Summarizing Quantitative Data: Part I

- ➢ Histograms
- ➢ Mean
- ➢ Standard Deviation
- ➢ Z-Scores
- ➢ Empirical Rule

University of Pittsburgh

Motivation: Describing Quantitative Data

- **Scenario:** Take a random sample of 13 songs from Spotify and record the length (in seconds)

138 162 178 197 204 209 216 222 231 245 262 273 297

- **Question:** What type of data has been collected?
- **Answer:** ______________
  - Data is ________ and it makes sense to ______________________
- **Question:** What population does this data represent?
- **Answer:** ______________________

Motivation: Describing Quantitative Data

- **Scenario:** Take a random sample of 13 songs from Spotify and record the length (in seconds)

138 162 178 197 204 209 216 222 231 245 262 273 297

Average in the sample is 218 seconds.

- **Question:** Can we conclude that the mean of all observations in the population is greater than 210 seconds?
- **Answer:** _____
  - While the sample may be ______________, it is __________ to definitively conclude that the population mean is greater than 210
  - This is an ________ question that can only be answered by looking at not only the ________, but also the _______ of the observations

DESCRIBING A SINGLE QUANTITATIVE VARIABLE

- **Shape:** reports observations that tend to be more or less common
  - Symmetry/Skewness
  - Modality
- **Center**: measure of a “typical” observation
  - Mean
  - Median
- **Spread:** measure of how different the observations tend to be
  - Standard deviation
  - Range
  - Interquartile range

HISTOGRAM

- **Histogram:** graphical display that summarizes quantitative data using bars of different heights to show frequencies in predetermined ranges
- To construct a histogram:
  1. Divide the range of data into __________ (or ______)
  2. Count the ________________________ that fall into each bin
  3. Draw a _____ equal to the __________________ for each bin using vertical axis for ________ and horizontal axis for ____________

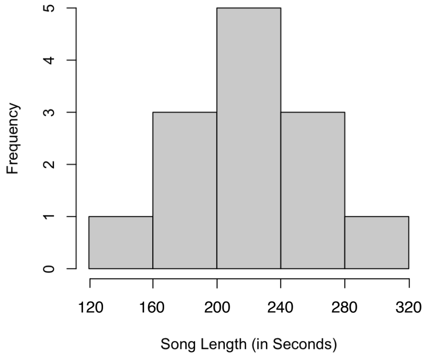
EXAMPLE: HISTOGRAM

- **Scenario:** Take a random sample of 13 songs from Spotify and record the length (in seconds)

  138 162 178 197 204 209 216 222 231 245 262 273 297

- **Question:** What does the histogram look like?
- **Answer:** ____ bins, each with a _____________ seconds

Bins	Count
120 to < 160	
160 to < 200	
200 to < 240	
240 to < 280	
280 to < 320	

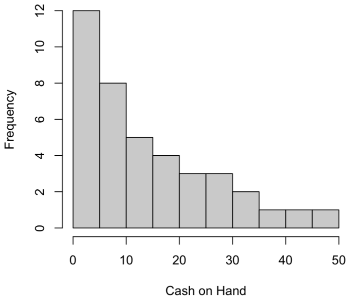


Note: Software typically does the best job of deciding how many bins to use and how wide they should be.

EXAMPLE: HISTOGRAM

- **Scenario:** Random sample of 40 people were asked how much cash they are currently carrying
- **Question:** What percentage of people were carrying between $10 and $25?
- **Answer:** ____ (________ = ____)

- **Question:** What percentage of people were carrying between $12 and $24?
- **Answer:** ___________________________
  - $12 and $24 are ___________ (i.e. are not values that _________ two bars)



DESCRIBING A HISTOGRAM

Symmetry

- **Symmetric:** similar number of observations on either side of the center
- **Left-skewed:** longer left tail
  - Indicates a few unusually small observations
- **Right-skewed:** longer right tail
  - Indicates a few unusually large observations

Modality

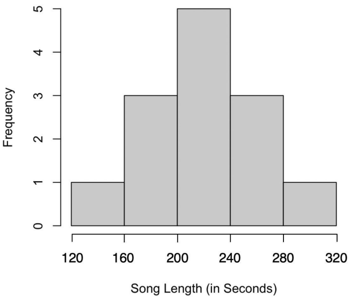
- **Unimodal:** one distinct peak
- **Bimodal:** two distinct peaks
- **Multimodel:** more than two distinct peaks
- **Uniform:** bars are all of similar height

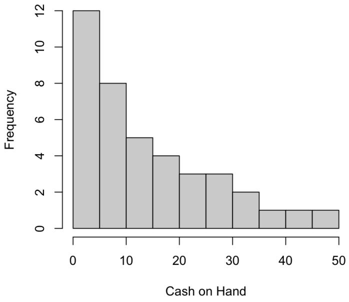
Important Special Case

Normal: symmetric and unimodal (bell-shaped)

EXAMPLE: HISTOGRAM

- **Scenario:** Histogram of song lengths on top; histogram of amount of cash on hand on bottom
- **Question:** What are the histogram shapes?
- **Answer:**
  - **Song Lengths:** _____________________
    - Typical song is about ____ minutes long, but ______ ______ to be _____________
    - Implies data is ________
  - **Cash on Hand:** _____________________
    - Most people carry _____________
    - As the amount of cash _________, the number of people carrying that amount _________





MEAN

- **Mean:** the average of all values in a set of quantitative data
  - Measure of the center of the data, or where a typical observation is likely to fall
  - **Population mean:** Denoted by μ (Greek letter "mu")
    - Parameter → Generally unknown
  - **Sample mean:** Denoted by $\bar{x}$ (Pronounced "x-bar")
    - Statistic → Can always be calculated for quantitative data
    - Good approximation of μ

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

EXAMPLE: MEAN

- **Scenario:** High temperature in Pittsburgh recorded each weekday during a week in June returned the following observations

75 76 78 80 81

- **Question:** What is the sample mean high temperature in June?
- **Answer:**

___ = ______________________ = _______ =

- **Question:** What is the population mean high temperature in June?
- **Answer:** _________
  - Only have a _______ of high temperatures→ ________________
  - Know probably ____________

STANDARD DEVIATION

- **Standard deviation:** measure of how "spread out" the data is
  - Can be used to measure how far away an observation is from the mean
  - **Population standard deviation:** Denoted by σ (Greek letter "sigma")
    - Parameter → Generally unknown
  - **Sample standard deviation:** Denoted by s
    - Statistic → Can always be calculated for quantitative data
    - Good approximation of σ

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

Example: Calculating Standard Deviation

- **Scenario:** High temperature in Pittsburgh recorded each weekday during a week in June returned the following observations

75 76 78 80 81

- **Question:** What is the sample standard deviation?
- **Answer:**

$$s = \sqrt{\frac{(\_\_\_\_\_\_)^2 + (\_\_\_\_\_\_)^2 + (\_\_\_\_\_\_)^2 + (\_\_\_\_\_\_)^2 + (\_\_\_\_\_\_)^2}{\_\_\_\_\_\_}}$$

$$= \sqrt{\frac{\_\_ + \_\_ + \_\_ + \_\_ + \_\_}{\_\_}}$$

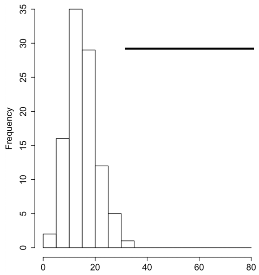
$$= \_\_\_\_\_\_ = \_\_\_\_\_\_$$

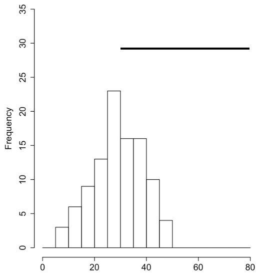
Example: Understanding Standard Deviation

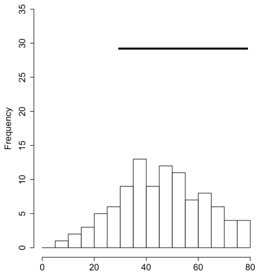
- **Scenario:** High temperature in Pittsburgh recorded each weekday during a week in June returned the following observations
- **Question:** What does the standard deviation tell us about the high temperatures in Pittsburgh in June?
- **Answer:** With a ____________ high temperature of ___, it would not be _________ to see the high temperature on a __________ as low as ___ or as high as ___.
  - Both are within _____________________ of the mean
- **Question:** What is the population standard deviation?
- **Answer:** ____________
  - Don't have high temperature for ________________
  - Probably ______________

Example: Understanding Standard Deviation

- **Scenario:** Commute time to work based on location of residence.
- **Task:** Classify each histogram as urban, suburban, or rural.





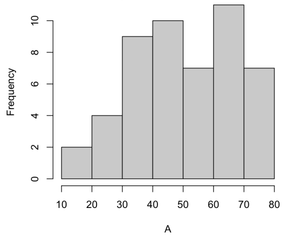


- **Task:** Approximate the mean and standard deviation of each group.
- **Answer:**

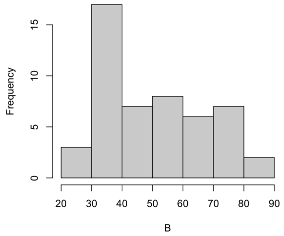
Variable	Urban	Suburban	Rural
Mean			
Standard Deviation			

Importance of Graphical and Numeric Displays

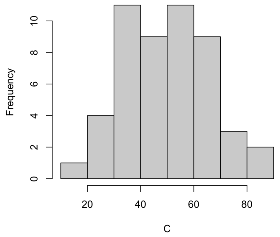
- **Question:** Why is it important to look at a graph of the data instead of only calculating the mean and standard deviation?



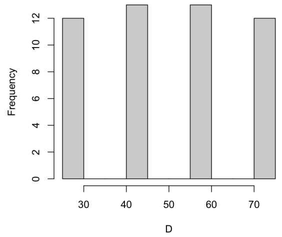
$\bar{x} = 50.34$
$s = 16.79$



$\bar{x} = 50.34$
$s = 16.79$



$\bar{x} = 50.34$
$s = 16.79$



$\bar{x} = 50.34$
$s = 16.79$

- **Answer:** Statistics do not ________________ the _______ of a dataset
  - All four graphs have the ________________________________, but drastically _________________
    - A: ____________
    - B: ____________
    - C: _______________
    - D: _______________

Standardizing and Z-Scores

- **Standardization:** the process of calculating how many standard deviations an observation is from its mean
  - Results in a Z-score

- **Z-score:** in a sample, a standardized value calculated as:

$$Z = \frac{x - \bar{x}}{s}$$

  - Positive Z-score → Observation _________ than mean
  - Negative Z-score → Observation ______ than mean
  - ___________ indicates how unusual the observation is relative to the rest of the data

Example: Using Z-Scores

- **Scenario:** SAT scores have a mean of 1000 and standard deviation 217 while ACT scores have a mean of 21 and standard deviation 5. Student A scored 1080 on the SAT, while Student B scored 28 on the ACT.
- **Question:** Which student performed better relative to the mean of the test they took?
- **Answer:** Student ___
  - **A:** $Z =$ ___________ = ____ standard deviations _______ the mean
  - **B:** $Z =$ ___________ = ____ standard deviations _______ the mean
- **Takeaway:** Cannot look at ________________ between an observation and the mean when comparing – must consider the _________.

Z-Score Interpretations and Outliers

- Observations that are:
  - Within 1 standard deviation of the mean
    - Quite common and not unusual in any way
  - Between 1 and 2 standard deviations from the mean
    - Relatively common, but slightly unusual
  - Between 2 and 3 standard deviations from the mean
    - Unusual, but not unheard of
  - More than 3 standard deviations from the mean
    - Exist, but are highly unusual and unlikely

- **Outlier:** an observation that is unusually far away from the rest of the data
  - ***Rule of Thumb #1:*** *More three standard deviations away from the mean*

Example: Z-Score Interpretations

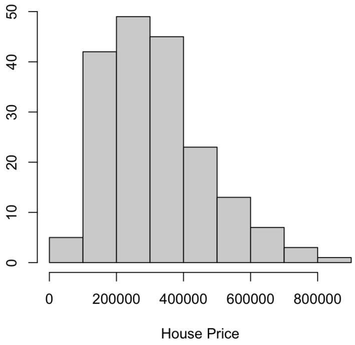
- **Scenario:** SAT scores have a mean of 1000 and a standard deviation of 217 while ACT scores have a mean of 21 and a standard deviation of 5. Student A scored 1080 on the SAT while Student B scored 28 on the ACT.
- **Question:** How should we describe each student's performance?
- **Answer:**
  - **A:** ____________________ but ______________
  - **B:** Slightly ________________ but still __________________

- **Question:** Would either score be considered an outlier?
- **Answer:** _____: Both are within __ standard deviations of their mean

Example: Impact of Unusual Observations

- **Scenario:** Histogram below displays assessed values of 188 houses in the same city.
- **Question:** What would happen to the mean and standard deviation if a house worth $2,000,000 were built?
- **Answer:** _____________________
  - **Mean:** ____ of house prices is now _____ ________ but sample size only increased by __
    - New value: ___________
  - **Standard Deviation:** Data is ________________
    - New value: ___________

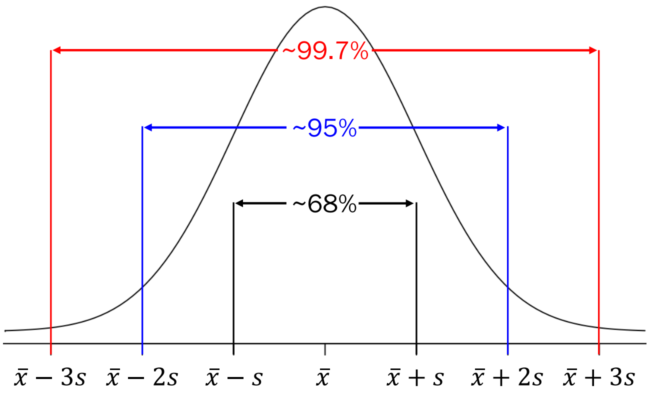
Variable	Mean	Std. Dev.
House Price	$317,277	$156,071



- **Takeaway:** Mean and standard deviation are ____________________ by outliers.

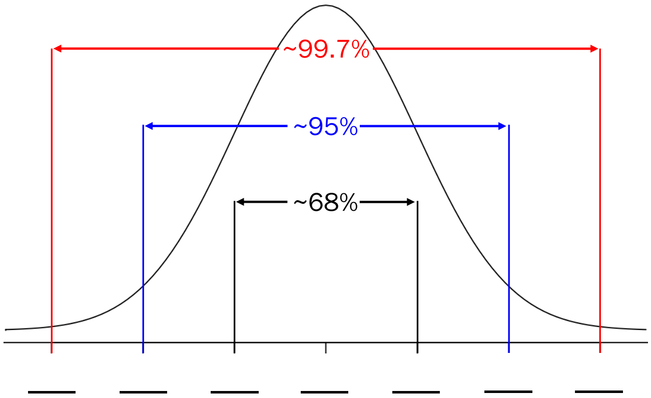
EMPIRICAL RULE

- **Empirical Rule:** In data that is **approximately normal**, about:
  - 68% of observations are within 1 standard deviation of the mean
  - 95% of observations are within 2 standard deviations of the mean
  - 99.7% of observations are within 3 standard deviations of the mean



EXAMPLE: EMPIRICAL RULE

- **Scenario:** IQ scores are normally distributed with a mean of 100 and a standard deviation of 15
- **Question:** What are the 7 values of the Empirical Rule?
- **Answer:**

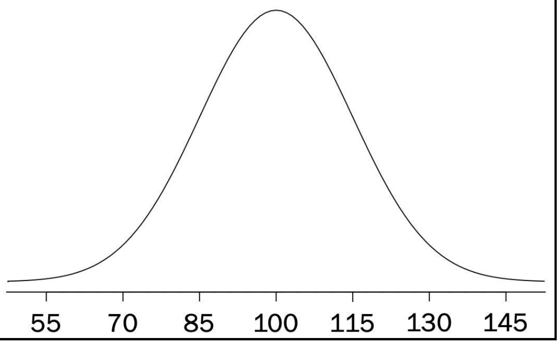


Interpretations:

- About ____% of IQ scores are between ____ and ____
- About ____% of IQ scores are between ____ and ____
- About ____% of IQ scores are between ____ and ____

EXAMPLE: EMPIRICAL RULE

- **Scenario:** IQ scores are normally distributed with a mean of 100 and a standard deviation of 15
- **Question:** What scores correspond to the highest 16% of IQs?
- **Answer:** Scores ____________
  - 68% of scores are ________________
  - ___% are in the tails (_____ or _____)
  - ______ (____) is in the upper tail
- **Question:** What percentage of IQ scores are between 82 and 124?
- **Answer:** __________________________
  - Empirical Rule is _________________
  - Can answer this using ___________


55
70
85
100
115
130
145