# STAT 103
# Statistical Thinking

Exam 2 Review

Professor Mengshi Zhou

# Chapter 10 Variable Type

- A **categorical variable** places individuals into categories

- A **quantitative variable** takes on numeric values for which arithmetic operations make sense

# Chapter 10 Distribution

- We can use frequency tables to summarize data

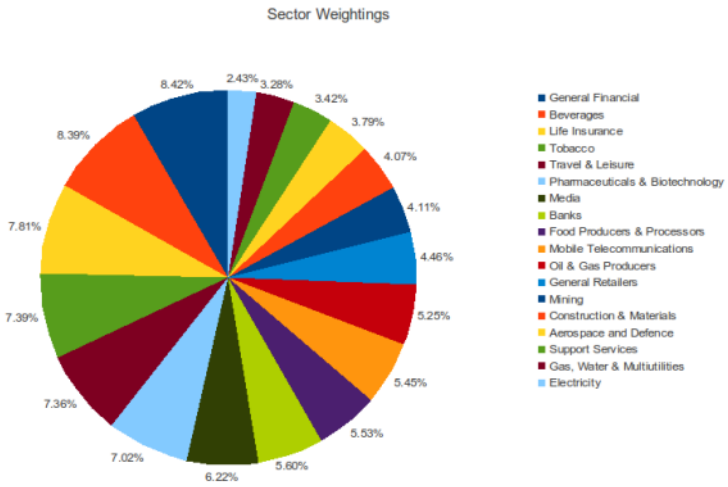
- The distribution of a variable:

Tells what values the variable can take and How often it takes those values.

| Position | Number of Players | Rates |
|---|---|---|
| Catcher | 34 | 7.7 |
| First Baseman | 19 | 4.3 |
| Outfielder | 83 | 18.9 |
| Pitcher | 215 | 49.0 |
| Second Baseman | 24 | 5.5 |
| Shortstop | 41 | 9.3 |
| Third Baseman | 23 | 5.2 |

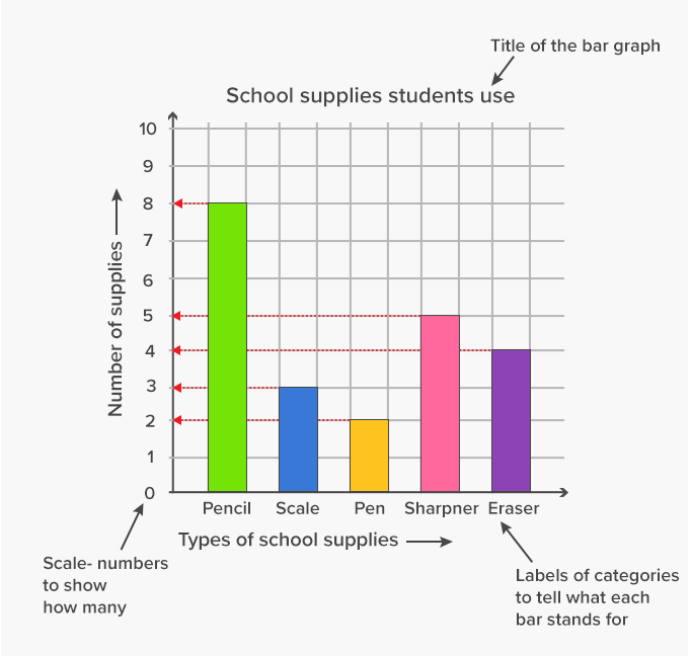
# Chapter 10 Bar graph and pie chart

- The wedges of the pie chart correspond to the rate of each value
- A pie chart can only be used when it represents all the parts of one whole
- We cannot use pie charts to describe the counts.



# Chapter 10 Bar graph and pie chart

- A bar graph is more versatile than a pie chart
- One axis represents all the possible value of a categorical variable
- The rates or counts will be displayed on the other axis and cover the whole range of possible values of the numbers counted for each category.



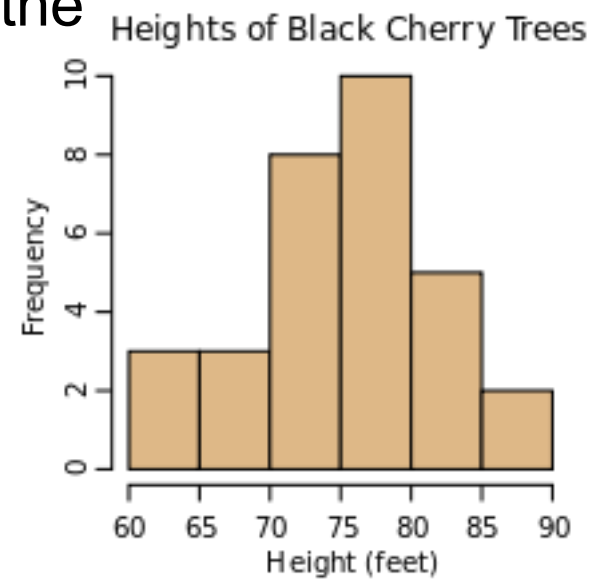
# Chapter 11 Histograms

- Quantitative variables can be illustrated with histograms

- A histogram divides the range of data into classes of equal widths and then counts the number of observations in each class

  – The classes of a histogram

  – Must be of equal width

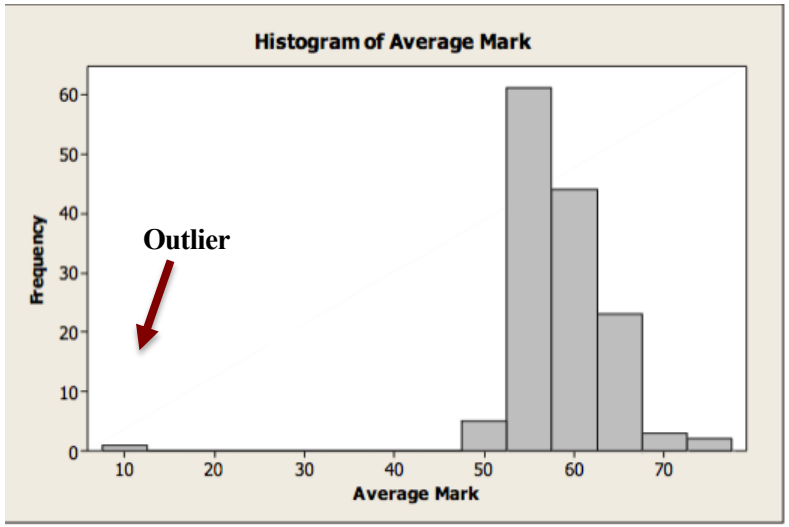


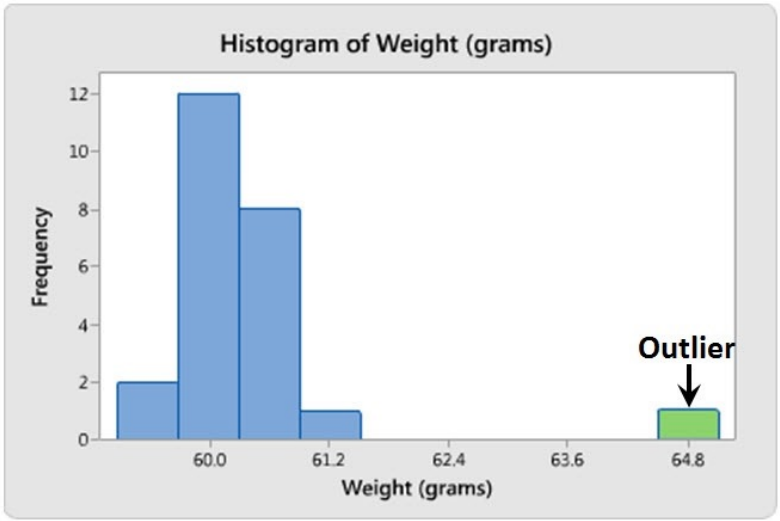
# Chapter 11 Histograms

- In a histogram, look for:
  - Skewness
  - The number of peaks
  - Any outliers that fall outside the pattern of the graph

# Outliers

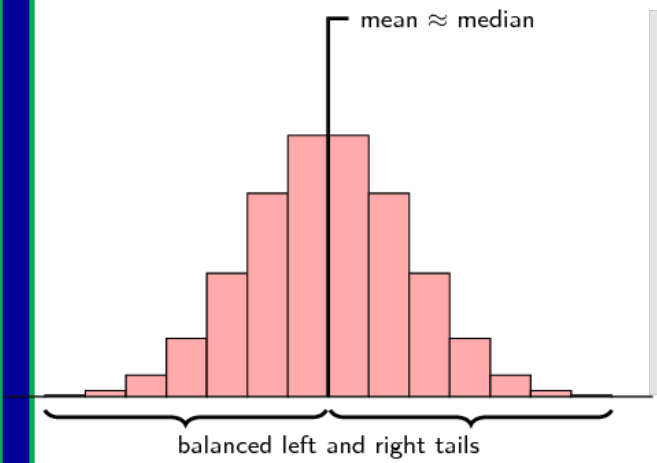
- Any data value away from the rest of the data a striking deviation

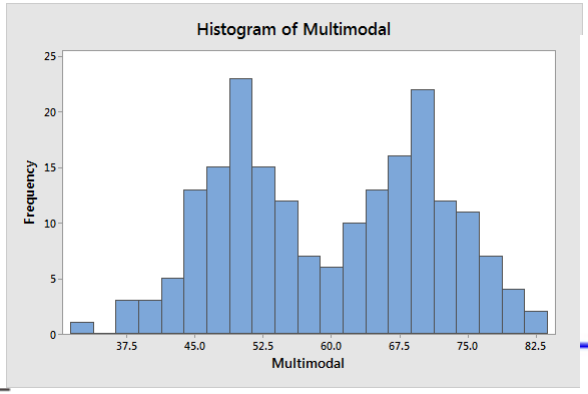


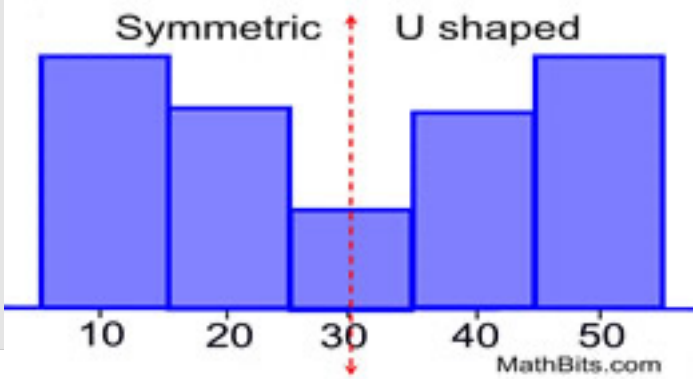



# Shape: Symmetric or Skewed?

- Symmetric: A distribution is set to be symmetrical if it can be divided into two equal sizes of the same shape

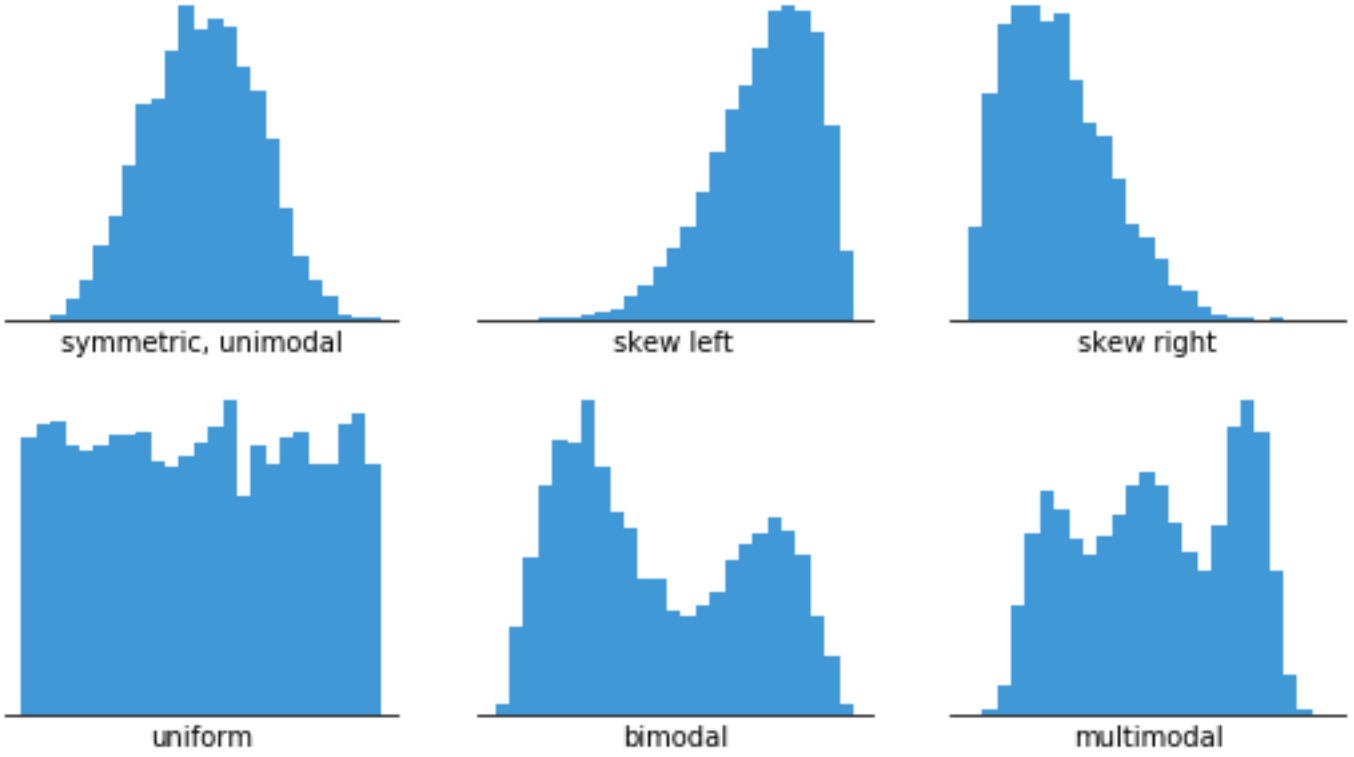


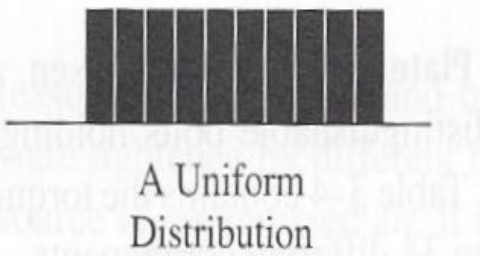
# Shape: Symmetric or Skewed?

- Skewed: a skewed distribution refers to asymmetry distribution

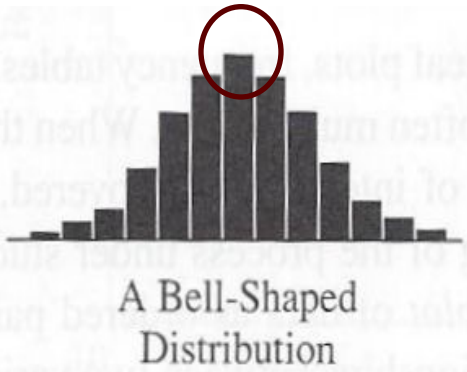


# Shape: Peaks

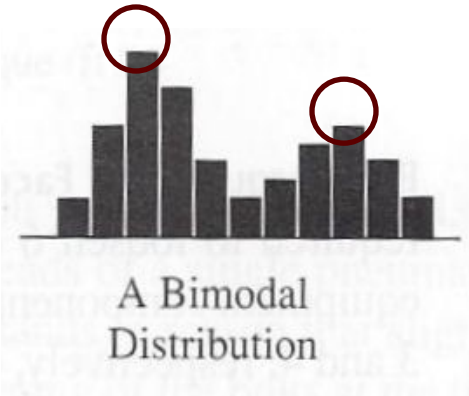
- Peaks occur when that data value is greater than its neighboring data points (on the left and right sides)




No peak-uniform




One peak-unimodal




Two peaks-bimodal

# Chapter 12 Measure the Center

- The **mean** (the numerical average): The mean is found by obtaining a sum of all the observations and dividing by the sample size (n):

$$\overline{X} = \frac{\text{sum of observations}}{n}$$

- The **median** (the 50th percentile): The median is the middle value of a sample when the observations are sorted from smallest to largest. 50% larger than the median, 50% lower than the median.

# median

- Half observations are less than median, the other half are larger than median
- The procedure of find **median (M)**
  - Order n observations, smallest to largest
  - Find the value that is in **location** $\frac{n+1}{2}$
    - If n is odd, median is the value in location $\frac{n+1}{2}$
    - If n is even, median is the average of the values in locations before and after $\frac{n+1}{2}$

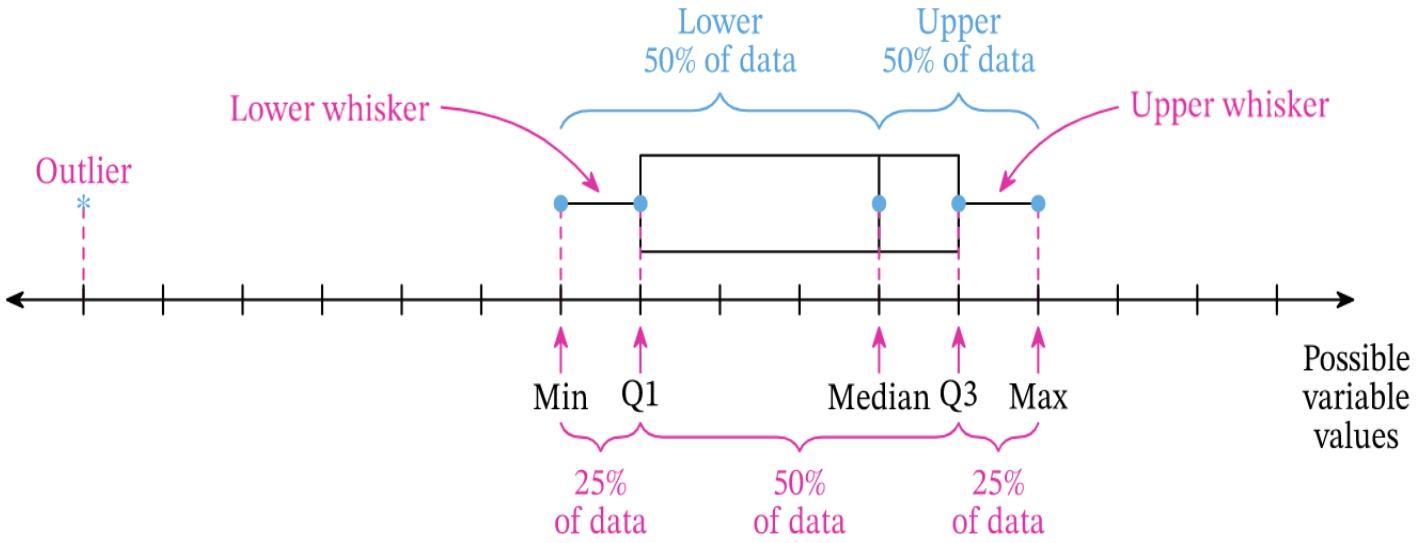
# Chapter 12 Measure the Spread

- The **standard deviation is the average distance from the mean**
  - As the observations become more spread from their mean, the standard deviation gets larger.
  - It is used to describe the spread only when the mean is used to describe the center
  - It is equal to zero only when there is no spread at all

# Chapter 12 Measure the Spread

- The first quartile (Q1) locates the middle of the lower half of the data. So, 25% of the data sit below Q1 and 75% of the data sit above Q1.
- The third quartile (Q3) locates the middle of the upper half of the data. So, 75% of the data sit below Q3 and 25% of the data sit above Q3.
- The interquartile range (IQR) = Q3 – Q1

# Chapter 12 Boxplot and five-number summary



# Chapter 12 Choose the right number of summarizing the distribution

- If a **sample has outliers and/or skewness**, resistant measures (median and IQR) are preferred over sensitive measures. This is because sensitive measures tend to overreact to the presence of outliers.

- If a sample is **reasonably symmetric with no outliers**, sensitive measures (mean and standard deviation) should be used. It is always better to use all of the observations in the sample when there are no problems with skewness and/or outliers.

# Chapter 12 Skewness

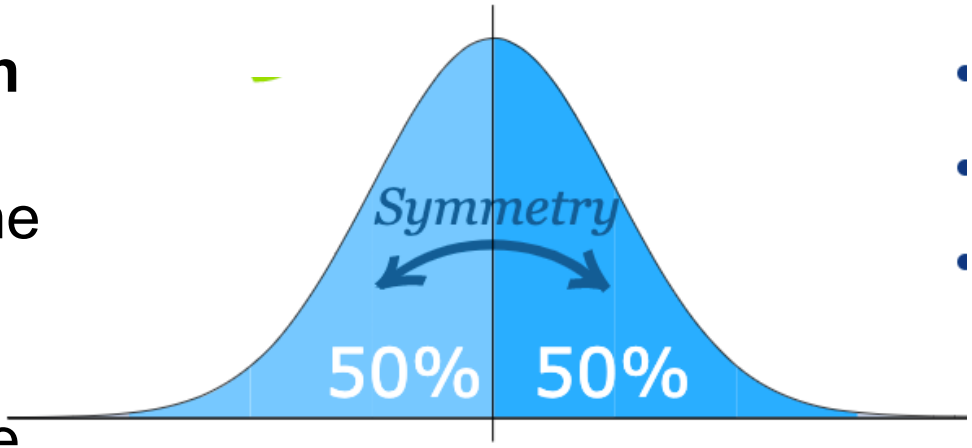
| Histogram Shape | Compare Two Measures of Centers |
|---|---|
| If symmetric | mean and median are approximately equal |
| If right skewed | mean is greater than the median |
| If left skewed | mean is less than median |

# Chapter 13

- Density Curve
- Normal Distribution
- 68-95-99.7 rule
- Standard score = $\frac{(observation - mean)}{(standard\ deviation)}$
- Percentiles

# Summary of Normal Distribution

- The normal distribution is **unimodal**
- The normal distribution is **symmetric about its mean**
- The curve is on or above the horizontal axis.
- The **mean** determine where the data tends to cluster
- the **standard deviation** determine how spread the distribution will be

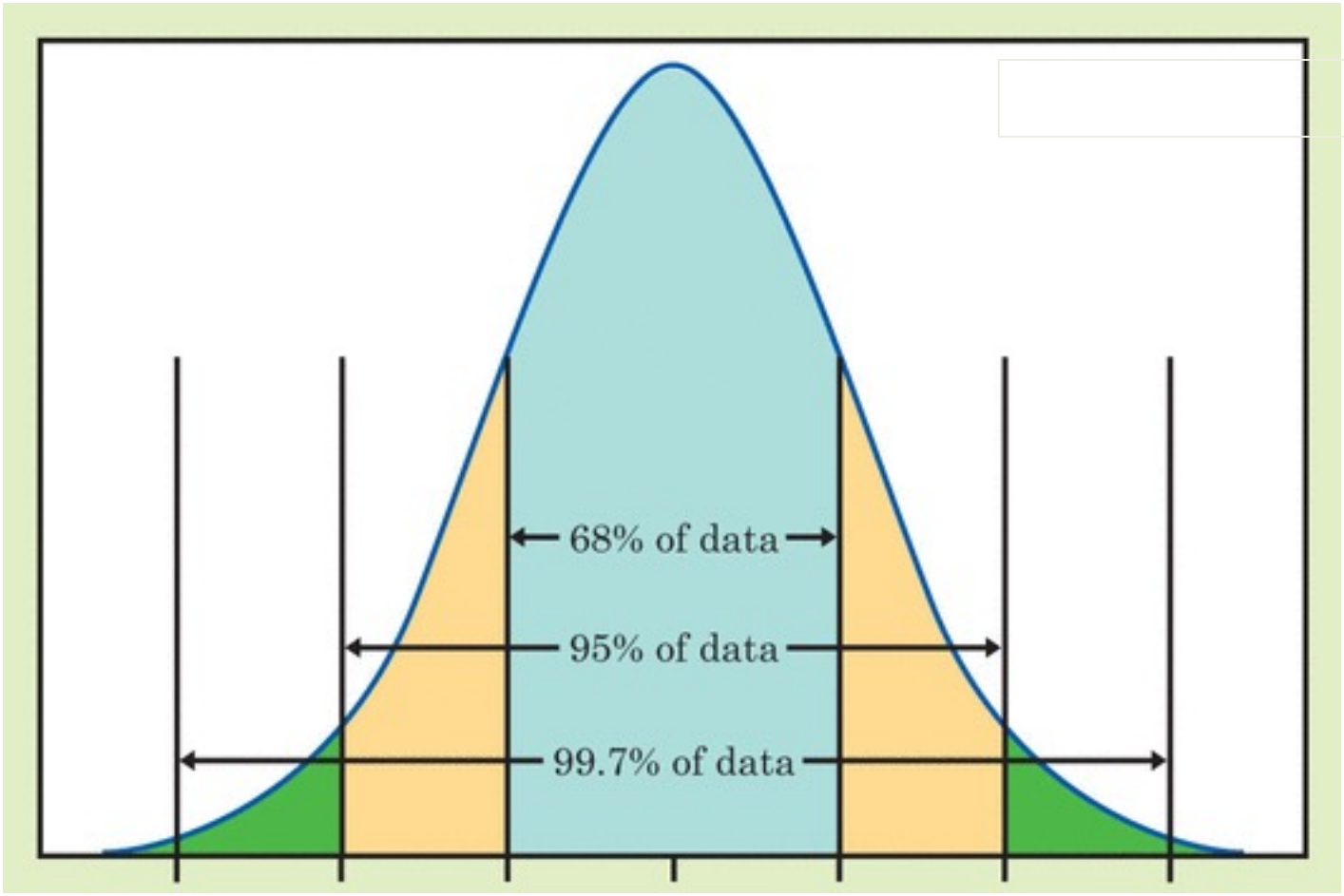


# What's so special about being "normal"?

- In any Normal distribution, approximately
  - 68% of all observations fall within 1 standard deviation from the mean *(mean – 1sd to mean + 1sd )*
  - 95% of all observations fall within 2 standard deviations from the mean *(mean – 2sd to mean + 2sd )*
  - 99.7% of all observations fall within 3 standard deviations from the mean *(mean – 3sd to mean + 3sd )*

*Only **approximately** true for any set of data*
*(No set of data is exactly described by a smooth curve)*

# 68-95-99.7 Rule



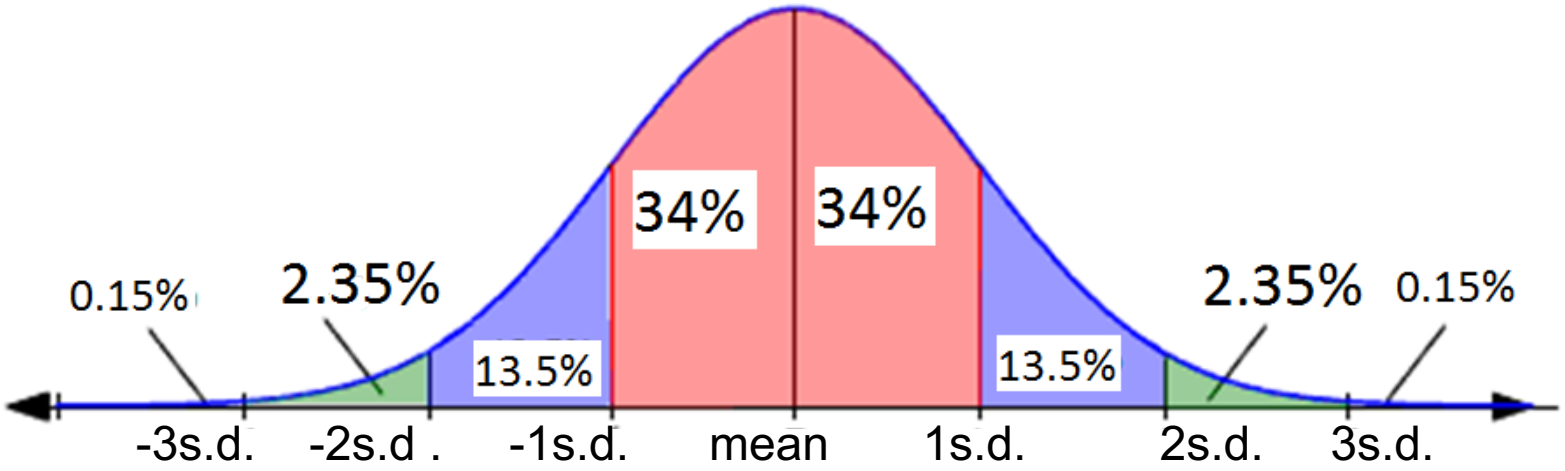
# Updated-68-95-99.7 rule

- Because the normal distribution is “symmetric”, we can find more information
- Examples:
  - 34% of all observations fall between mean to mean + 1 standard deviation
  - 2.35%+0.15% = 2.50% of all observations are larger than mean + 2 standard deviation
  - How many observations are smaller than mean - 2 standard deviation?

2.35%+0.15% = 2.50%

How many observations are larger than mean + 1 standard deviation?

13.5% + 2.25% + 0.15% = 16%



# Standard Scores

- We measure the distance an observation is away from the mean in standard deviation units.
  - *"this observation is 1 standard deviation above the mean"*
  - *"it falls 1.7 standard deviations below the mean"*

- The standardized score measure how many standard deviation a data point is above/below the mean
  - **Standard score** = $\frac{(observation - mean)}{(standard\ deviation)}$

- Most often called a **z-score**
- Use z-score to compare values, find percentages or percentiles
- The larger the z-score, the less likely for that value to occur

# Percentiles

- Percentile: the $n^{th}$ percentile is a value such that n percent of the observations lie below it (and the rest lie above it)
- If a person is in the $99^{th}$ percentile of height, this people is taller than 99% of the rest of the population
- If a number is in the $50^{th}$ percentile of height, this number is larger than 50 % of the observations, and smaller than 50 % of the rest of the observations
- If a number is in the $75^{th}$ percentile of height, this number is larger than 75 % of the observations, and smaller than 25 % of the rest of the observations