at birth in 1980-81 from four Western Michigan hospitals.
- Researchers deemed the participants representative of the community in social class, maternal education, age, marital status, and sex of infant.
- Children were followed-up at age 4 and assessed using the General Cognitive Index (GCI)
  - A measure of the child's intellectual functioning
- Researchers surveyed parents and recorded if the child had been breastfed during infancy.

# Breastfeeding and Intelligence

- Explanatory and response variables.
  - **Explanatory variable:** Whether the baby was breastfed. (Categorical)
  - **Response variable:** Baby's GCI at age 4. (Quantitative)

- Is this an experiment or an observational study?
- Can cause-and-effect conclusions be drawn in this study?

# Breastfeeding and Intelligence

- **Null hypothesis:** There is no relationship between breastfeeding during infancy and GCI at age 4.
- **Alternative hypothesis:** There is a relationship between breastfeeding during infancy and GCI at age 4.

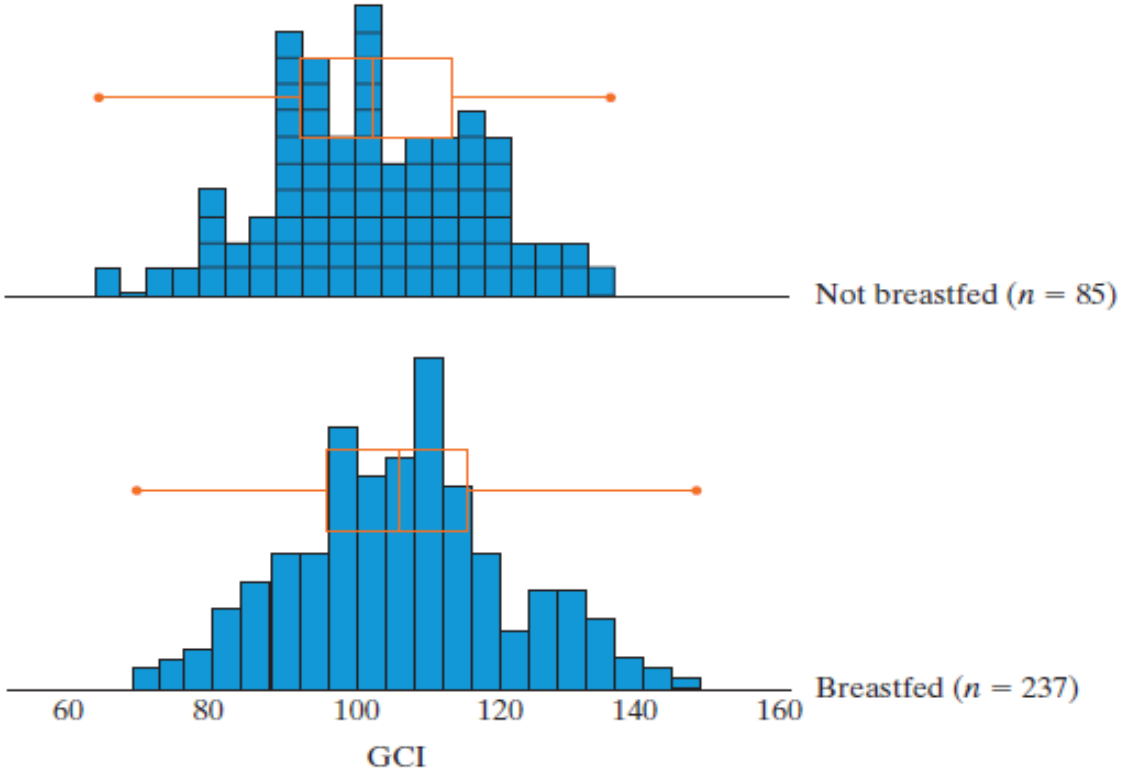
# Breastfeeding and Intelligence

- $\mu_{breastfed}$ = Average GCI at age 4 for breastfed children
- $\mu_{not}$ = Average GCI at age 4 for children not breastfed

- **$H_0$:** $\mu_{breastfed} = \mu_{not}$
- **$H_a$:** $\mu_{breastfed} \neq \mu_{not}$

# Breastfeeding and Intelligence

| Group | Sample size, n | Sample mean | Sample SD |
|---|---|---|---|
| Breastfed | 237 | 105.3 | 14.5 |
| Not BF | 85 | 100.9 | 14.0 |



# Breastfeeding and Intelligence

The difference in means was 4.4.

- If breastfeeding is not related to GCI at age 4:
  - Is it **possible** a difference this large could happen by chance alone? Yes
  - Is it **plausible (believable, fairly likely)** a difference this large could happen by chance alone?
    - We can investigate this with simulations.
    - Alternatively, we can use theory-based methods.

# T-statistic

- If we can assume the draws are iid and the populations are normal, with unknown sds, then t-statistic is used.
- It is the number of standard deviations our statistic is above or below the mean under the null hypothesis.
- $t = \frac{statistic - hypothesized\ value}{SE} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- Here, t = $\frac{105.3 - 100.9}{\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}} = 2.46.$
- p-value ~ 1.4 or 1.5%. [2 * (1-pnorm(2.46))], or use pt.

# Breastfeeding and Intelligence

Meaning of the p-value:

- If breastfeeding were not related to GCI at age 4, then the probability of observing a difference of 4.4 or more or -4.4 or less just by chance is about 1.4%.

- A 95% CI can also be obtained using the t-distribution. The SE is $\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}$ = 1.79. So the margin of error is multiplier x SE.

# Breastfeeding and Intelligence

- The SE is $\sqrt{(\frac{14.5^2}{237} + \frac{14.0^2}{85})}$ = 1.79. The margin of error is multiplier x SE.
- The multiplier should technically be obtained using the t distribution, but for large sample sizes you get almost the same multiplier with t and normal. Use 1.96 for a 95% CI to get 4.40 +/- 1.96 x 1.79 = 4.40 +/- 3.51 = (0.89, 7.91).
- The book uses 2 instead of 1.96, and the applet uses 1.9756 from the t-distribution. Just use 1.96 for this class.

# Breastfeeding and Intelligence

- We have strong evidence against the null hypothesis and can conclude the association between breastfeeding and intelligence here is statistically significant.
- Breastfed babies have statistically significantly higher average GCI scores at age 4.
- We can see this in both the small p-value (0.015) and the confidence interval that says the mean GCI for breastfed babies is 0.89 to 7.91 points higher than that for non-breastfed babies.

# Breastfeeding and Intelligence

- Can you conclude that breastfeeding improves average GCI at age 4?
  - No.  The study was not a randomized experiment.
  - We cannot conclude a cause-and-effect relationship.
- There might be alternative explanations for the significant difference in average GCI values.
- What might some confounding factors be?

# Breastfeeding and Intelligence

- Can you conclude that breastfeeding improves average GCI at age 4?
  - No.  The study was not a randomized experiment.
  - We cannot conclude a cause-and-effect relationship.
- There might be alternative explanations for the significant difference in average GCI values.
  - Maybe better educated mothers are more likely to breastfeed their children
  - Maybe mothers that breastfeed spend more time with their children and interact with them more.
  - Some mothers who do not breastfeed are less healthy or their babies have weaker appetites and this might slow down development in general.

# Breastfeeding and Intelligence

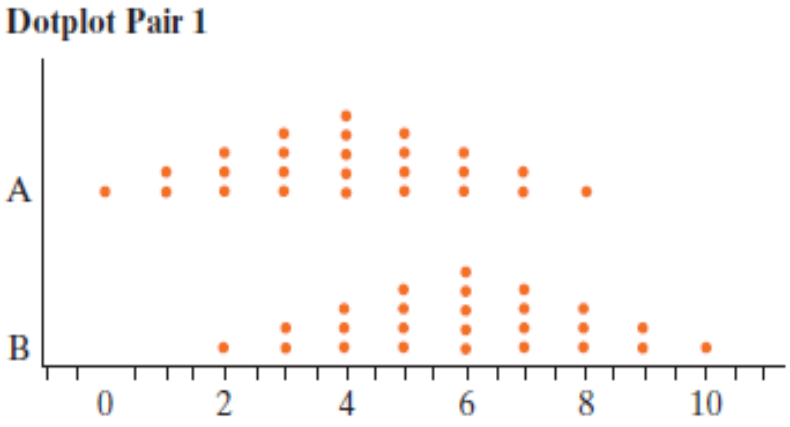
- Could you design a study that allows drawing a cause-and-effect conclusion?
  - We would have to run an experiment using random assignment to determine which mothers breastfeed and which would not. (It would be impossible to double-blind.)
  - Random assignment roughly balances out all other variables.
- Is it feasible/ethical to conduct such a study?

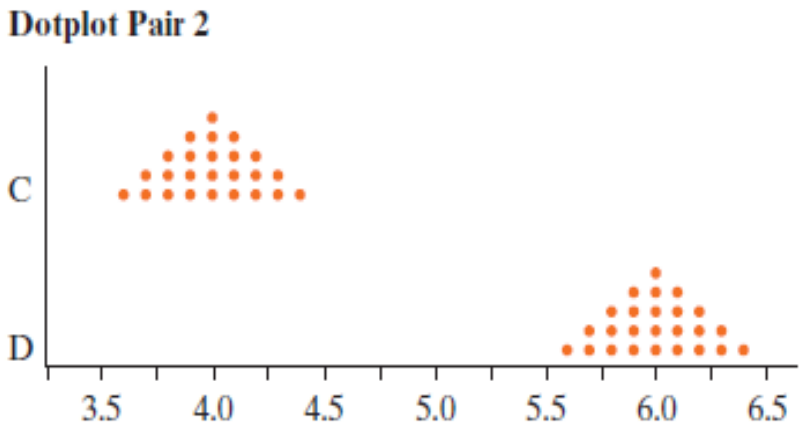
# Strength of Evidence

- We already know:
  - As sample size increases, the strength of evidence increases.
  - Just as with proportions, as the sample means move farther apart, the strength of evidence increases.

# More Strength of Evidence

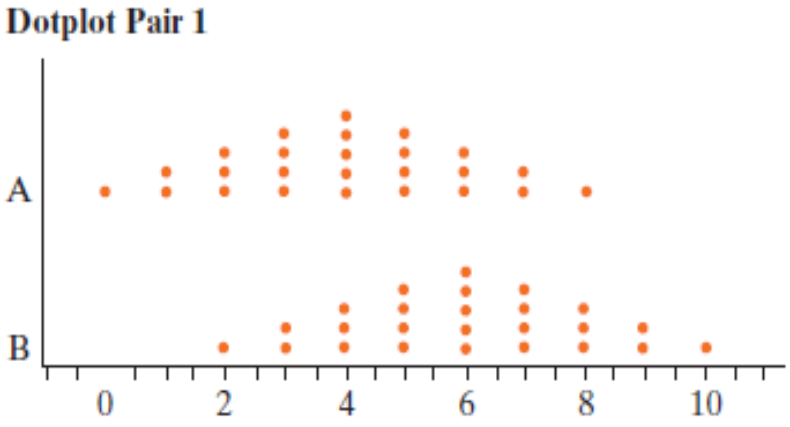
- If the means are the same distance apart, but the standard deviations change, then the strength of evidence changes too.
- Which gives stronger evidence against the null?

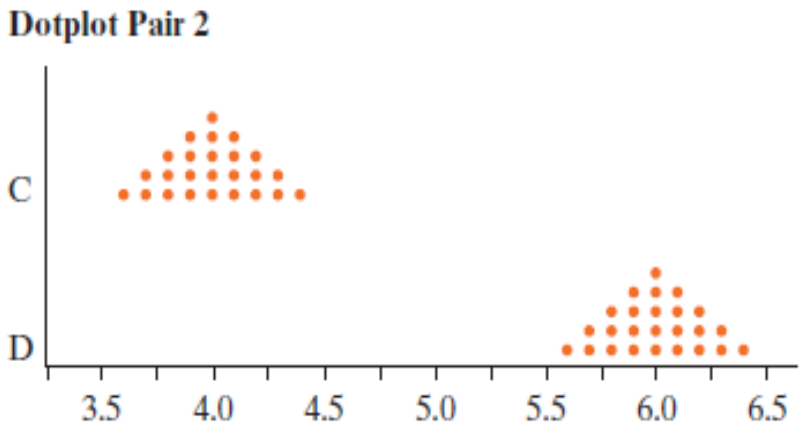


# More Strength of Evidence

- If the means are the same distance apart, but the standard deviations change, then the strength of evidence changes too.
- Which gives stronger evidence against the null?




- Smaller SDs lead to stronger evidence against the null.

# Effects on Width of Confidence Intervals

- Just as before:
  - As sample size increases, confidence interval widths tend to decrease.
  - As confidence level increases, confidence interval widths increase.
  - The difference in means will not affect the width (margin of error) but will affect the center of the CI.
- As we saw with a single mean, as the SDs of the samples increase, the width of the confidence interval will increase.

# Paired Data.

Chapter 7

# Introduction

- The paired data sets in chapter 7 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.
  - The big idea is with paired data, just view the *differences* between each pair of scores as your data. Now you have one variable and can just analyze it using the methods we already know.
  - When you analyze paired data this way, person to person variability gets removed so you get more power when testing, smaller p-values and smaller margins of error.

# 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.

# Simulation-Based Approach for Analyzing Paired Data, and rounding first base example.

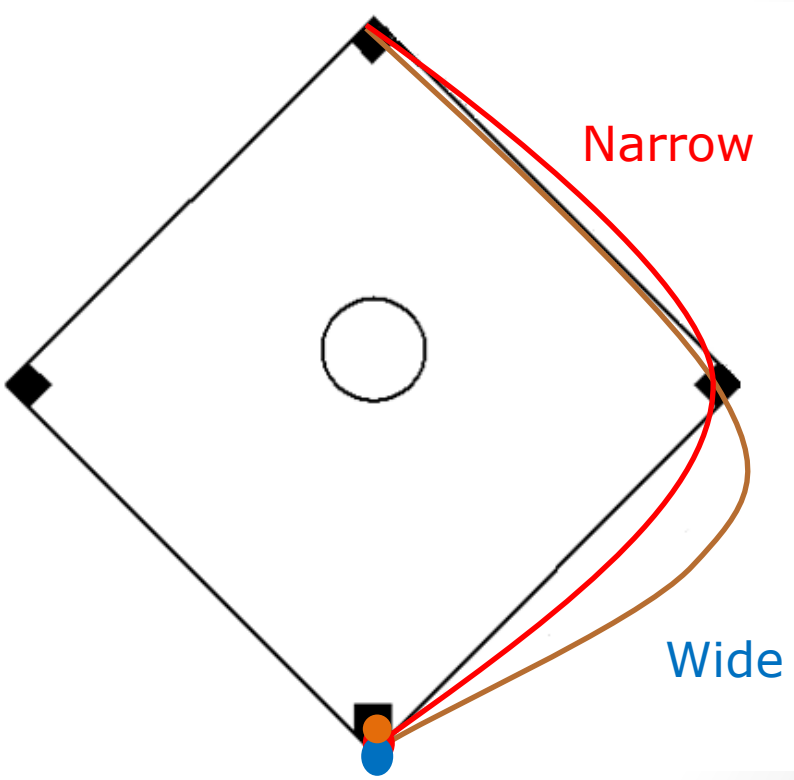
Section 7.2

# 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.**

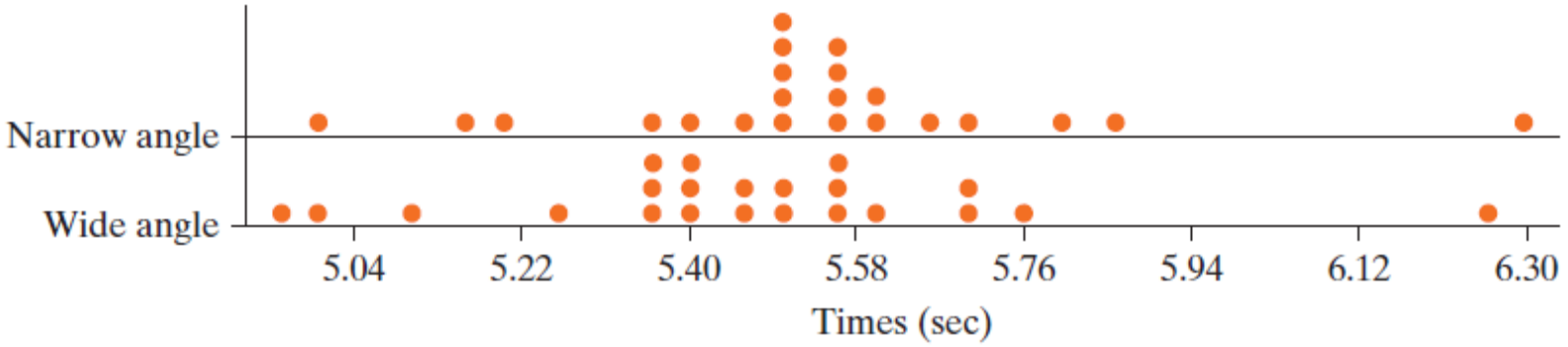
# First Base

- 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 | ... |



# Paired data and rounding first base example.

- There is a lot of overlap in the distributions and substantial variability.

| | Mean | SD |
|---|---|---|
| Narrow | 5.534 | 0.260 |
| Wide | 5.459 | 0.273 |

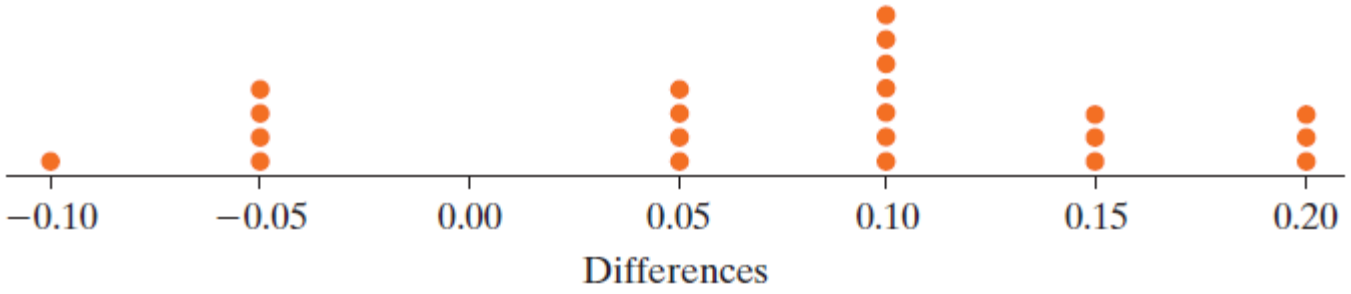
- It is difficult to detect a difference between the methods when these is so much variation.
-

# Rounding First Base

- 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 | … |

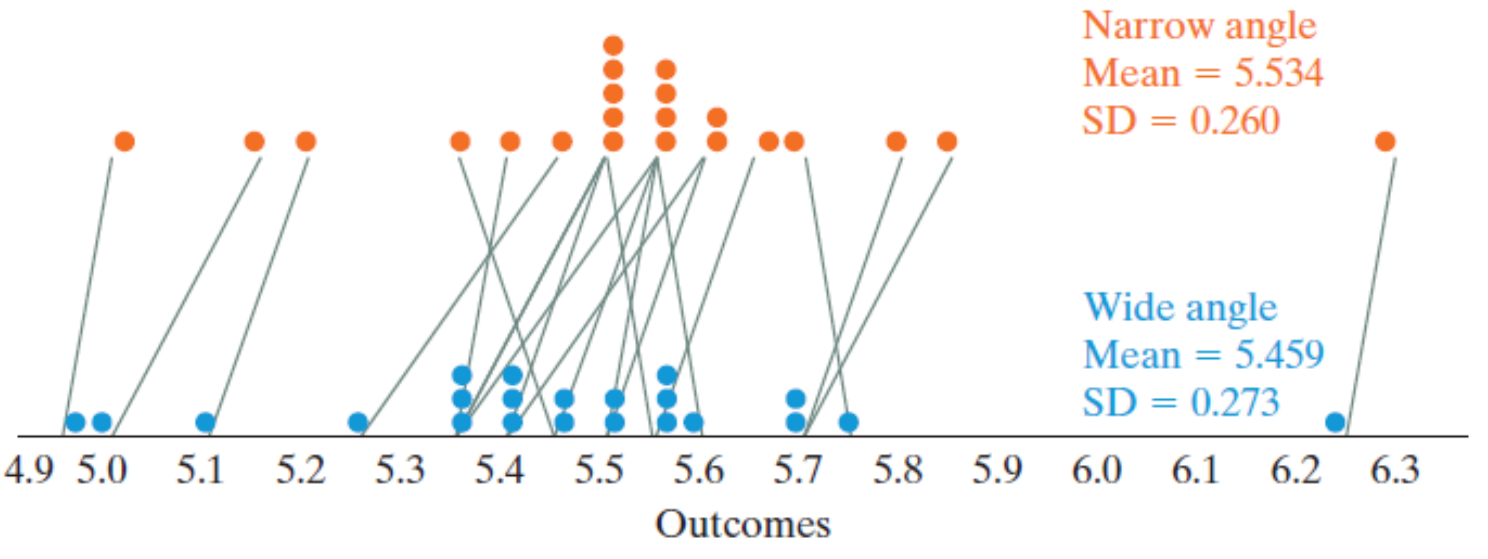


# The Differences in Times

- 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, $\mu_d$, 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

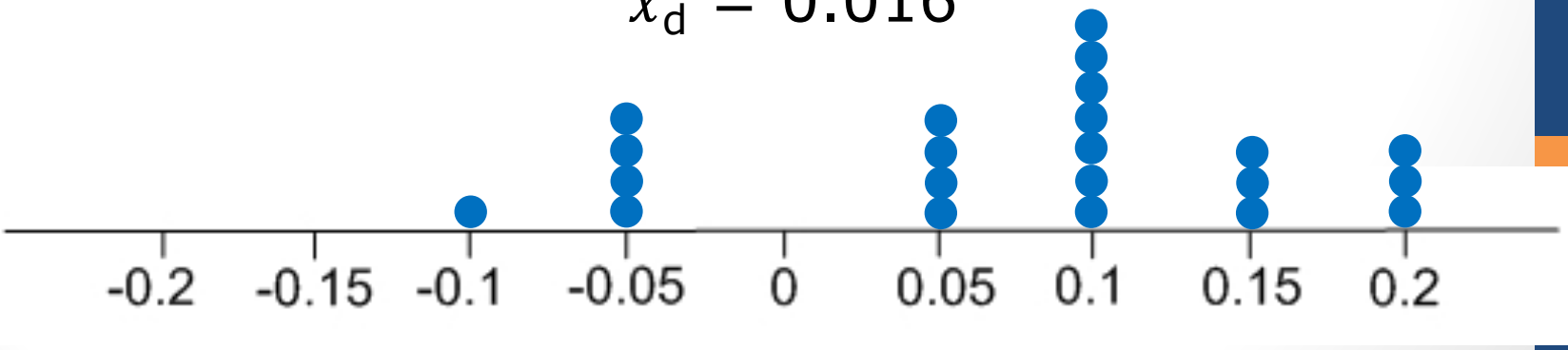
How can we use simulation-based methods to find an approximate p-value?

- 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

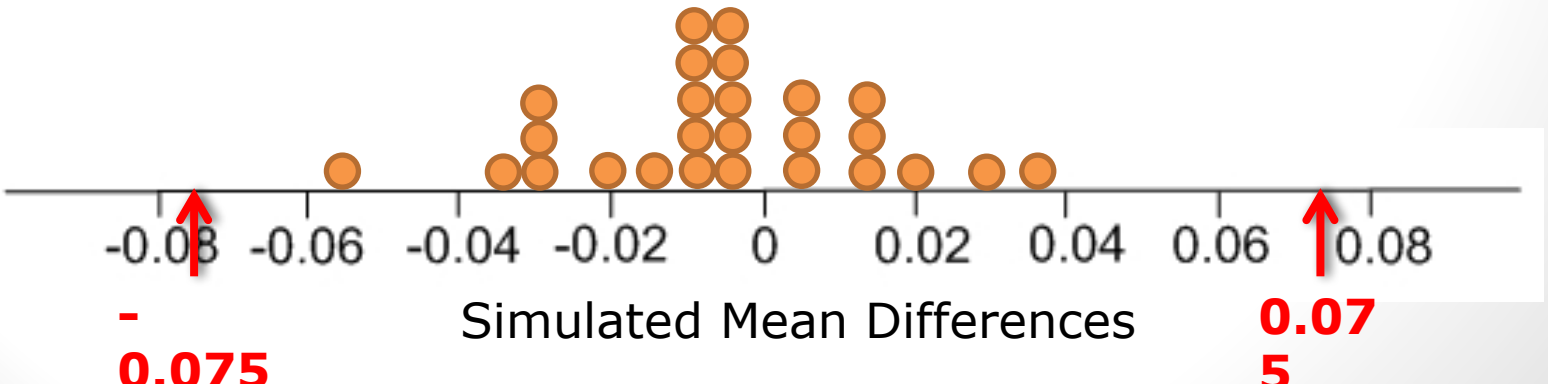
| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| narrow angle | 5.50 | 5.70 | 5.60 | 5.50 | 5.85 | 5.55 | 5.40 | 5.50 | 5.15 | 5.80 | … |
| wide angle | 5.55 | 5.75 | 5.50 | 5.40 | 5.70 | 5.60 | 5.35 | 5.35 | 5.00 | 5.70 | … |
| diff | 0.05 |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

$\bar{x}_d = 0.016$



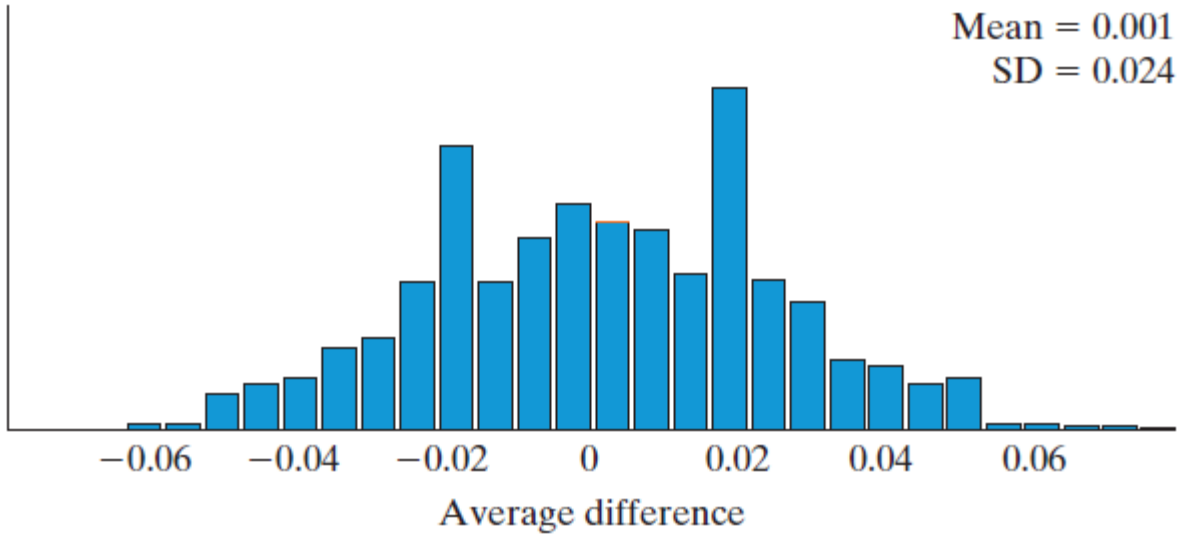
# More Simulations

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



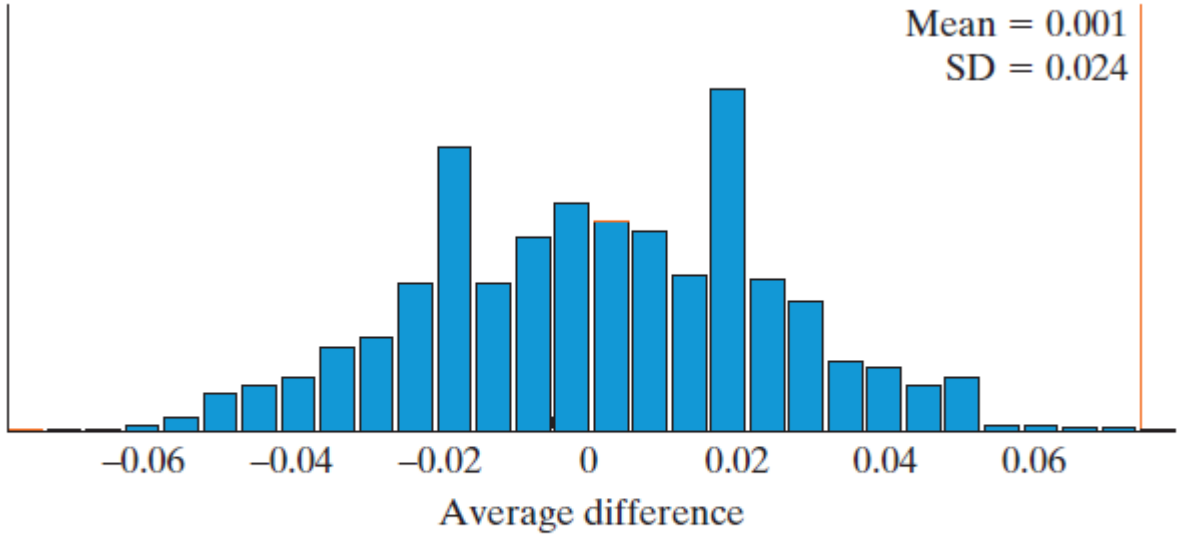
# First Base

- 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?



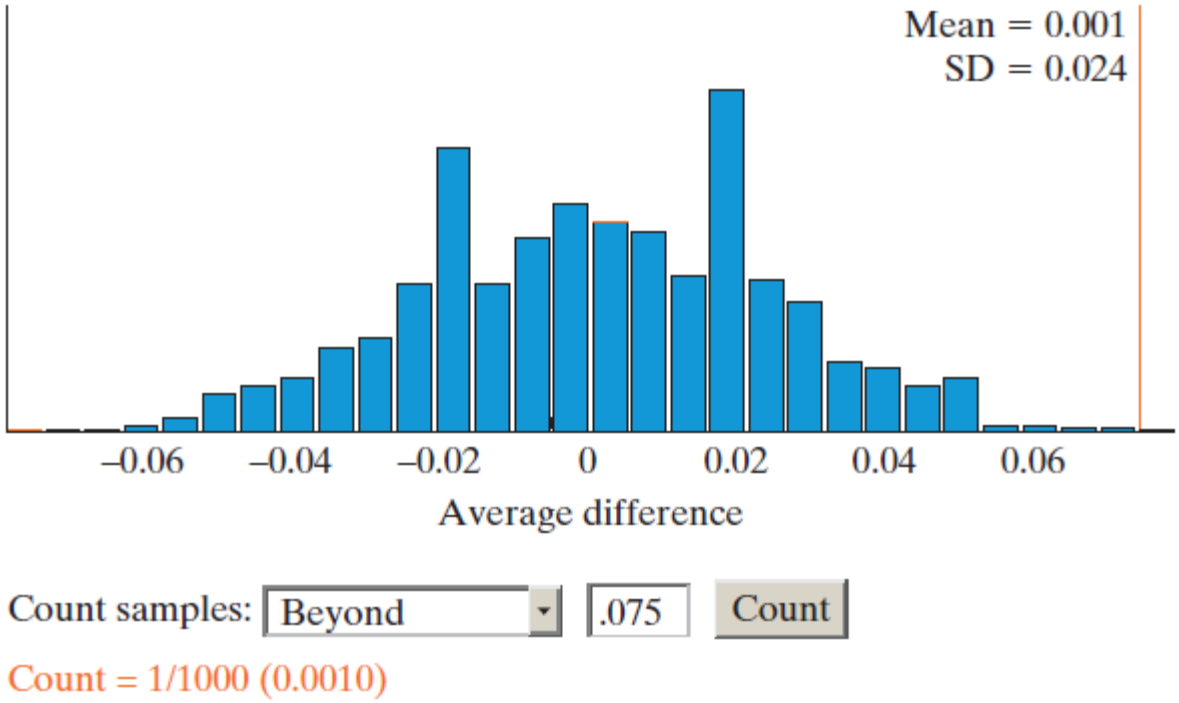
# First Base

- 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.

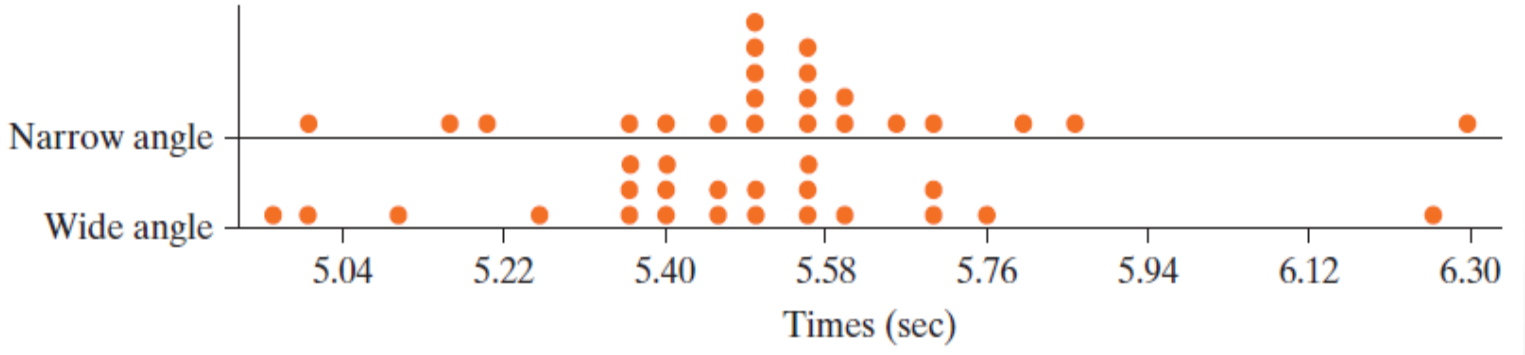
# First Base

- Approximate a 95% confidence interval for $\mu_d$:
  - 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**

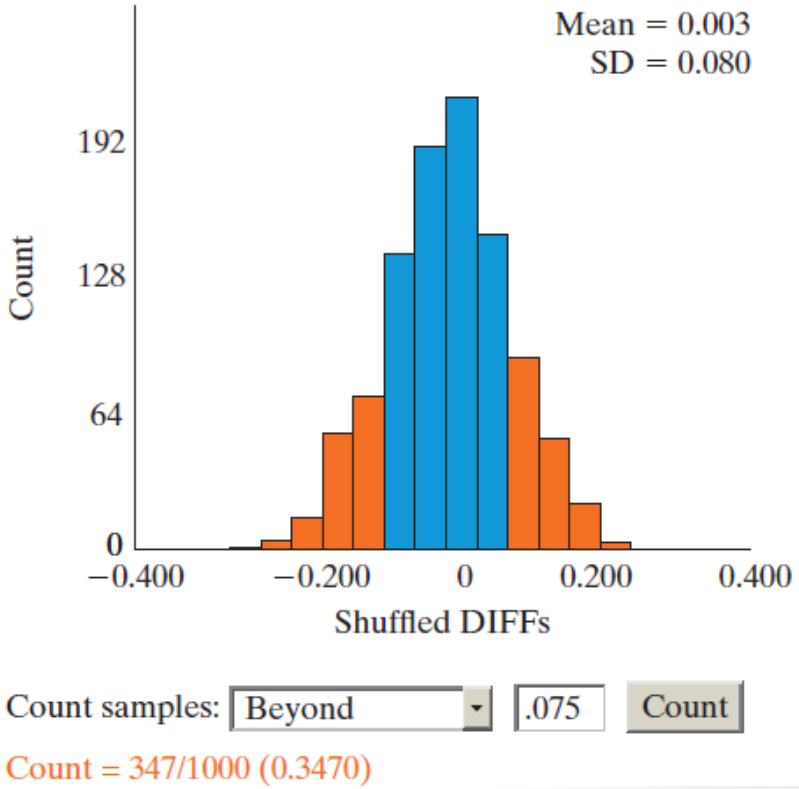
- What do you think would happen if we wrongly analyzed the data using a 2 independent samples procedure? (i.e. The researcher selected 22 runners to use the wide method and an independent sample of 22 other runners to use the narrow method, obtaining the same 44 times as in the actual study.



# First Base

Using an applet which tests a difference between these two means, ignoring the fact that it is paired data, we get a p-value of 0.3470.

This p-value is much larger than the one we obtained earlier, when we ignored the fact that the data were paired.



# 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:

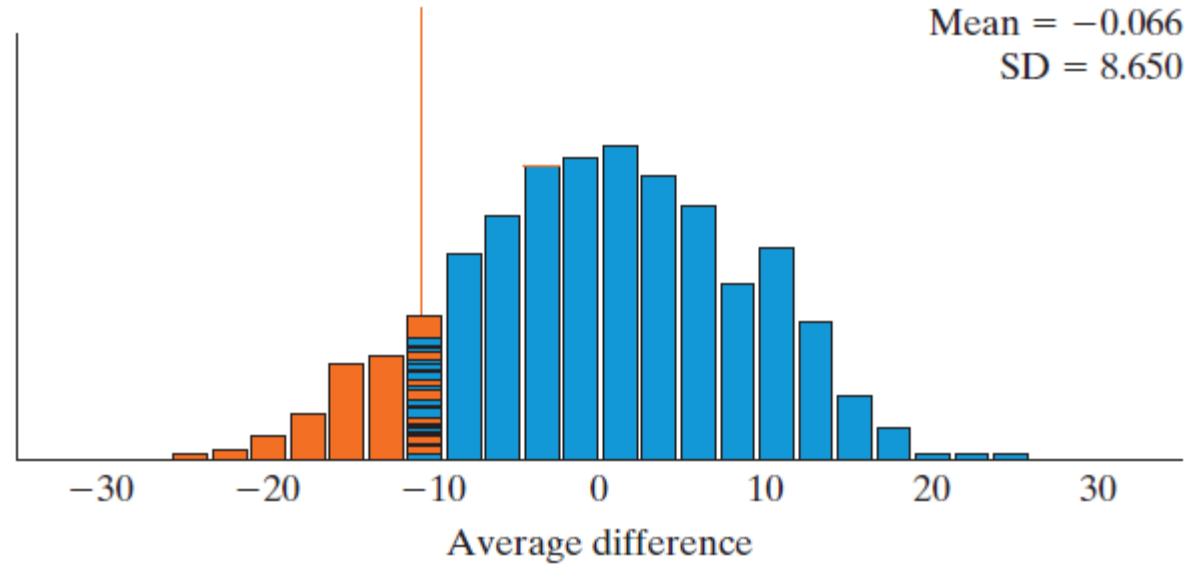
- $H_0$: $\mu_d = 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is 0.
- $H_a$: $\mu_d < 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is less than 0.

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

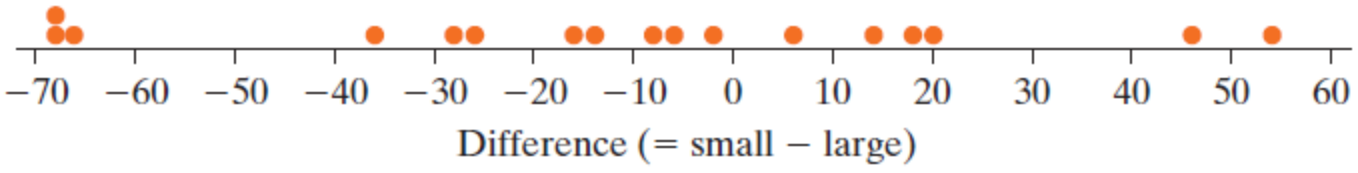
# How Many M&Ms Would You Like?

- 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

- If we can assume the differences for each person are iid and normal with unknown sd, then with a theory based test we calculate the t-statistic:

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

- 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

Calculate

☑ Confidence interval

confidence level 95 % Calculate CI

(−29.5435, 7.7835)

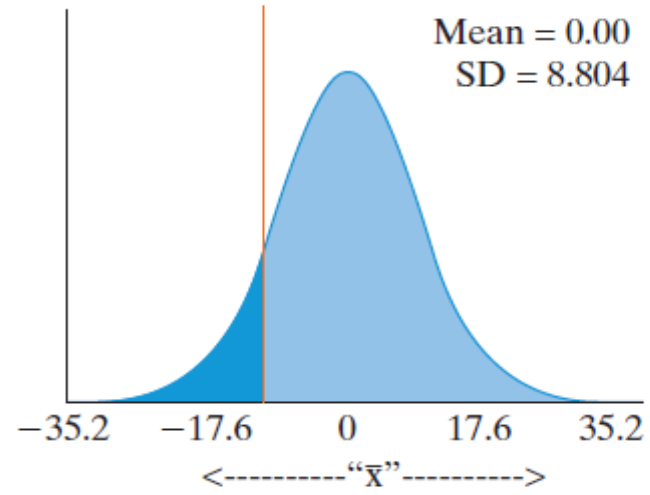
Theory-based inference

☑ 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.

# 5. When to use which formula.

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

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

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

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

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

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

s = $\sqrt{[p(1-p)]}$, $\sigma = \sqrt{[\pi(1-\pi)]}$.

# 5. When to use which formula.

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

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

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

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

s = √[p(1-p)], $\sigma = \sqrt{[\pi(1-\pi)]}$.

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

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

What if n is small and the draws are not normal?

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

# 5. When to use which formula.

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

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

As with one sample, if $\sigma_1$ is known, replace $s_1$ with $\sigma_1$, and the same for $\sigma_2$. And as with one sample, if $\sigma_1$ and $\sigma_2$ are unknown, the sample sizes are small, and the distributions are roughly normal, then use $t_{mult}$ instead of 1.96. If the sample sizes are small, the distributions are normal, and $\sigma_1$ and $\sigma_2$ are known, then use 1.96.

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

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

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

# When to use which formula.

e. Matched pairs data, iid observations, want a 95% CI for μ.

Look at differences (score with treatment minus score with control) and treat differences as ordinary numerical data according to parts a or b.

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

$n \geq 30$ is often considered large enough to use 1.96.

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