

SUMMARIZING QUANTITATIVE DATA:

PART 2

- Median and Quartiles
- Five Number Summary
- Boxplots
- Comparing Histograms and Boxplots
- Independent Comparisons
- Dependent Comparisons



PERCENTILE

- **Percentile:** value at or below which a given percentage of a variable's values fall
 - If the population mean is known, then exact percentiles can be found
 - If we do not know the population mean and only have a sample, then percentiles are only approximations
- To calculate a percentile in a sample by hand:
 - Sort the data from smallest to largest
 - Calculate the location of percentile: $L_p = (n + 1) \times \frac{p}{100}$
 - Count to position L_p in ordered data
 - If L_p is an integer, use that value as the percentile
 - If L_p is not an integer, weight the values around L_p according to the decimal part

EXAMPLE: PERCENTILES

- **Scenario:** Sample of 20 final exam scores from a large class
33 54 59 62 65 67 69 71 73 74 76 77 79 82 83 85 88 91 94 98
- **Question:** What is the 60th percentile?
- **Answer:**
 - Location: $L_{60} = \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$
 - 12th and 13th Largest Scores: $\underline{\hspace{2cm}}$
 - 60th Percentile: $P_{60} = \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$
- **Question:** What is the interpretation of this value?
- **Answer:** $\underline{\hspace{1cm}}$ of students scored $\underline{\hspace{2cm}}$ or equal to $\underline{\hspace{1cm}}$ while $\underline{\hspace{1cm}}$ scored $\underline{\hspace{2cm}}$ or equal to $\underline{\hspace{1cm}}$.

SPECIAL PERCENTILES

4

- **Median:** measure of center equal to the 50th percentile
 - Half of observations lie on either side
 - Odd sample size: Middle number
 - Even sample size: Average of middle two numbers
- **1st quartile (Q_1):** value equal to the 25th percentile
 - 25% of data is smaller while 75% is larger
- **3rd quartile (Q_3):** value equal to the 75th percentile
 - 75% of data is smaller while 25% is larger
- **Quicker Method:** Rather than calculating percentiles by hand, we can use the `summary` function in R.

SPECIAL PERCENTILES

5

- **Scenario:** Sample of 20 final exam scores from a large class

33	54	59	62	65	67	69	71	73	74	76	77	79	82	83	85	88	91	94	98
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

- **Question:** What are the the median, 1st quartile, and 3rd quartile?

- **R Code:** `summary(scores$score, quantile.type = 6)`

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
33.0	65.5	75.0	74.0	84.5	98.0

- **Answer:**
 - Median: _____
 - First Quartile: _____
 - Third Quartile: _____

MEASURES OF SPREAD

6

- **Range:** difference between maximum and minimum values
 - **Drawback:** One outlier can drastically skew the perception of how spread out the data is
- **Interquartile range:** measures the spread of the middle 50% of the data
 - $IQR = Q_3 - Q_1$
 - Difference between third and first quartiles
 - **Benefit:** Ignores outliers

EXAMPLE: RANGE AND IQR

7

- **Scenario:** Sample of 20 final exam scores from a large class

33 54 59 62 65 67 69 71 73 74 76 77 79 82 83 85 88 91 94 98

- **Question:** What is the range?

- **Answer:** _____ = _____

- **Question:** What is the IQR?

- **Answer:** _____ = _____

- **Question:** Why does the range give us the impression that the spread of exam scores is so large?

- **Answer:** One _____ observation (_____)

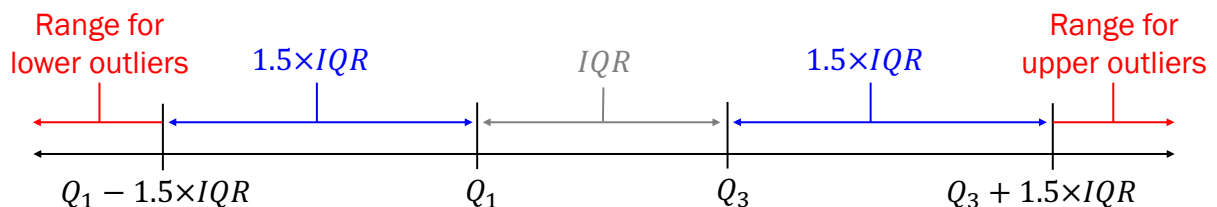
OUTLIERS

8

- **Outlier:** an observation that is unusually far away from the rest of the data

- **Rule of Thumb #2:** Observations are considered outliers if they are either:

- Less than $Q_1 - 1.5 \times IQR$
- Greater than $Q_3 + 1.5 \times IQR$



EXAMPLE: OUTLIERS

9

- **Scenario:** Sample of 20 final exam scores from a large class

33 54 59 62 65 67 69 71 73 74 76 77 79 82 83 85 88 91 94 98

- **Question:** Are there any outliers?

- **Answer:** Yes

- **Lower Cutoff:** $Q_1 - 1.5 \times IQR = \text{_____} = \text{_____}$
- **Upper Cutoff:** $Q_3 + 1.5 \times IQR = \text{_____} = \text{_____}$
- _____ is the only outlier

FIVE NUMBER SUMMARY AND BOXPLOT

10

- **Five number summary:** quick, straightforward way of summarizing a set of quantitative data that consists of the
 - Minimum, 1st quartile, median, 3rd quartile, and maximum
- **Boxplot:** graphical display of quantitative data that displays the five number summary and outliers
 - **Bottom whisker:** Ranges from smallest non-outlier to first quartile
 - **Bottom of box:** Drawn at first quartile
 - **Line through box:** Drawn at median
 - **Top of box:** Drawn at third quartile
 - **Top whisker:** Ranges from third quartile to largest non-outlier
 - **Outliers:** represented by asterisks (*)

EXAMPLE: BOXPLOT

11

- **Scenario:** Sample of 20 final exam scores from a large class

33 54 59 62 65 67 69 71 73 74 76 77 79 82 83 85 88 91 94 98

The five number summary is 33, 65.5, 75, 84.5, 98.

- **Task:** Construct the boxplot.
- **Answer:**

EXAMPLE: CHANGING AN OBSERVATION

12

- **Scenario:** Suppose the student in the sample who scored 33 was replaced by a student who scored 63.
- **Question:** What would happen to the mean, median, and standard deviation?
- **Answer:**
 - **Mean:** _____
 - **Standard Deviation:** _____ (Observations now _____)
 - **Median:** _____ (_____ of observations on either side of 75)

54 59 62 63 65 67 69 71 73 74 76 77 79 82 83 85 88 91 94 98

COMPARING GRAPHS AND MEASURES OF CENTER

13

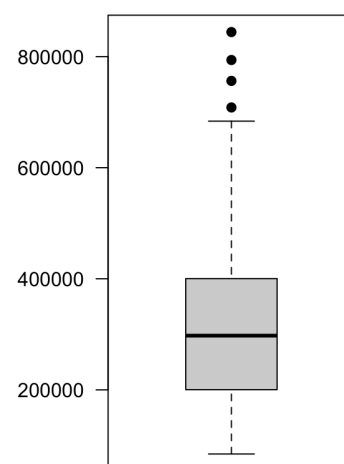
Histogram	Boxplot	Mean vs. Median
Approximately symmetric	Whiskers about same length	Mean $\approx$ Median
Left-skewed	Lower whisker longer	Mean < Median
Right-skewed	Upper whisker longer	Mean > Median

- **Note:** Use the mean to describe the center unless the data is severely skewed.
- **Question:** Why does the skewness impact the relationship between the mean and median?
- **Answer:**
 - If the data is _____-skewed, small values pull the mean _____
 - If the data is _____-skewed, large values pull the mean _____

EXAMPLE: COMPARING GRAPHS AND MEASURES OF CENTER

14

- **Scenario:** Boxplot below displays assessed values of 188 houses in the same city.
- **Question:** What would the shape of the histogram be?
- **Answer:** _____
 - Many _____ assessed values
 - _____ values tend to be _____ spread out
- **Question:** What is the relationship between the mean and median?
- **Answer:** _____
 - Mean: \$ _____; Median: \$297,618
 - _____ observations in _____ tail pull mean _____



COMPARING ONE QUANTITATIVE AND ONE CATEGORICAL

15

- Three situations may arise when comparing the relationship between one quantitative and one categorical variable:
 - Two-sample comparison
 - Several sample comparison
 - Paired sample comparison
- To compare the distributions of the quantitative variable across levels of a categorical variable:
 - **Graphically:** Side-by-side boxplots
 - **Numerically:**
 - Five number summaries
 - Means and standard deviations

EXAMPLE: TWO-SAMPLE COMPARISON

16

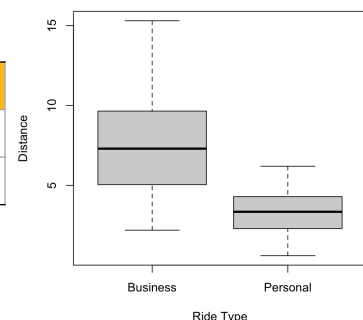
- **Scenario:** Uber driver tracks the lengths of his passengers' rides for one month and whether the passenger took the ride for business or personal reasons.
- **Question:** Which variable is of each type?
- **Answer:**
 - Quantitative: _____
 - Categorical: _____
- **Question:** Do the observations in the categories overlap?
- **Answer:** _____
 - Each ride is a _____ and can only be taken for _____ of the two reasons
- **Takeaway:** These two samples are said to be _____.

EXAMPLE: TWO-SAMPLE COMPARISON

17

- **Scenario:** Uber driver tracks the lengths of his passengers' rides for one month and whether the passenger took the ride for business or personal reasons.

Group	Mean	Std. Dev.	Q1	Median	Q3
Business	7.59	3.29	5.08	7.30	9.52
Personal	3.33	1.36	2.33	3.35	4.28



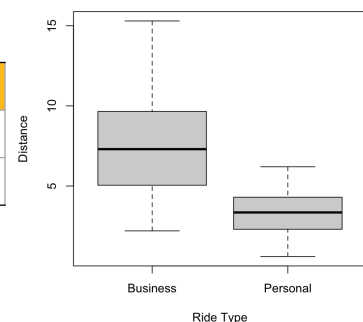
- **Question:** What do the graphs and statistics reveal about the centers of the groups?
- **Answer:** Typical Uber ride for _____ reasons is _____ than a typical Uber ride for _____ reasons
 - _____ are both larger
 - Longest personal rides are _____ ride

EXAMPLE: TWO-SAMPLE COMPARISON

18

- **Scenario:** Uber driver tracks the lengths of his passengers' rides for one month and whether the passenger took the ride for business or personal reasons.

Group	Mean	Std. Dev.	Q1	Median	Q3
Business	7.59	3.29	5.08	7.30	9.52
Personal	3.33	1.36	2.33	3.35	4.28



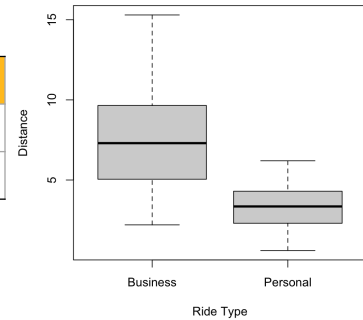
- **Question:** What do the graphs and statistics reveal about the spreads of the groups?
- **Answer:** Lengths of business rides are _____ than the lengths of personal rides
 - _____ and _____ (4.44 vs. 1.95) are _____
 - Business ranges from _____ compared to _____ for personal

EXAMPLE: TWO-SAMPLE COMPARISON

19

- **Scenario:** Uber driver tracks the lengths of his passengers' rides for one month and whether the passenger took the ride for business or personal reasons.

Group	Mean	Std. Dev.	Q1	Median	Q3
Business	7.59	3.29	5.08	7.30	9.52
Personal	3.33	1.36	2.33	3.35	4.28



- **Question:** What do the graphs and statistics reveal about the shapes of the distributions?
- **Answer:** Business rides are _____ while personal rides are _____
 - **Business:** _____ whisker _____ than _____ whisker
 - **Personal:** Both whiskers _____

EXAMPLE: SEVERAL SAMPLE COMPARISON

20

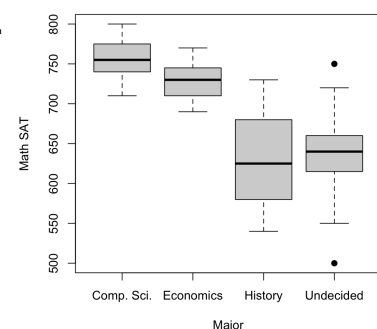
- **Scenario:** Compare math SAT scores for a random sample of Pitt students from three majors (computer science, economics, and history) as well as a group who are undecided.
- **Question:** Which variable is of each type?
- **Answer:**
 - Quantitative: _____
 - Categorical: _____
- **Question:** Do the observations in the categories overlap?
- **Answer:** _____
 - Each student is a _____ with _____ major (or _____ major)
 - _____ (unless a student is _____)
- **Takeaway:** All four samples are _____

EXAMPLE: SEVERAL SAMPLE COMPARISON

21

- **Scenario:** Compare math SAT scores for a random sample of Pitt students from three majors (computer science, economics, and history) as well as a group who are undecided.

Group	Mean	Std. Dev.	Q1	Median	Q3
Comp. Sci.	756	24.79	740	755	772.5
Economics	728.5	23.23	710	730	742.5
History	627	57.50	580	625	675
Undecided	639	54.09	617.5	640	660

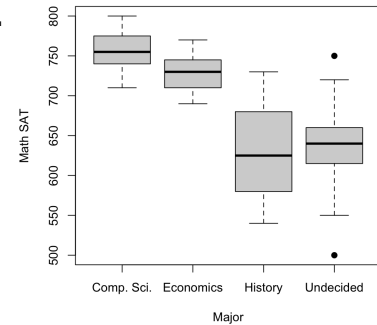


- **Question:** What do the graphs and statistics reveal about the centers of the groups?
- **Answer:** Computer science and economics are _____ than history and undecided
 - _____ slightly higher than _____; _____ slightly higher than _____

EXAMPLE: SEVERAL SAMPLE COMPARISON

- **Scenario:** Compare math SAT scores for a random sample of Pitt students from three majors (computer science, economics, and history) as well as a group who are undecided.

Group	Mean	Std. Dev.	Q1	Median	Q3
Comp. Sci.	756	24.79	740	755	772.5
Economics	728.5	23.23	710	730	742.5
History	627	57.50	580	625	675
Undecided	639	54.09	617.5	640	660



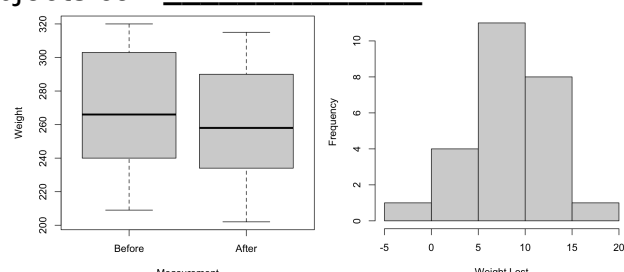
- **Question:** What do the graphs and statistics reveal about the spreads of the groups?
- **Answer:** Computer science and economics have _____ than history and undecided
 - CS and economics are _____; history and undecided are _____

EXAMPLE: PAIRED SAMPLE COMPARISON

- **Scenario:** Want to test effectiveness of a new weight loss program. Recruit 25 people who are weighed at the beginning of the study, follow the plan for 6 weeks, and get weighed again at the end.
- **Question:** Which variable is of each type?
- **Answer:**
 - Quantitative: _____
 - Categorical: _____ (_____)
- **Question:** Do the observations in the categories overlap?
- **Answer:** _____
 - Each subject was measured _____ and has _____ in the dataset
 - The “_____” value depends on the “_____” value
- **Takeaway:** The samples are _____.

EXAMPLE: PAIRED SAMPLE COMPARISON

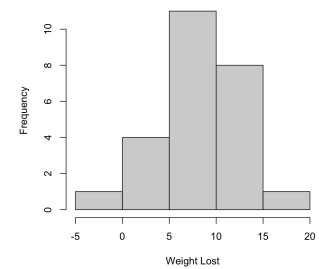
- **Scenario:** Want to test effectiveness of a new weight loss program. Recruit 25 people who are weighed at the beginning of the study, follow the plan for 6 weeks, and get weighed again at the end.
- **Question:** Which graph is most effective at displaying the effectiveness of the program?
- **Answer:** _____
 - Want to learn how much weight subjects can _____
 - Side-by-side boxplots only display _____ weights, but do not _____



EXAMPLE: PAIRED SAMPLE COMPARISON

- **Scenario:** Want to test effectiveness of a new weight loss program. Recruit 25 people who are weighed at the beginning of the study, follow the plan for 6 weeks, and get weighed again at the end.
- **Question:** What do the graph and statistics reveal about the center, spread, and shape?

Group	Mean	Std. Dev.	Q1	Median	Q3
Weight Lost	8.80	4.42	6	10	12



- **Answer:**

- **Center:** A typical subject lost _____ pounds
- **Spread:** Weight loss amounts tended to _____ by about _____
- **Shape:** Distribution is _____ and _____
 - _____