STA250 Probability and Statistics

Lecture 2: Measures of Tendency and Dispersion

Asst. Prof. Abdullah YALÇINKAYA

Ankara University, Faculty of Science, Department of Statistics

ayalcinkaya@ankara.edu.tr

2023

STA250 Probability and Statistics

Reference Books

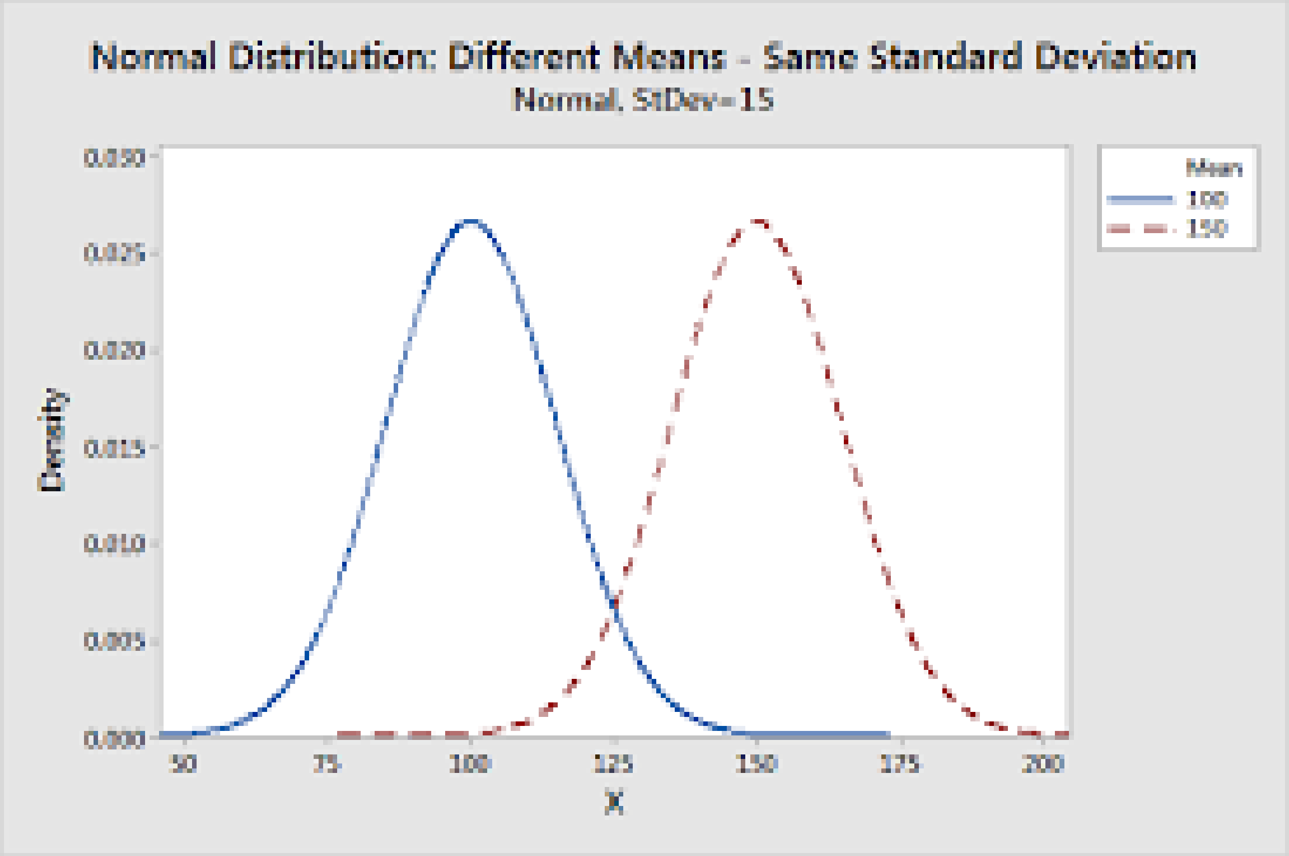
This lecture notes are prepared according to the contents of

«PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS»
by Walpole, Myers, Myers and Ye

«APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS»
by Montgomery and Runger

«Statistics for Biomedical Engineers and Scientists How to Visualize and Analyze Data» by Andrew P. King and Robert J. Eckersley

Central Tendency refers to the Middle of the Distribution



Central Tendency Measures

- **Mode**
- **Median**
- **Mean**

1.MODE

- When it's unique, the mode is the value that appears the most often in a data set and it can be used as a measure of central tendency
- But sometimes, there is no mode or there is more than one mode.
- There is no mode when all observed values appear the same number of times in a data set.
- There is more than one mode when the highest frequency was observed for more than one value in a data set.
- In both of these cases, the mode can’t be used to locate the centre of the distribution.

1.Mode (continues)

- The mode can be used to summarize categorical variables, while the mean and median can be calculated only for numeric variables.
- This is the main advantage of the mode as a measure of central tendency.
- It’s also useful for discrete variables and for continuous variables when they are expressed as intervals.

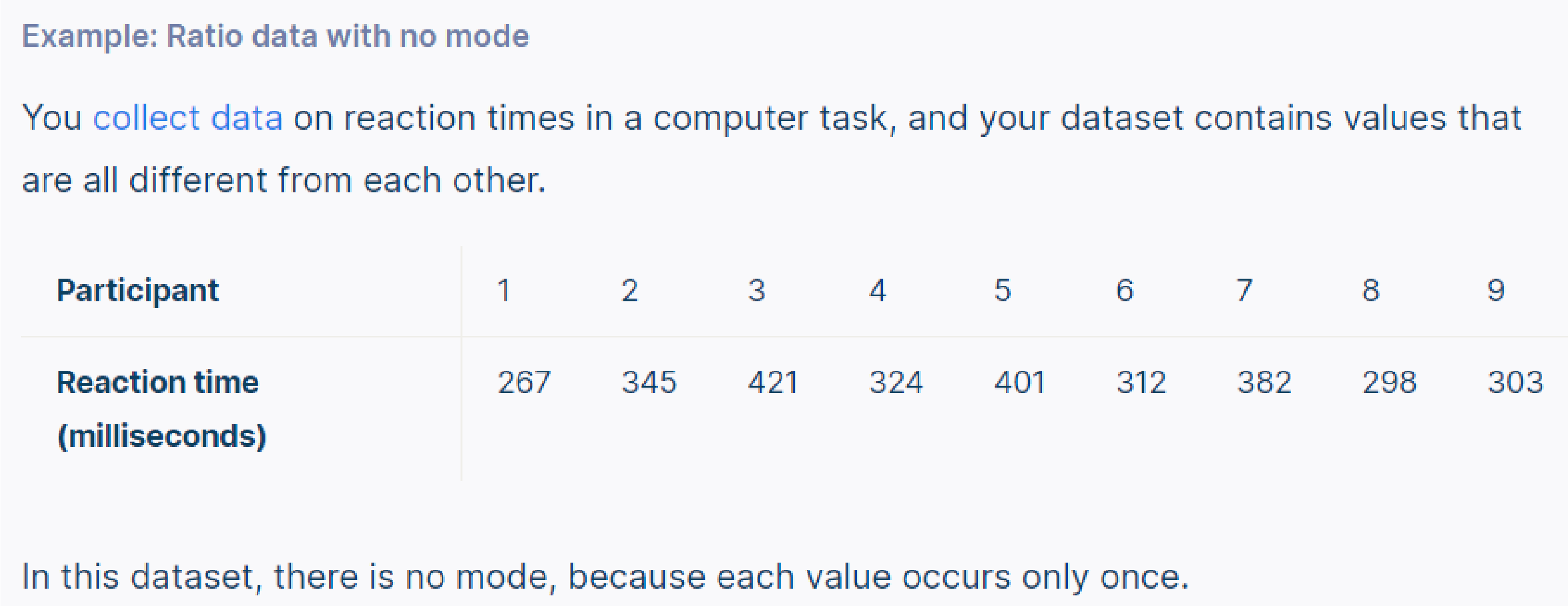
Example (mode)

- During a hockey tournament, a player scored 7, 5, 0, 7, 8, 5, 5, 4, 1 and 5 points in 10 games.

Number of points scored	Frequency
0	1
1	1
4	1
5	4
7	2
8	1

- After summarizing the data in a frequency table, you can easily see that the mode is 5 because this value appears 4 times so the value 5 appears the most often in the data set.
- The mode can be considered a measure of central tendency for this data set because it's unique.

Example (mode)



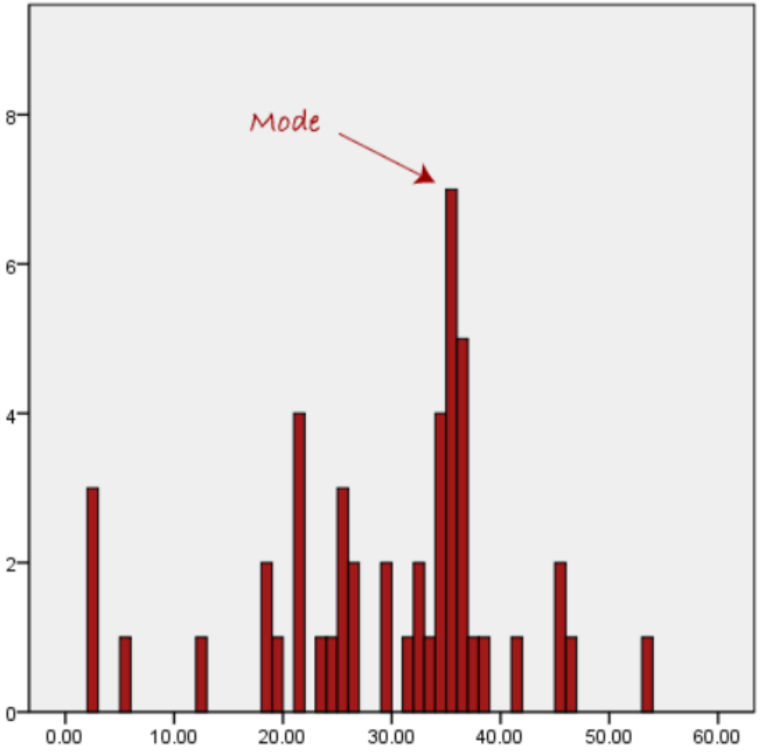
Example: Ratio data with no mode

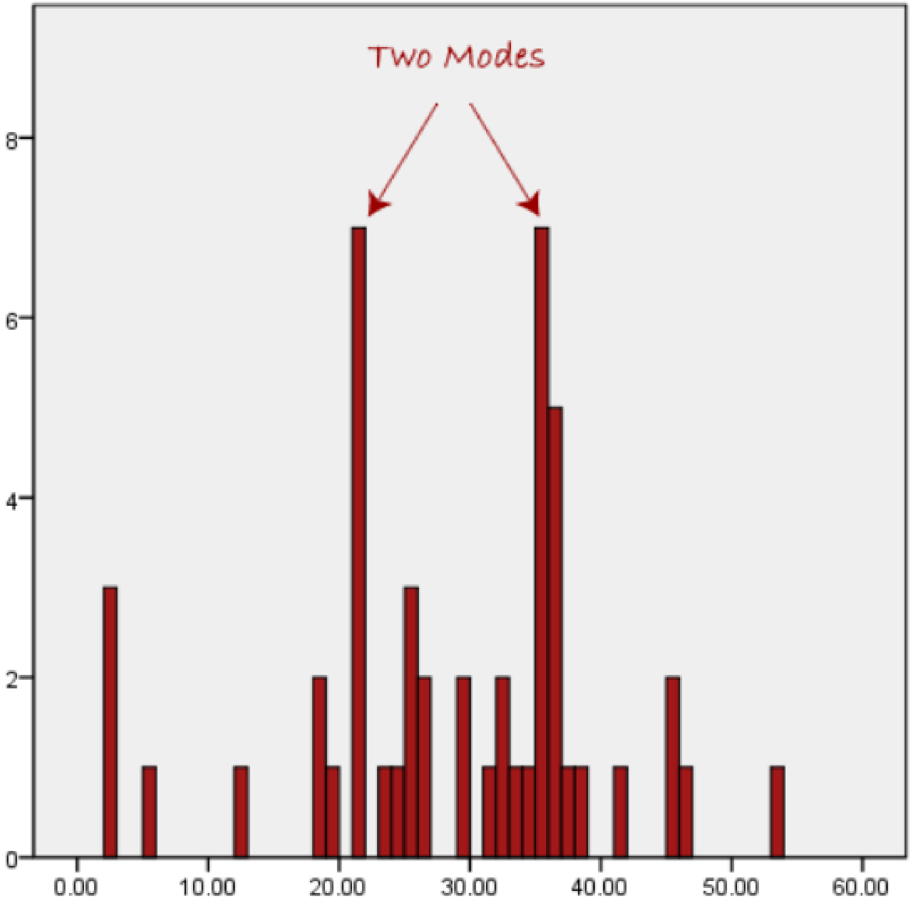
You collect data on reaction times in a computer task, and your dataset contains values that are all different from each other.

Participant	1	2	3	4	5	6	7	8	9
Reaction time (milliseconds)	267	345	421	324	401	312	382	298	303

In this dataset, there is no mode, because each value occurs only once.

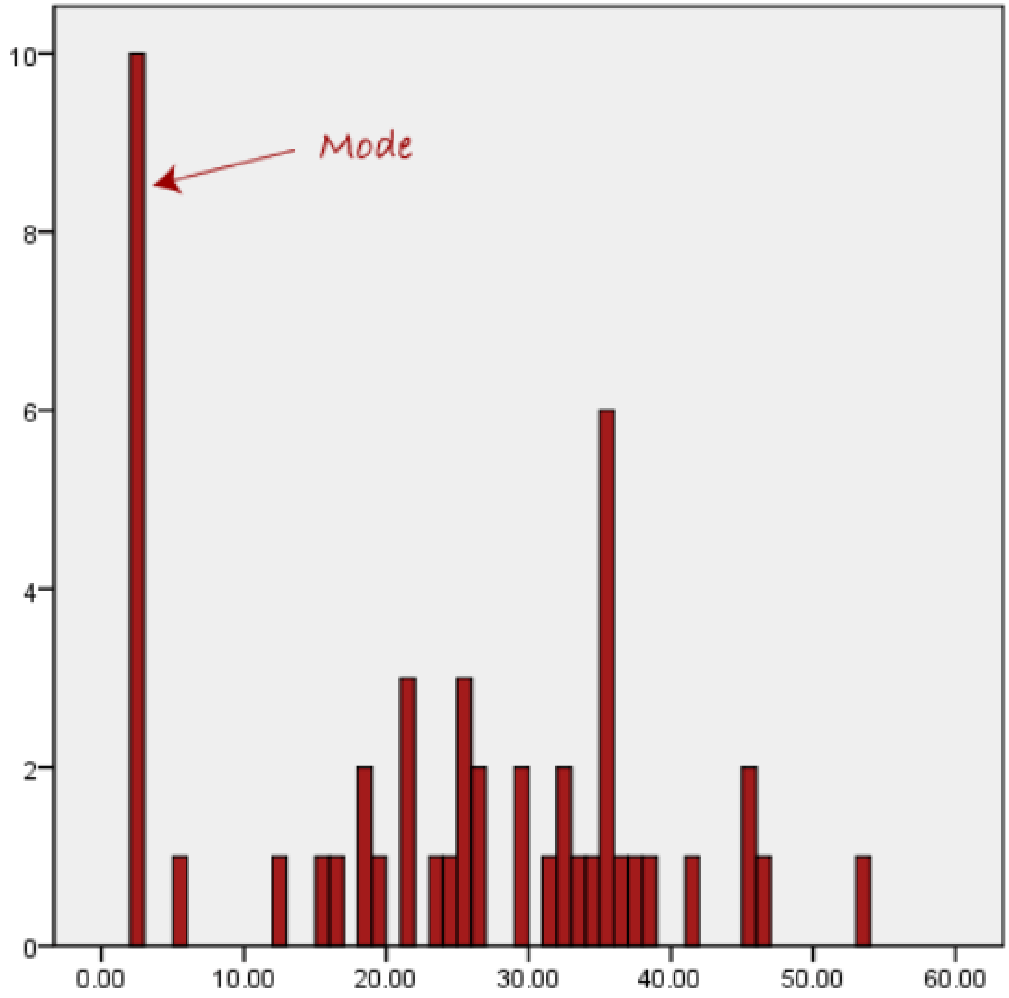
Examples (mod)





Source: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

Example (mod)



Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram.

Source: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

Measures of Location: The Sample Median

- **The <u>sample median</u> is another measure of location.**
  - The median is the midpoint of a observations.
  - Firstly, arrange all observation in order of size, from smallest to largest.
  - If the *n* observations are $x_1, x_2, \ldots, x_n$, the sample median is calculated as follows:

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & if n\ is\ odd, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2)+1}\right) & if\ n\ is\ even, \end{cases}$$

Measures of Location: The Sample Median

- ***Example 1***
  - *data: 2 4 6 Median (M) = 4*

- ***Example 2***
  - *data: 2 4 6 8 Median = 5 (ave. of 4 and 6)*

- ***Example 3***
  - *data: 6 2 4 Median* $\neq$ *2*
  - *(order the values: 2 4 6 , so Median = 4)*

Measures of Location: The Sample Mean

- **The sample mean is the most common measure of central tendency in statistics.**
  - The sample mean is an estimate of the population mean.
  - To find mean of set of *n* observations,
    - add their values and
    - divide by the number observations.
  - If the *n* observations are x_1, x_2, …, x_n , the sample mean is calculated as follows:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

  - Or, in more compact notation,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

Two samples of 10 northern red oak seedlings are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen. The stem weights in grams were recorded after the end of 140 days.

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

No Nitrogen	Nitrogen
0.28	0.26
0.32	0.43
0.36	0.46
0.37	0.47
0.38	0.49
0.42	0.52
0.43	0.62
0.43	0.75
0.47	0.79
0.53	0.86

Solution

- **Mean**

$$\bar{x}_{no\ nitrogen} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{0.32+0.53+\cdots+0.43}{10} = 0.399$$

$$\bar{x}_{nitrogen} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{0.26+0.43+\cdots+0.86}{10} = 0.565$$

- **Median**

$$X_{no\ nitrogen} = \frac{0.38+0.42}{2} = 0.400$$

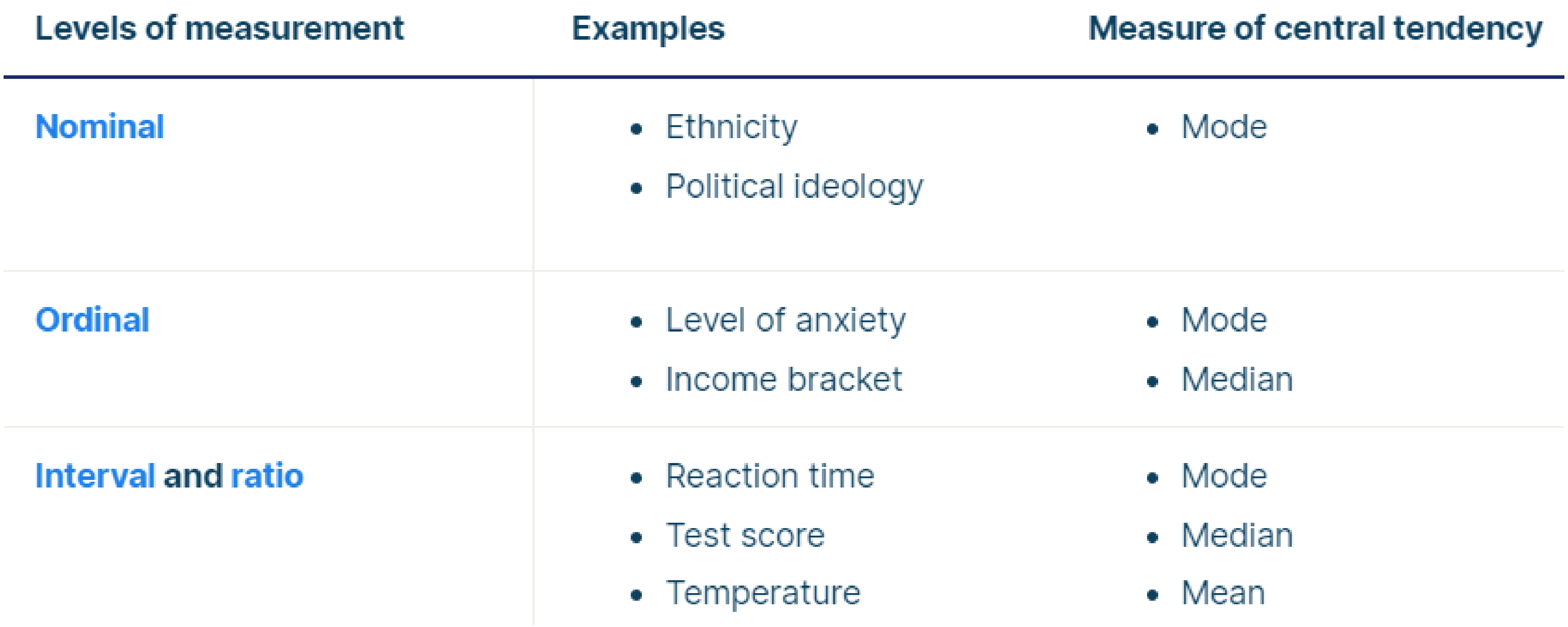
$$X_{nitrogen} = \frac{0.49+0.52}{2} = 0.505$$

When should you use the mean, median or mode?

- The 3 main measures of central tendency are best used in combination with each other because they have complementary strengths and limitations. But sometimes only 1 or 2 of them are applicable to your dataset, depending on the level of measurement of the variable.
- The mode can be used for any level of measurement, but it's most meaningful for nominal and ordinal levels.
- The median can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.
- The mean can only be used on interval and ratio levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

When should you use the mean, median or mode?



Levels of measurement	Examples	Measure of central tendency
Nominal	• Ethnicity • Political ideology	• Mode
Ordinal	• Level of anxiety • Income bracket	• Mode • Median
Interval and ratio	• Reaction time • Test score • Temperature	• Mode • Median • Mean

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

When should you use the mean, median or mode?

- To decide which measures of central tendency to use, you should also consider the distribution of your dataset.
- For normally distributed data, all three measures of central tendency will give you the same answer so they can all be used.
- In skewed distributions, the median is the best measure because it is unaffected by extreme outliers or non-symmetric distributions of scores. The mean and mode can vary in skewed distributions.

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

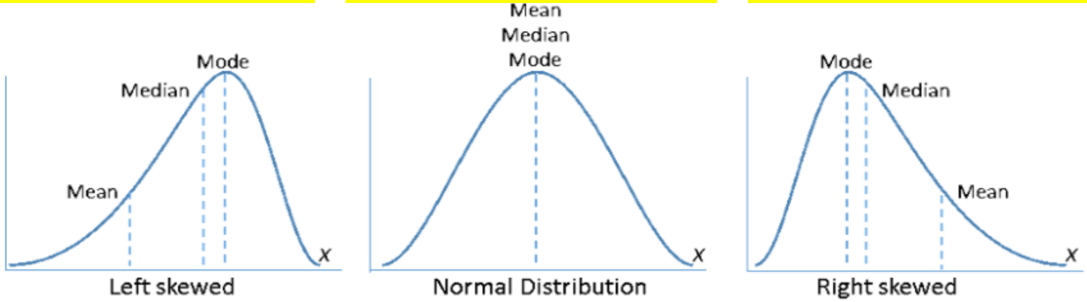
Comparing The Sample Mean and The Sample Median

- The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is usually farther out in the long tail than is the median.
- The mean is affected by outliers while sample median is not. For example, if one house is extremely expensive, then the mean will rise. The median would ignore that outlier.

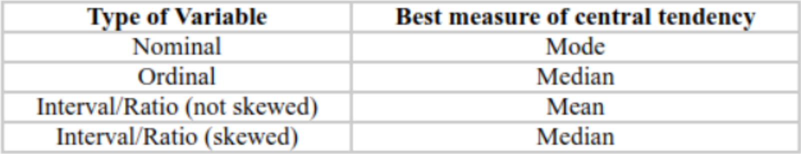
Mean<Median<Mode

Mean=Median=Mode

Mean>Median>Mode

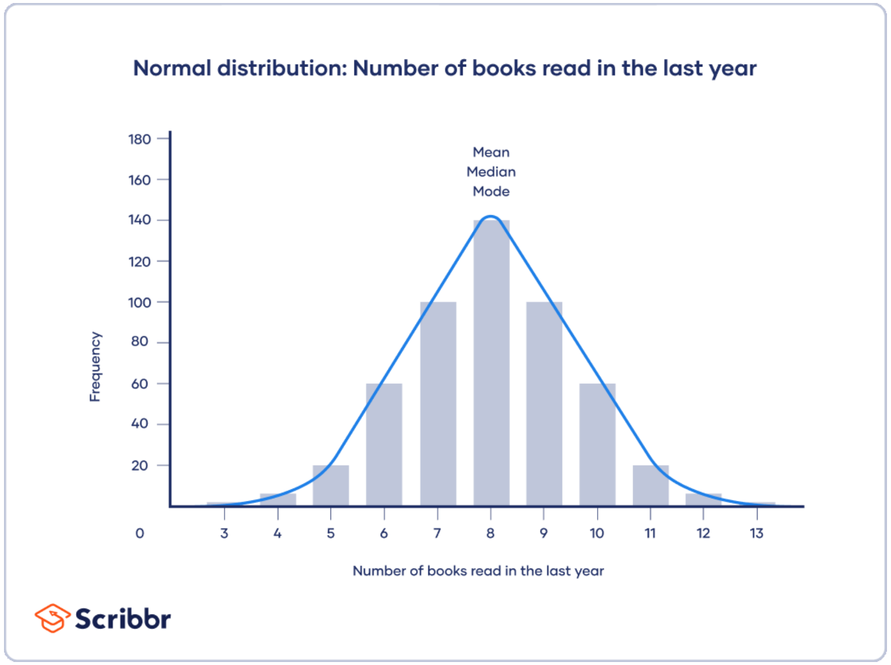


The table below shows the best measure of central tendency:



Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Comparing The Sample Mean and The Sample Median

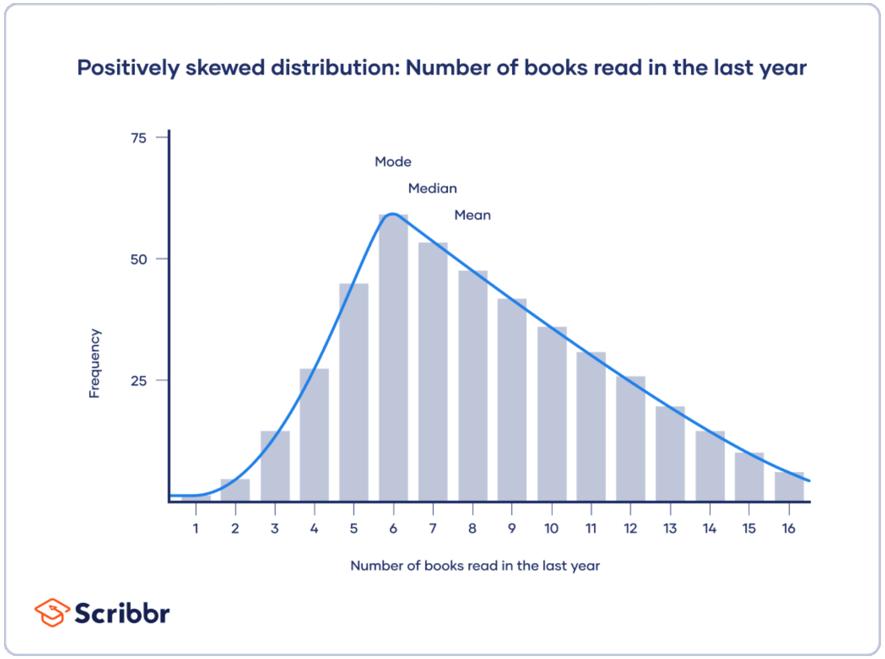


For symmetric distributions: Mean=Median=Mode

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

Comparing The Sample Mean and The Sample Median

- **Right skewed distributions: The direction of this tail tells you the side of the skew In a positively skewed distribution, there’s a cluster of lower scores and a spread out tail on the right.**

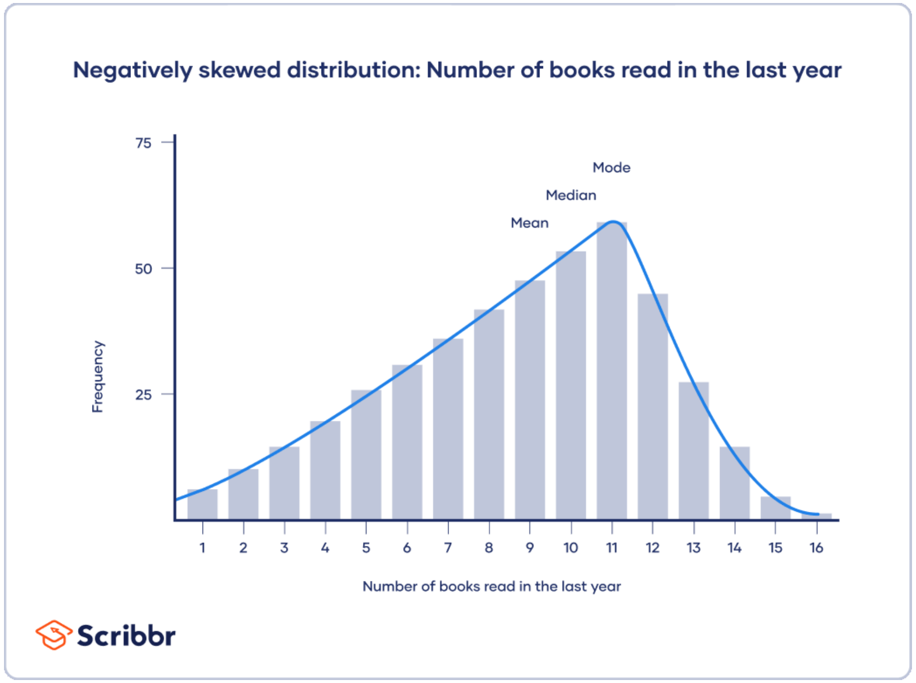


For right skewed distributions: Mode<Median<Mean

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

Comparing The Sample Mean and The Sample Median

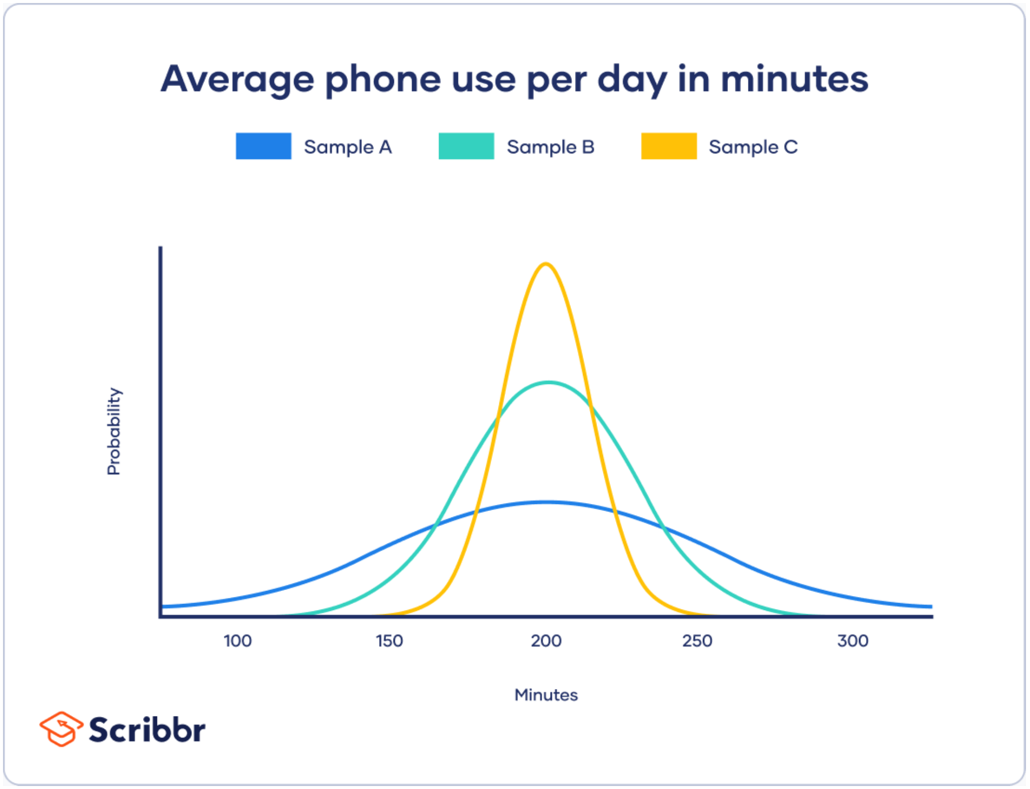
- **Left skewed distributions: In a negatively skewed distribution, there's a cluster of higher scores and a spread out tail on the left.**



For Left Skewed distributions: Mean<Median<Mode

Source: Bhandari, P. (2022, November 18). *Central Tendency | Understanding the Mean, Median & Mode.* Scribbr. Retrieved February 21, 2023, from https://www.scribbr.com/statistics/central-tendency/

Variability is about the spread.



Variability Measures

- **Sample variability is critical to statistical analysis.**
- **We will discuss the following measures of variability or spread.**
  - Range
  - Quartiles
  - Variance and Standard Deviation
  - Coefficient of Variation

Variability Measures: Range and Quartiles

- **The range is the simplest one. It considers only the largest and smallest value in the dataset.**
- **The range is strongly affected by outliers.**

$$\boldsymbol{Range = max - min}$$

- **Three numbers which divide the ordered data into four equal sized groups.**

  - Q_1 has 25% of the data below it.
  - Q_2 has 50% of the data below it. (MEDIAN)
  - Q_3 has 75% of the data below it.

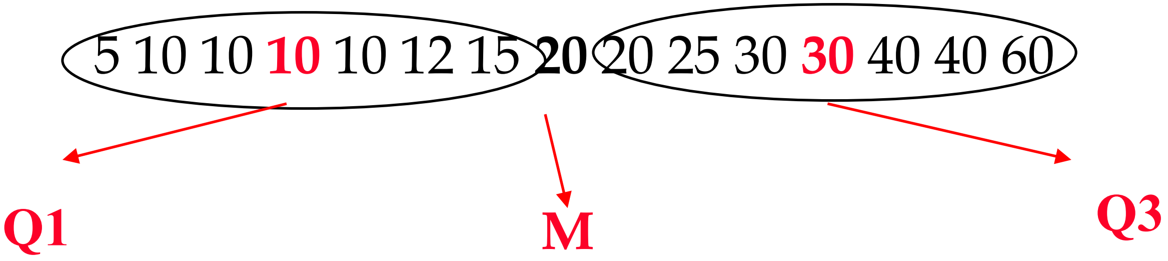
Variability Measures: Range and Quartiles

- **How do we obtain the quartiles?**

- Firstly, the data is sorted by ascending.
- The first **quartile Q1** is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
- The second **quartile Q2** is the median of the observations.
- The third **quartile Q3** is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Example

- Our North Carolina sample of 15 workers’ travel times, arranged in increasing order, is



5 10 10 **10** 10 12 15 **20** 20 25 30 **30** 40 40 60

Q1 **M** **Q3**

- There is an odd number observations, so the median is the middle one . Median= 20.
- Location of Q1: $\frac{n+1}{2} = \frac{7+1}{2} = 4$ this is the 4th of these 7 observations, so Q1=10 minutes.
- The third quartile is the median of the 7 observations to the right of the median, Q3=30 minutes.

The Five Number Summary and Box Plot

- **To get a quick summary of both center and spread , combine all five numbers. These are listed below.**

  1. Minumum
  2. Q1
  3. Median
  4. Q3
  5. Maximum

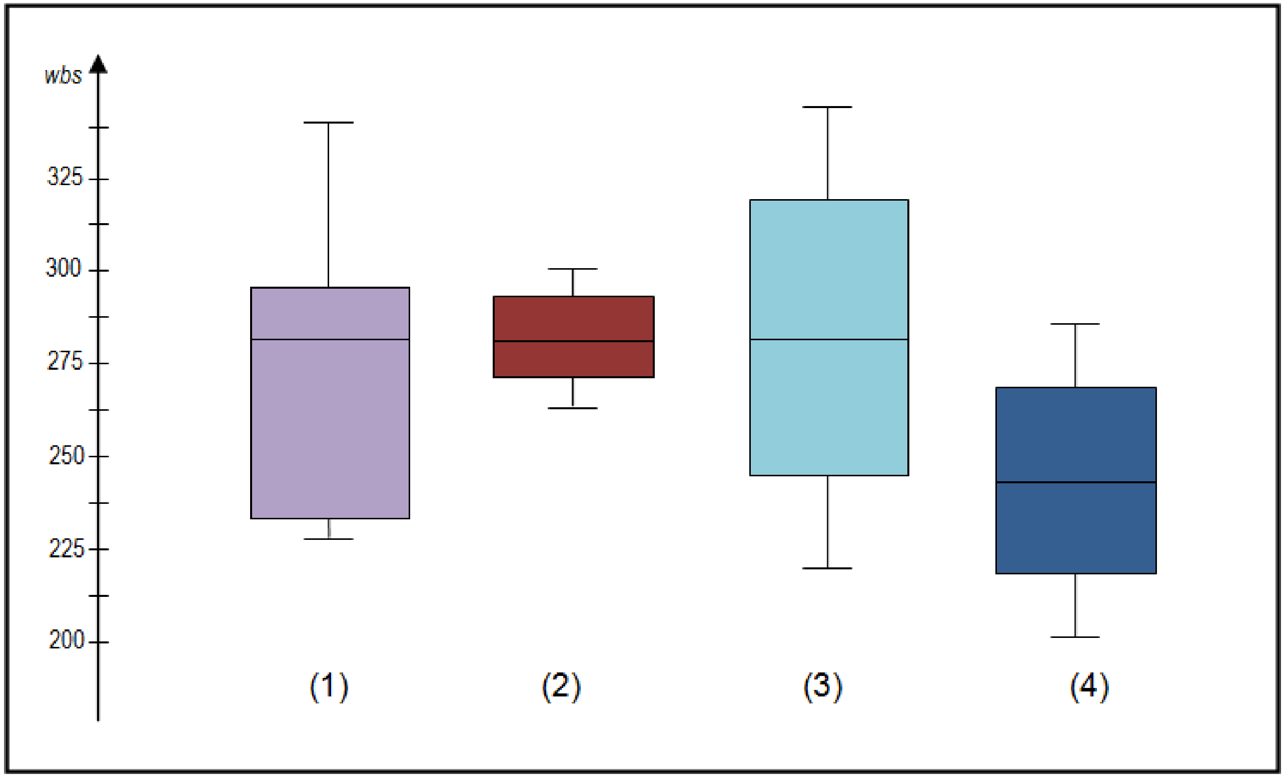
- **A boxplot is a graph of the five number summary.**
  - A center box spans the quartiles Q1 and Q3.
  - A line in the box marks the median M
  - Lines extend from the box out to the smallest and largest observations.

BOXPLOTS

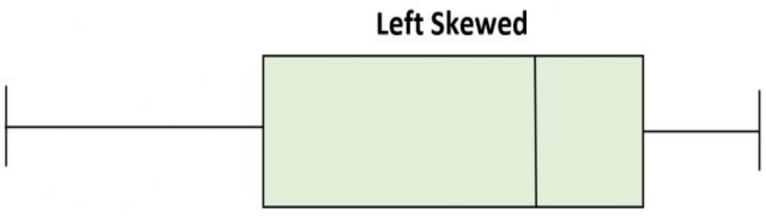
A boxplot is **a standardized way of displaying the distribution of data based on a five number summary** (“minimum”, first quartile [Q1], median, third quartile [Q3] and “maximum”).

It can tell you about your outliers and what their values are.

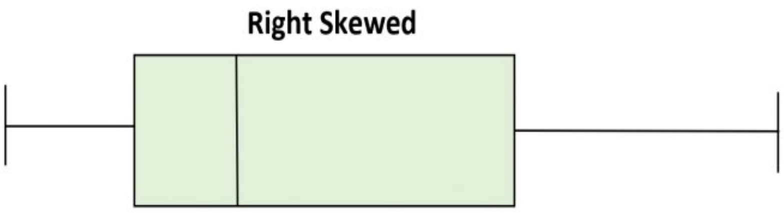
Interpreting box plots



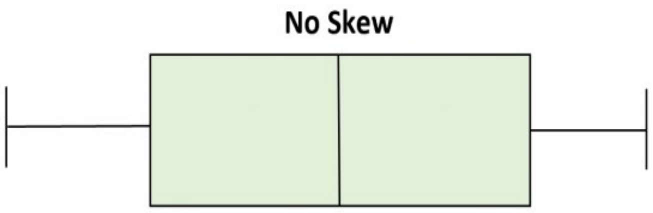
Interpreting box plots



When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).



When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).



When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.

Example

- Here are the travel times to work of the 20 New York workers and 15 North Carolina workers.

- North Caroline : 5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

- New York : 5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85

- The five number summary of travel times,

	Min	Q1	Median	Q3	Max
NC	5	10	20	30	60
NYC	5	15	22.5	42.5	85

Example

- Boxplots are best used for side-by-side comparison of more than one distribution, as in Figure 2.1.

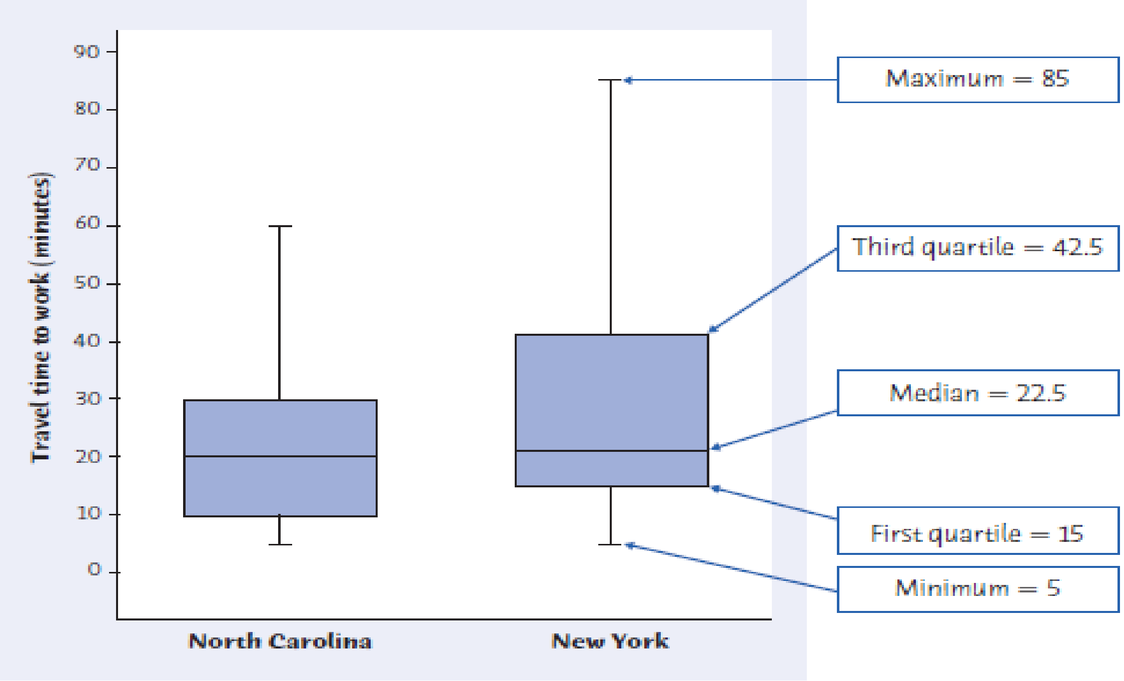


FIGURE 2.1
Boxplots comparing the travel times to work of samples of workers in North Carolina and New York.

The five-number summary of a distribution leads to a new graph, the *boxplot*. Figure 2.1 shows boxplots comparing travel times to work in North Carolina and New York.

Moore SD, Notz IW, Flinger MA. The Basic Practice of Statistics,New York, 6t.Edtion.

Example

- We see from Figure 2.1 that travel times to work are in general a bit longer in New York than in North Carolina.
- The median, both quartiles, and the maximum are all larger in New York.
- New York travel times are also more variable, as shown by the span of the box and the spread between the extremes.
- The New York data are more strongly right-skewed.

Outliers

- Look again at the boxplot of travel times to work in New York in Example.
- The five-number summary for this distribution is

15 22.5 42.5 85

How shall we describe the spread of this distribution?

- The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread. This distance is called **interquartile range**
- the **interquartile range (IQR)** is the distance between the first and third quartiles,

$$\boldsymbol{IQR = Q3 - Q1}$$

The *1.5 x IQR* Rule for Outliers

- The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.
- Call an observation a suspected outlier if it falls more than 1.5 *IQR* above the third quartile or below the first quartile.

Example:Using the *1.5 x IQR* rule

- For the New York travel time data, *IQR* =27.5 and

$$1.5\; x\; IQR = \; 1.5\; x\; 27.5 \; = \; 41.25$$

Any values not falling between

$$Q1 - (1.5x\; IQR) = \; 15.0 - 41.25 = -\,26.25$$
$$Q3 + (1.5x\; IQR) = \; 42.5 \; + 41.25 = \; 83.75$$

- Look again at the boxplot in Example: the only suspected outlier is the longest travel time, 85 minutes. *the 1.5 IQR rule* suggests that the three next-longest travel times (60 and 65 minutes) are just part of the long right tail of this skewed distribution.

Variability Measures: Variance and Standart Deviation.

- **Sample variance and standard deviation are the most important measures of variability.**
  - For a set of n observations, x_1, x_2, …, x_n , the sample variance, s^2, is calculated as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

  - n –1 is called the degrees of freedom associated with the variance. This is the number independent squared deviations, or pieces of information that make up s^2.

- **The standard deviation, s, is the square root of the variance.**

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

The usefulness of the standard deviation:

- s measures *spread about the mean* and should be used only when the mean is chosen as the measure of center.
- s is *always zero or greater than zero.* $s=0$ only when there is no spread. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out their mean, s gets larger.
- s has the *same units of measurement as the original observations.* For example, if you measure weight in kilograms, both the mean x and the standard deviation s are also in kilograms.
- Like the mean x, s is *not resistant.* A few outliers can make s very large.

For example; the standard deviation of the travel times for the 15 North Carolina workers in the Example is 15.23 minutes. If we omit the high outlier, the standard deviation drops to 11.56 minutes.

Example:

- Georgia Southern University had 2417 students with regular admission in their freshman class of 2010. For each student, data are available on their SAT and ACT scores (if taken), high school GPA, and the college within the university to which they were admitted.[7]In Exercise 3.49, the full data set for the SAT Critical Reading scores will be examined. Here are the first five observations from that data set

650 490 580 450 570

- We will compute *x* and *s* for these students. First, find the mean:

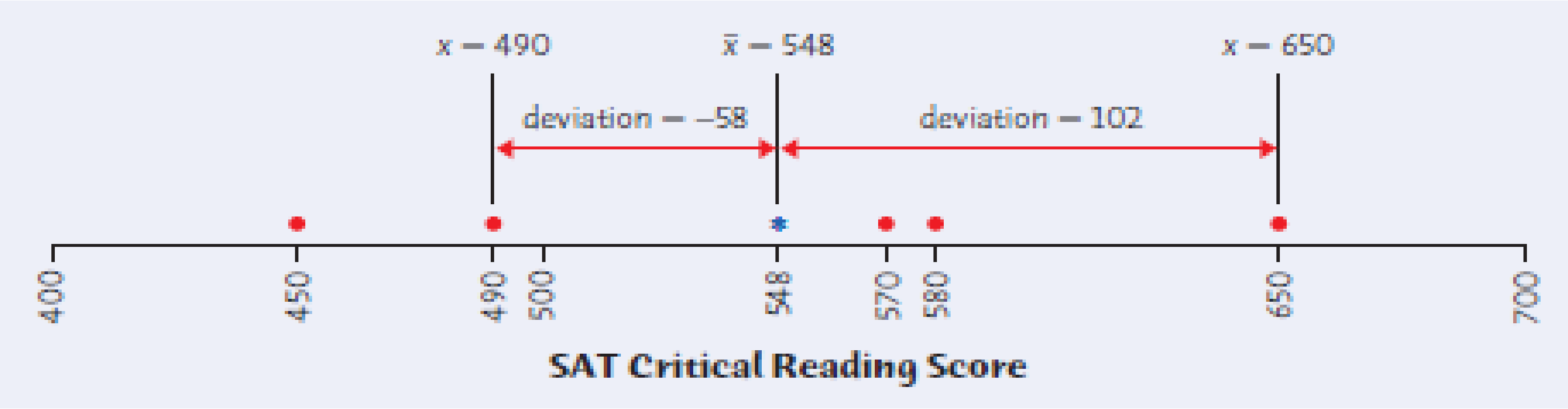
$$\bar{x} = \frac{650 + 490 + 580 + 450 + 570}{5}$$

$$= \frac{2740}{5} = 548$$

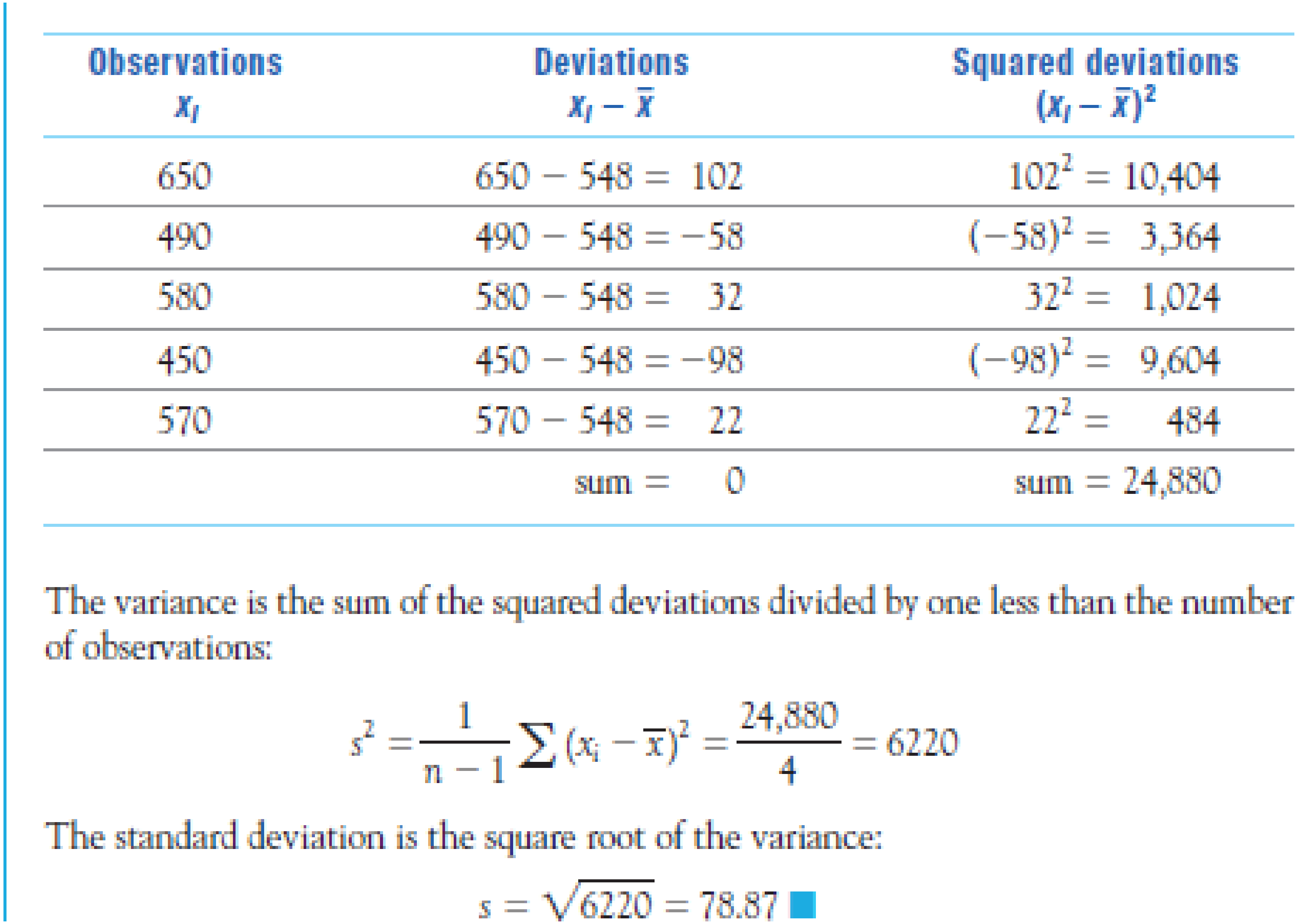
Example:

650 490 580 450 570

- This figure displays the data as points above the number line, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean. The deviations how how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.



Example:



Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
650	$650 - 548 = 102$	$102^2 = 10{,}404$
490	$490 - 548 = -58$	$(-58)^2 = 3{,}364$
580	$580 - 548 = 32$	$32^2 = 1{,}024$
450	$450 - 548 = -98$	$(-98)^2 = 9{,}604$
570	$570 - 548 = 22$	$22^2 = 484$
	sum $= 0$	sum $= 24{,}880$

The variance is the sum of the squared deviations divided by one less than the number of observations:

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{24{,}880}{4} = 6220$$

The standard deviation is the square root of the variance:

$$s = \sqrt{6220} = 78.87$$ ■

Example

- Let's calculate the variance of the follow data set: 2, 7, 3, 12, 9.
- The first step is to calculate the mean. The sum is 33 and there are 5 data points. Therefore, the mean is 33/5 = 6.6. Then you take each value in data set, subtract the mean and square the difference. For instance, for the first value:
- $(2 - 6.6)^2 = 21.16$
- The squared differences for all values are added:
- 21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.20
- The sum is then divided by (n-1) 69.20/4 = 17.3
- The variance is 17.3. To get the standard deviation, you calculate the square root of the variance, which is 4.16.

Variability Measures: Coefficient of Variation

- The coefficient of variation (CV) is defined as the ratio of the standard deviation. It shows the extent of variability in relation to the mean of the population.

$$CV = \frac{s}{\bar{x}}.100$$

- If the coefficient of variation approaches 0%, the variability is decreased, if it approaches above 25% indicates that the variability is quite increasing.
- Coefficient of variation is used to compare the variability of data sets with different means and data sets with difference in measurement units.

Coefficient of Variation: Example

- The following table gives the values of mean and variance of heights and weights of the 10 students of a school. Which is more varying than the other

	Height	Weight
Mean	155 cm	46.50 kg
Variance	72.25 cm	28.09 kg

- For comparing two data, first we have to find their coefficient of variations.

$$CV = \frac{s}{\bar{x}}.100$$

$$\text{CV1} = \frac{\sigma_1}{\bar{x}_1}.100 = \frac{8.5}{155}.100 = 5.48\%$$

$$\text{CV2} = \frac{\sigma_2}{\bar{x}_2}.100 = \frac{5.3}{46.50}.100 = 11.40\%$$

Since CV2>CV1, the weight of the students is more varying than the height.

Homework

- Amount spent (euros) by customers in a restaurant when exposed to odors
- **NO ODOR**

15.9 18.5 15.9 18.5 18.5 21.9 15.9 15.9 15.9 15.9 15.9 18.5 18.5 18.5 20.5 18.5 18.5 15.9 15.9 15.9 18.5 18.5 15.9 18.5 15.9 18.5 15.9 25.5 12.9 15.9

- **LEMON ODOR**

18.5 15.9 18.5 18.5 18.5 15.9 18.5 15.9 18.5 18.5 15.9 18.5 21.5 15.9 21.9 15.9 18.5 18.5 18.5 18.5 25.9 15.9 15.9 15.9 18.5 18.5 18.5 18.5

- **LAVENDER ODOR**

21.9 18.5 22.3 21.9 18.5 24.9 18.5 22.5 21.5 21.9 21.5 18.5 25.5 18.5 18.5 21.9 18.5 18.5 24.9 21.9 25.9 21.9 18.5 18.5 22.8 18.5 21.9 20.7 21.9 22.5

- (a) Give the five-number summary of this distribution.
- (b) Use the five-number summary to draw a boxplot of the data. What is the shape of the distribution?
- (c) Which observations does the 1.5 *IQR* rule flag as suspected outliers?

See You Next Lesson

Have a good week, best wishes…

See you ☺