DESCRIPTIVE STATISTICS

Descriptive Statistics: Understanding Data

- Descriptive statistics are summary statistics that quantitatively describe or summarize features of a dataset.

- They provide simple summaries about the sample and the measures. These summaries can be either quantitative (numerical) or visual (charts, graphs).

- Helps in simplifying large amounts of data in an easily understandable way.

Key Types of Descriptive Statistics

Measures of Central Tendency

- Mean: The average of all data points.
- Median: The middle value in a data set.
- Mode: The most frequently occurring value(s).

Measures of Variability

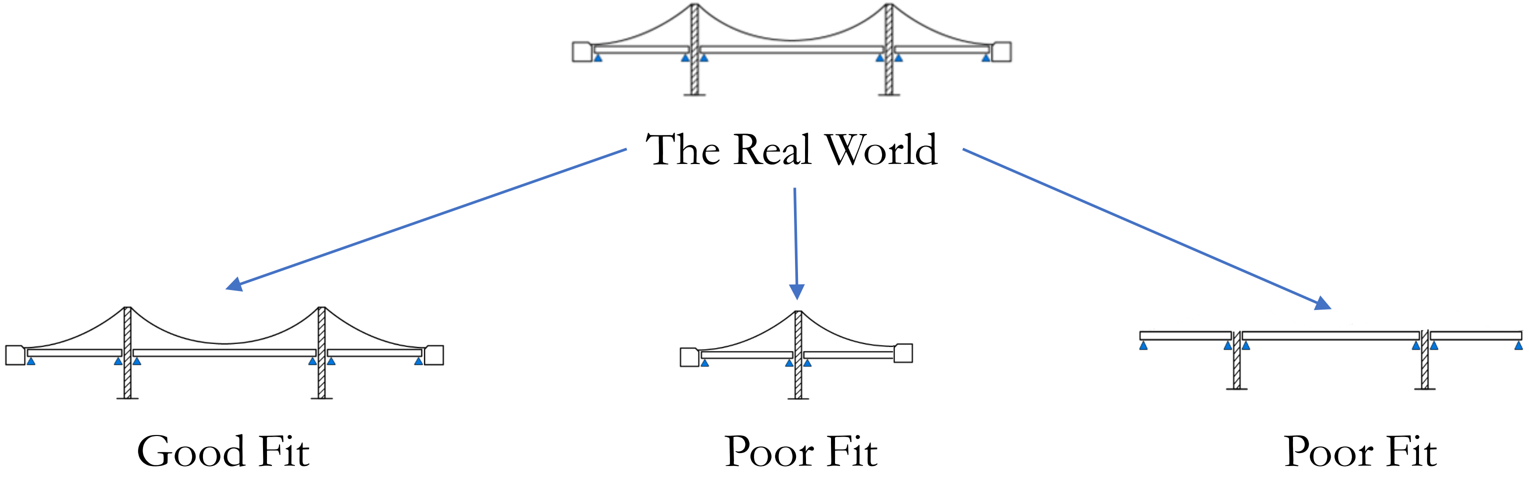
- Range: The difference between the highest and lowest values.
- Variance: Measures how spread out the numbers are around the mean.
- Standard Deviation: The square root of the variance.

Building Statistical Models

- Scientists are interested in discovering something about a phenomenon that we assume exists.

- We don’t have access to the real-world situation, so we can only “infer” based upon the models we build.

Building Statistical Models



The Real World

Good Fit

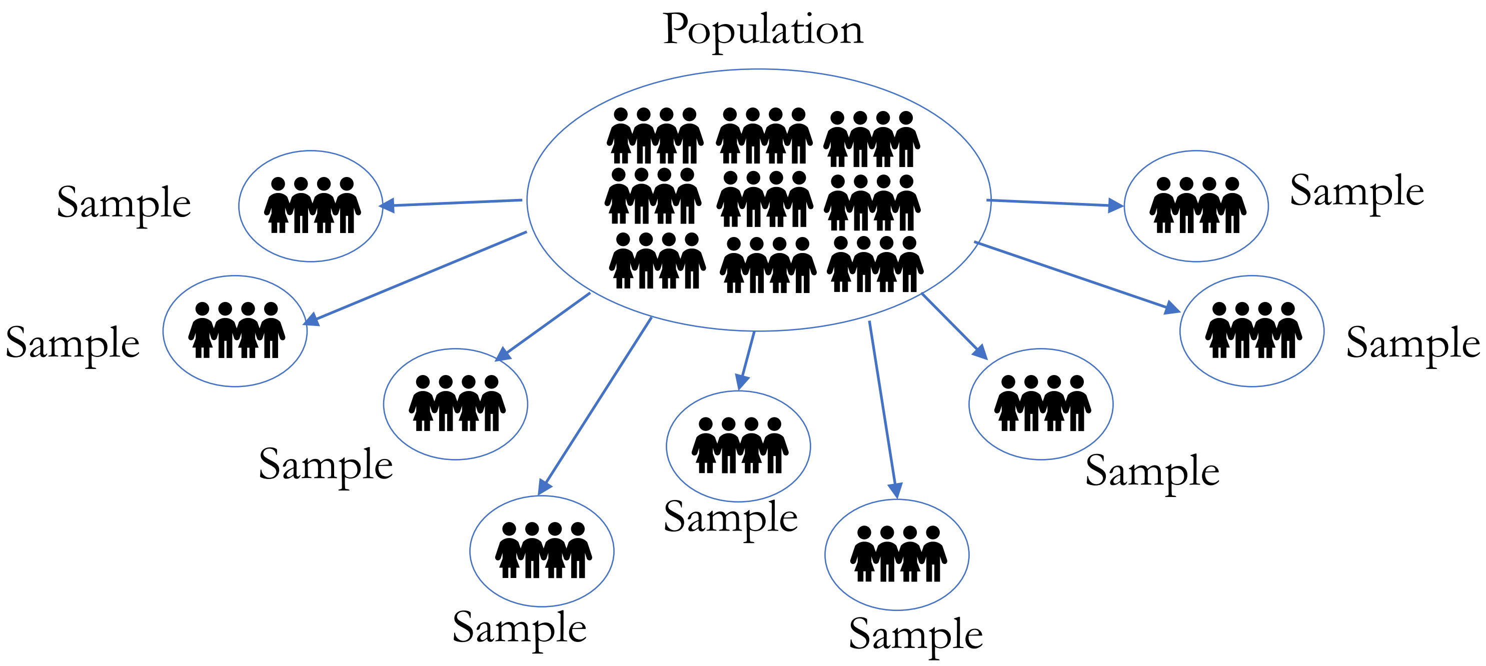
Poor Fit

Poor Fit

- If we want our inferences to be accurate, the statistical models we build must represent the data collected as closely as possible, known as *fit* of the model.

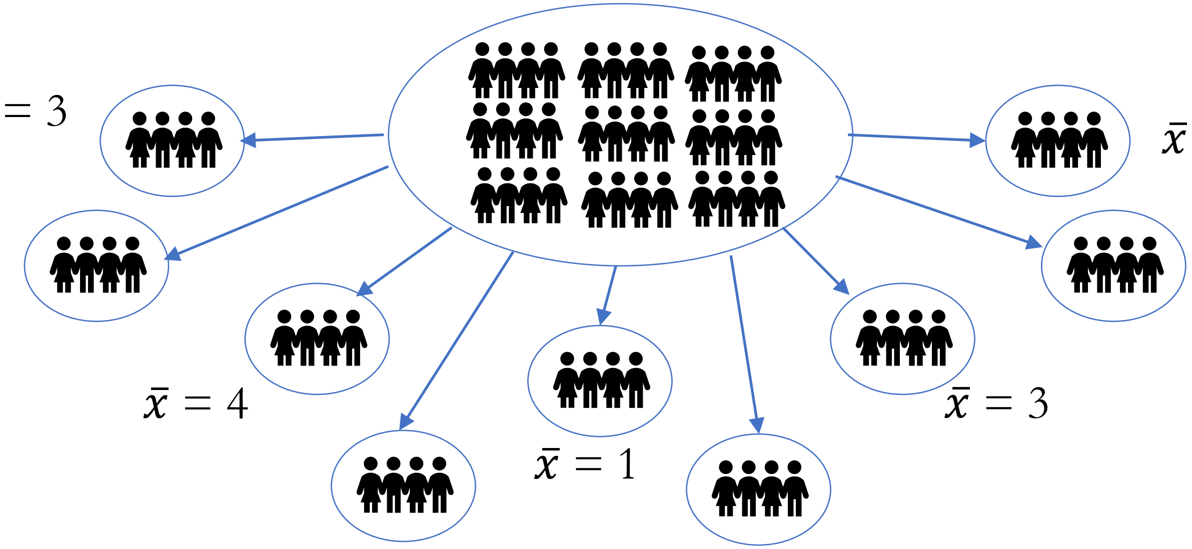
Population and Samples

- If we take several random samples from the population, each of these samples will give us slightly different results, but on average they will be representative of the population.



Sampling Distribution of Means

Population, μ = 3



Sample, $\bar{x} = 3$

$\bar{x} = 2$

$\bar{x} = 4$

$\bar{x} = 3$

$\bar{x} = 1$

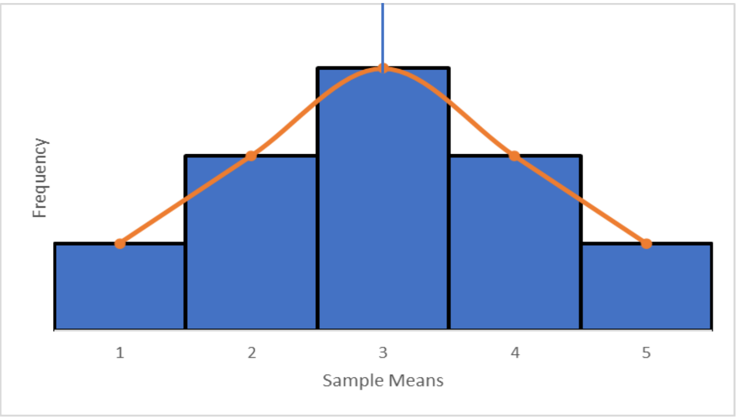
$\bar{x} = 4$

$\bar{x} = 3$

$\bar{x} = 5$

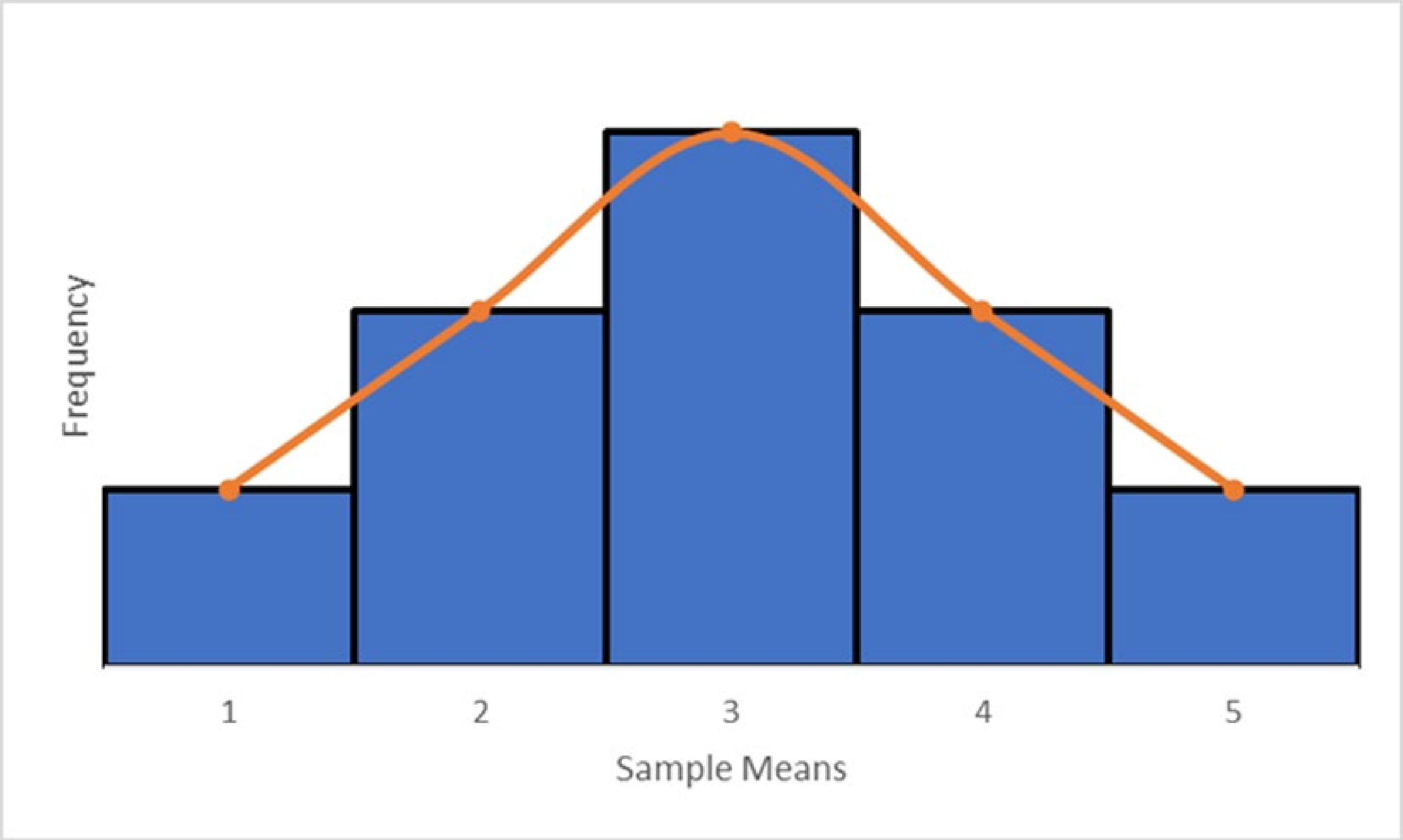
$\bar{x} = 2$

Mean = 3

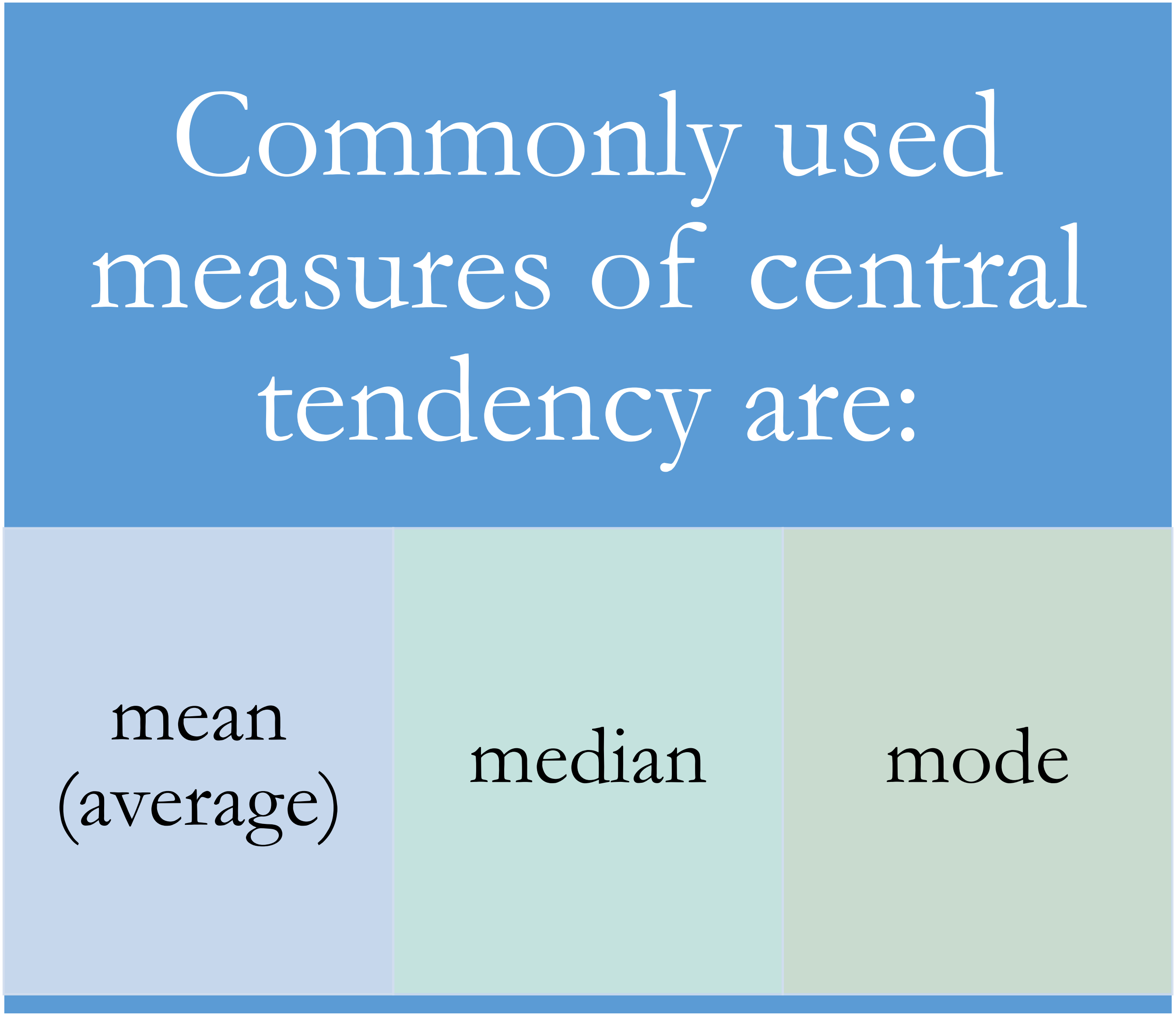


Central Tendency

Measures of central tendency are summary statistics that represent the center point or typical value of a dataset.



Measures of Central Tendency



Commonly used measures of central tendency are:

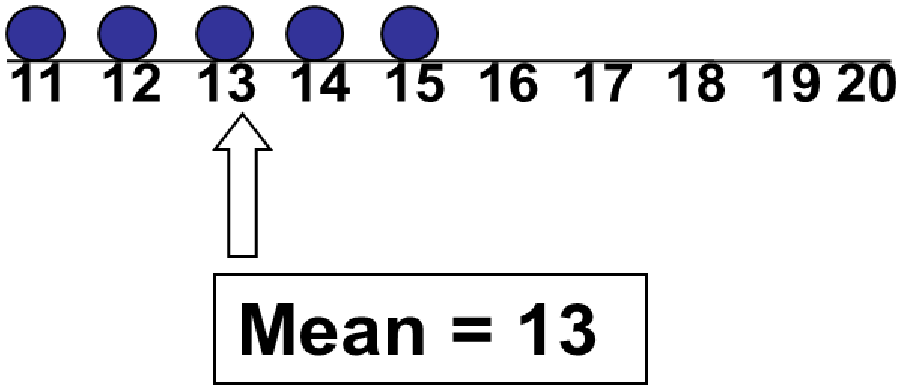
mean (average)

median

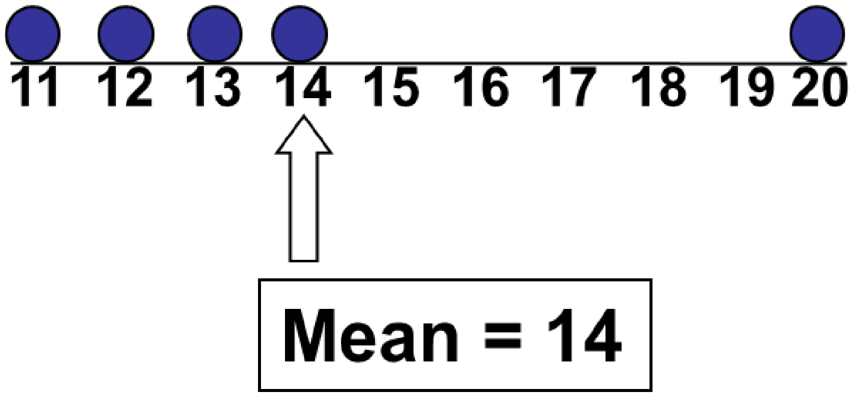
mode

Measures of Central Tendency: The Mean

- The most common measure of central tendency.
- Mean = sum of values divided by the number of values.
- Affected by extreme values (outliers). It is a good measurement if there are no extreme values.



Mean = 13



Mean = 14

$$\frac{11+12+13+14+15}{5}=\frac{65}{5}=13$$

$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

Measures of Central Tendency: Median

- Median is the number in the middle of the data values when the values are in numerical order (smallest to largest).
- Good measurement if there are extreme values

31, 36, 53, 54, 55, 76, 76

- If the number of values is odd, the median is the middle number (center value).

In this example set, the median is **54**

Measures of Central Tendency: Median

20, 31, 36, 53, 54, 55, 76, 76

- If the number of values is even, the median is the average of the two middle numbers. (When n is an even number, the median is the average two center values.)

First find the position: $\text{Median position} = \dfrac{n+1}{2}$ position in the ordered data

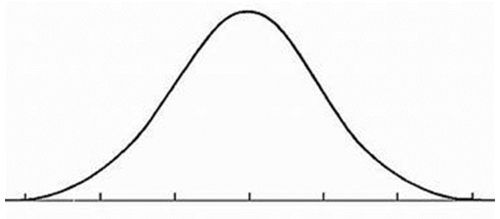
Note: $\dfrac{n+1}{2}$ is **not** the value of the median, only the position of the median in the ranked data.

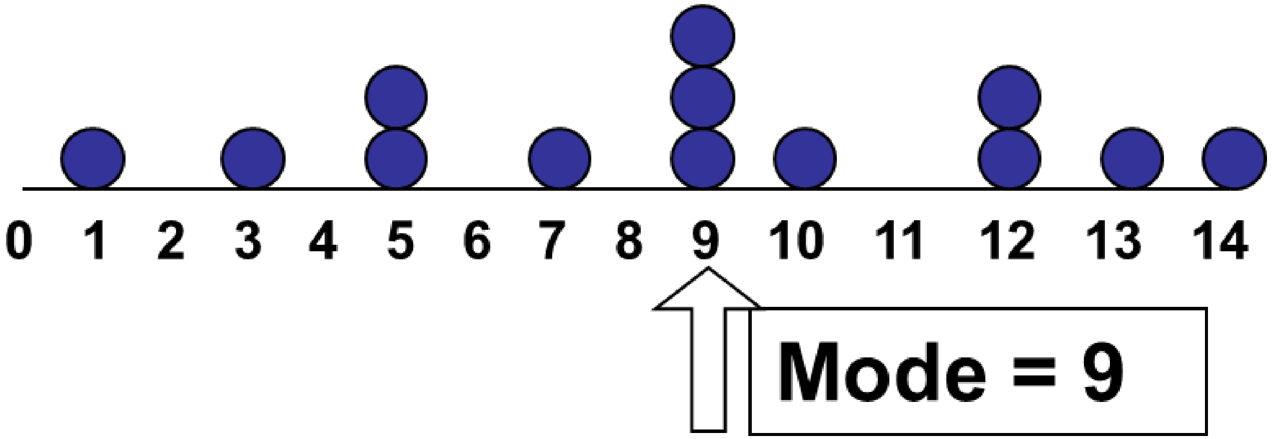
$$20, 31, 36, \underbrace{53, 54,}_{} 55, 76, 76$$

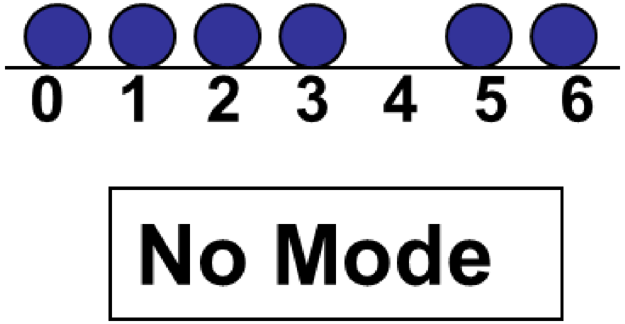
In this example set, the position is $\dfrac{8+1}{2} = 4.5$ Then, the median is $\dfrac{53+54}{2} = \mathbf{53.5}$

Measures of Central Tendency: The Mode

- Mode is the value that occurs most often.
- There may be no mode. There may be several modes.
- Good measurement if the data set is bell shaped.







Measures of Central Tendency: Review Example

$$\begin{array}{rr} \textbf{House Prices:} & \\ & \$2{,}000{,}000 \\ & \$\quad 500{,}000 \\ & \$\quad 300{,}000 \\ & \$\quad 100{,}000 \\ & \underline{\$\quad 100{,}000} \\ \text{Sum} & \$3{,}000{,}000 \end{array}$$

- **Mean:** $\left(\frac{\$3,000,000}{5}\right)$ = **\$600,000**
- **Median:** middle value of ranked data **\$300,000**
- **Mode:** most frequent value **\$100,000**

Measures of Central Tendency: Review Example

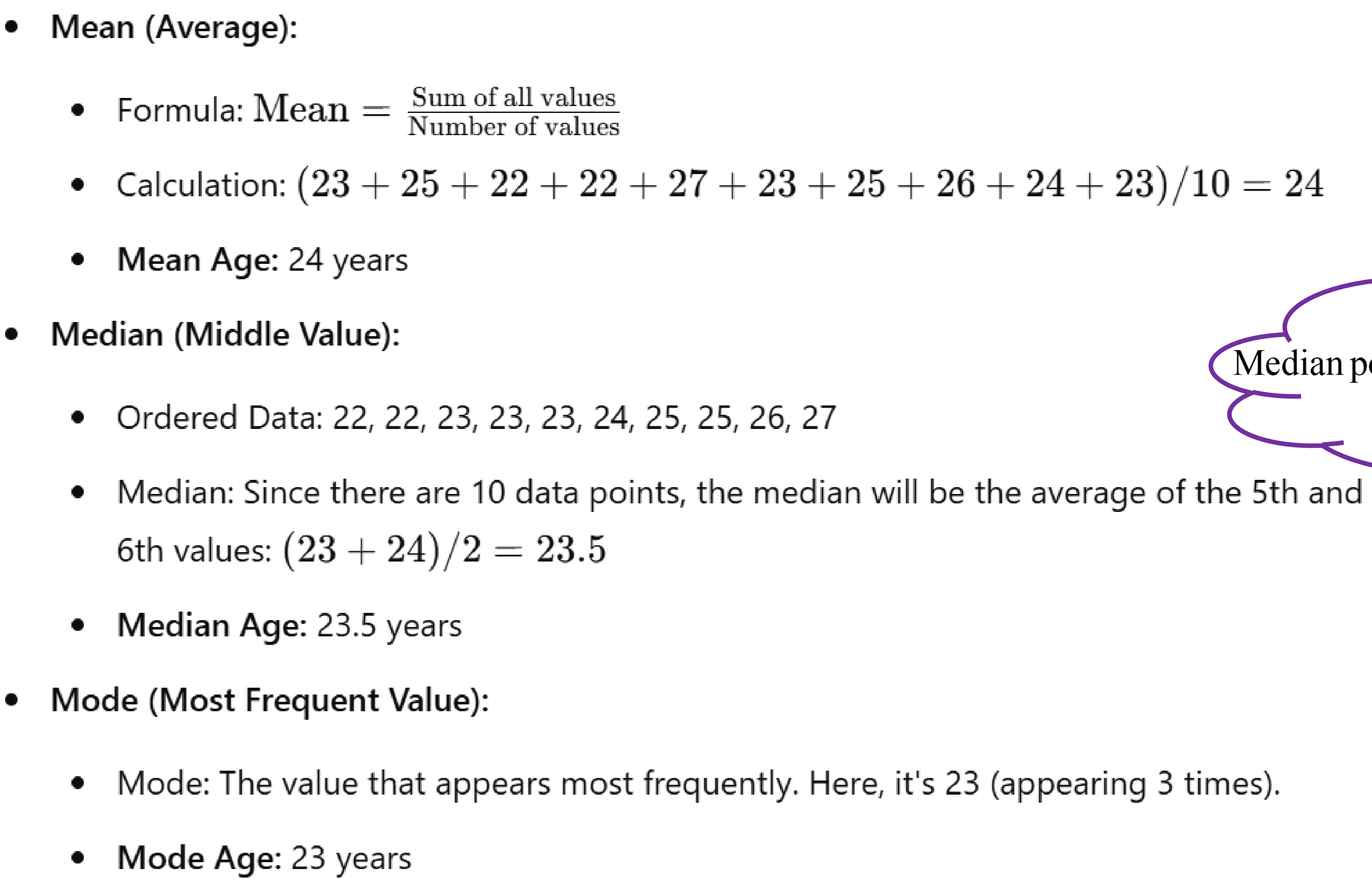
Example Data Set: Ages of participants in a workshop: 23, 25, 22, 22, 27, 23, 25, 26, 24, 23.

Using this data, find:

- Mean?
- Median?
- Mode?

Measures of Central Tendency: Review Example

Example Data Set: Ages of participants in a workshop: 23, 25, 22, 22, 27, 23, 25, 26, 24, 23.



- **Mean (Average):**
  - Formula: $\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$
  - Calculation: $(23 + 25 + 22 + 22 + 27 + 23 + 25 + 26 + 24 + 23)/10 = 24$
  - **Mean Age:** 24 years
- **Median (Middle Value):**
  - Ordered Data: 22, 22, 23, 23, 23, 24, 25, 25, 26, 27
  - Median: Since there are 10 data points, the median will be the average of the 5th and 6th values: $(23 + 24)/2 = 23.5$
  - **Median Age:** 23.5 years
- **Mode (Most Frequent Value):**
  - Mode: The value that appears most frequently. Here, it's 23 (appearing 3 times).
  - **Mode Age:** 23 years

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

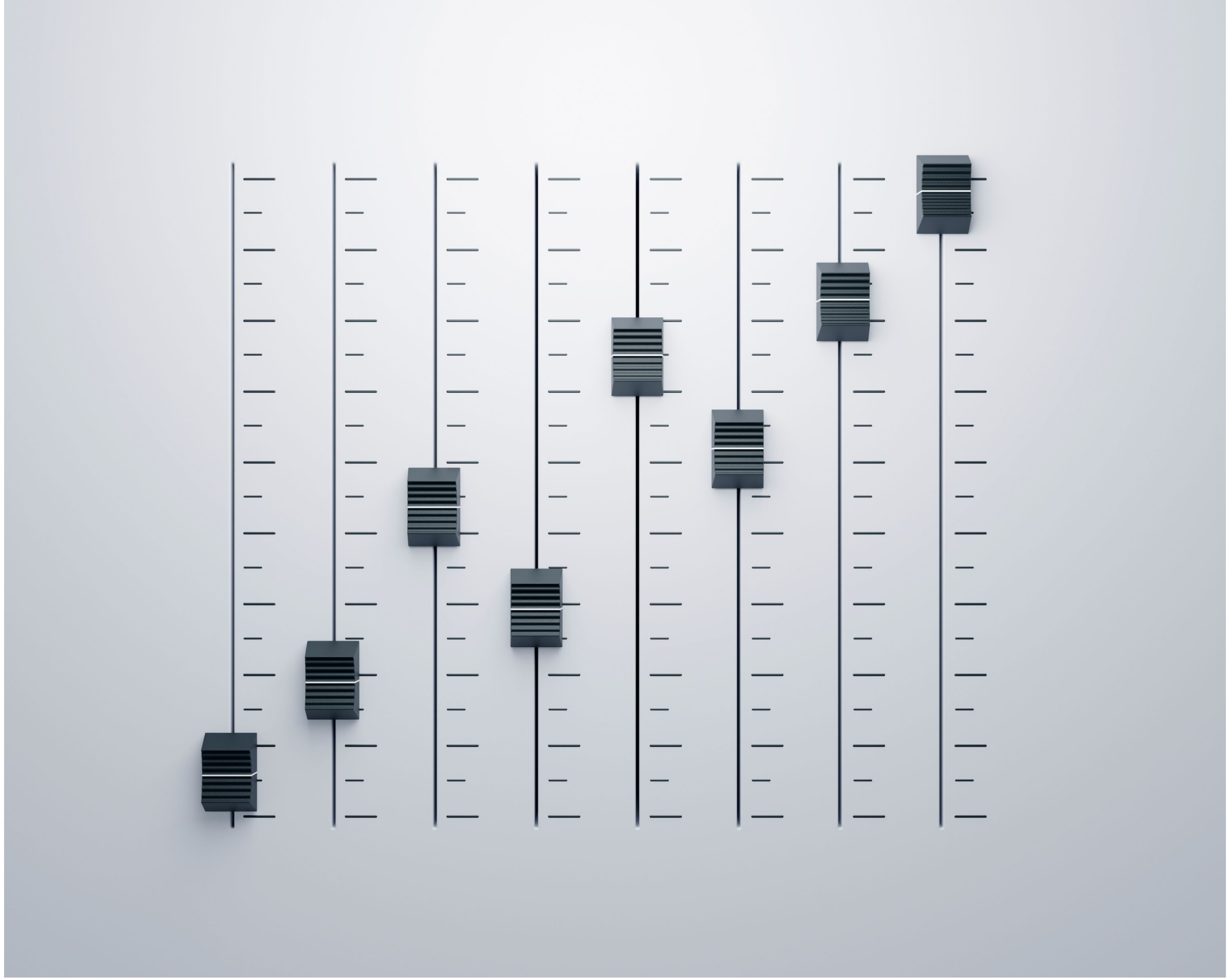
Measures of Central Tendency: Summary

Median = Where data is cut in half (50% of data points to left and 50% to the right)

Mean = The Average of the data values

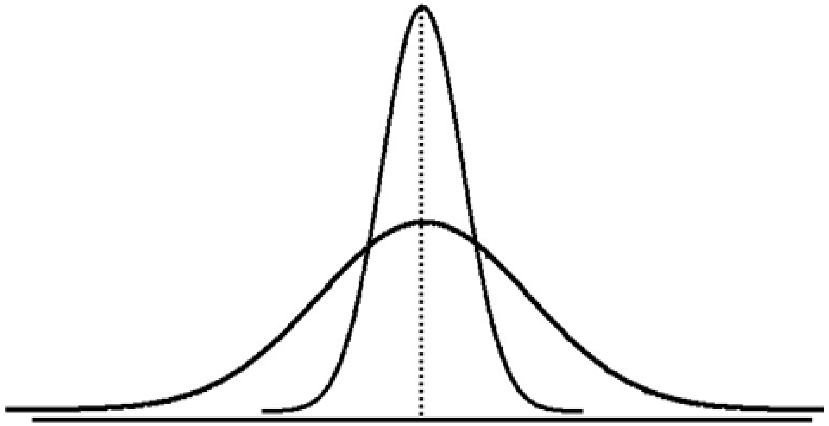
Mode = The Most Common data value

Measures of Variability



Measures of Variation

- Measures of variation give information on the **spread** or **variability** of the data values.

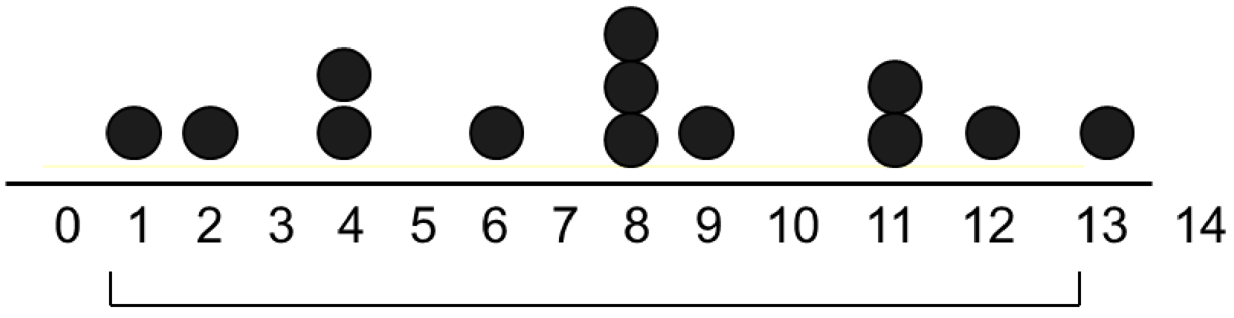


Same center, different variation

Measures of Variation: The Range

- Difference between the largest and the smallest values.

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$



Range = 13 − 1 = 12

Measures of Variation: The Sample Variance

- Variance is a measure of variability that quantifies the spread of a set of data points around their mean value. It tells us how much the data points, on average, differ from the mean. The larger the variance, the more spread out the data points are from their mean value.

- Variance Formula:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

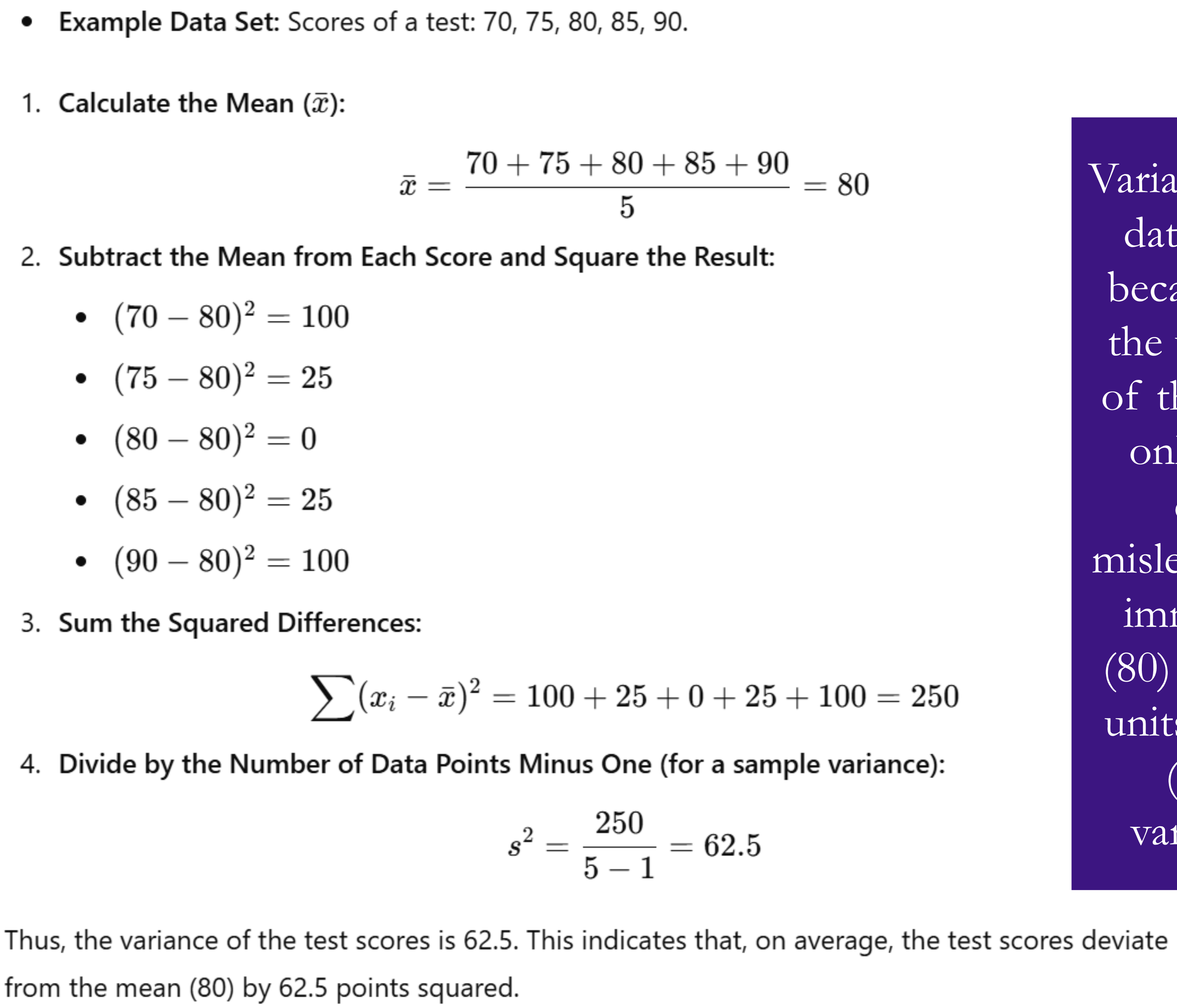
Where

$\overline{X} =$ arithmetic mean

$n =$ sample size

$X_i = i^{\text{th}}$ value of the variable X

Variance Example



- **Example Data Set:** Scores of a test: 70, 75, 80, 85, 90.

1. **Calculate the Mean ($\bar{x}$):**

$$\bar{x} = \frac{70 + 75 + 80 + 85 + 90}{5} = 80$$

2. **Subtract the Mean from Each Score and Square the Result:**
   - $(70 - 80)^2 = 100$
   - $(75 - 80)^2 = 25$
   - $(80 - 80)^2 = 0$
   - $(85 - 80)^2 = 25$
   - $(90 - 80)^2 = 100$

3. **Sum the Squared Differences:**

$$\sum (x_i - \bar{x})^2 = 100 + 25 + 0 + 25 + 100 = 250$$

4. **Divide by the Number of Data Points Minus One (for a sample variance):**

$$s^2 = \frac{250}{5 - 1} = 62.5$$

Thus, the variance of the test scores is 62.5. This indicates that, on average, the test scores deviate from the mean (80) by 62.5 points squared.

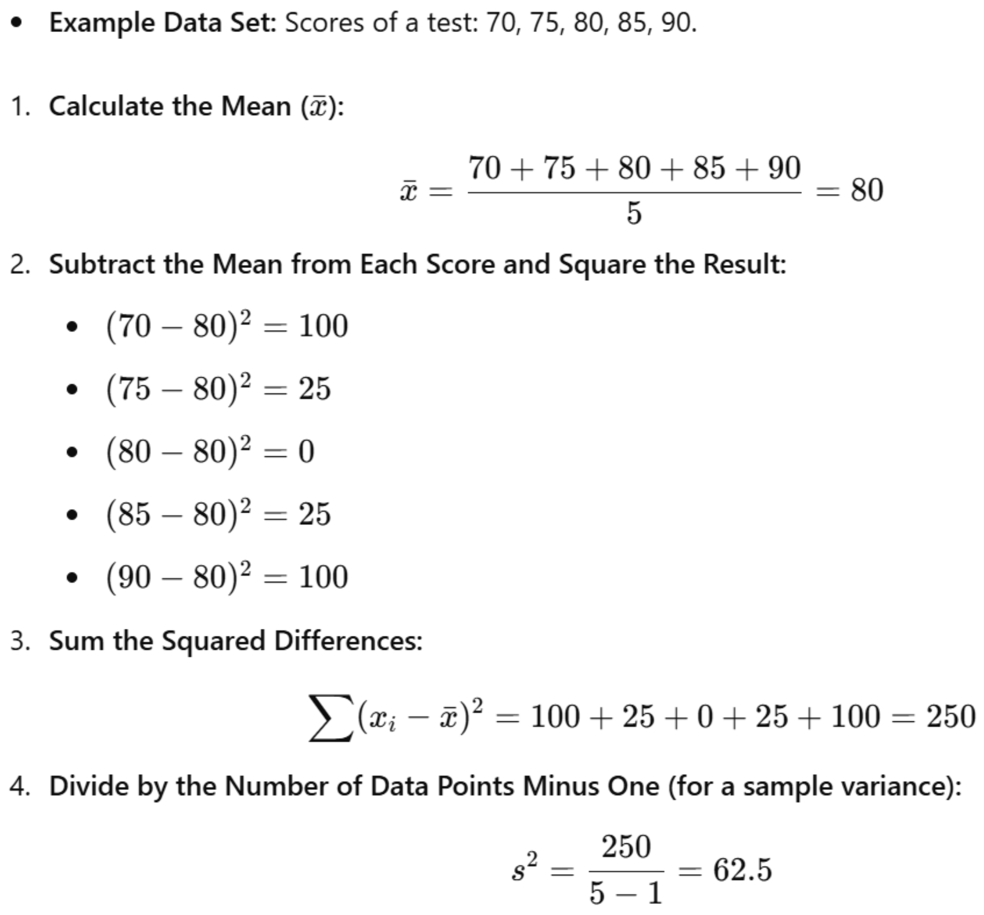
Variance measures how spread out the numbers in a dataset are from their average (mean). However, because we square the differences from the mean, the units of variance are not the same as the units of the data. This squaring is done to make sure we only have positive values to sum up, as negative differences could cancel out positive ones, misleading the true spread. But this number doesn't immediately tell us how far away from the mean (80) the scores typically are, because it's in squared units. To bring it back to the same unit as our data (like scores), we take the square root of the variance, which gives us the standard deviation.

Measures of Variation: The Sample Standard Deviation

- Most commonly used measure of variation.
- Shows variation about the mean.
- It is the square root of the variance.
- The standard deviation is easier to understand and directly tells us about the spread of data around the mean in the original units of measurement.
  - Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

Standard Deviation Example



- **Example Data Set:** Scores of a test: 70, 75, 80, 85, 90.

1. **Calculate the Mean ($\bar{x}$):**

$$\bar{x} = \frac{70 + 75 + 80 + 85 + 90}{5} = 80$$

2. **Subtract the Mean from Each Score and Square the Result:**
   - $(70 - 80)^2 = 100$
   - $(75 - 80)^2 = 25$
   - $(80 - 80)^2 = 0$
   - $(85 - 80)^2 = 25$
   - $(90 - 80)^2 = 100$

3. **Sum the Squared Differences:**

$$\sum (x_i - \bar{x})^2 = 100 + 25 + 0 + 25 + 100 = 250$$

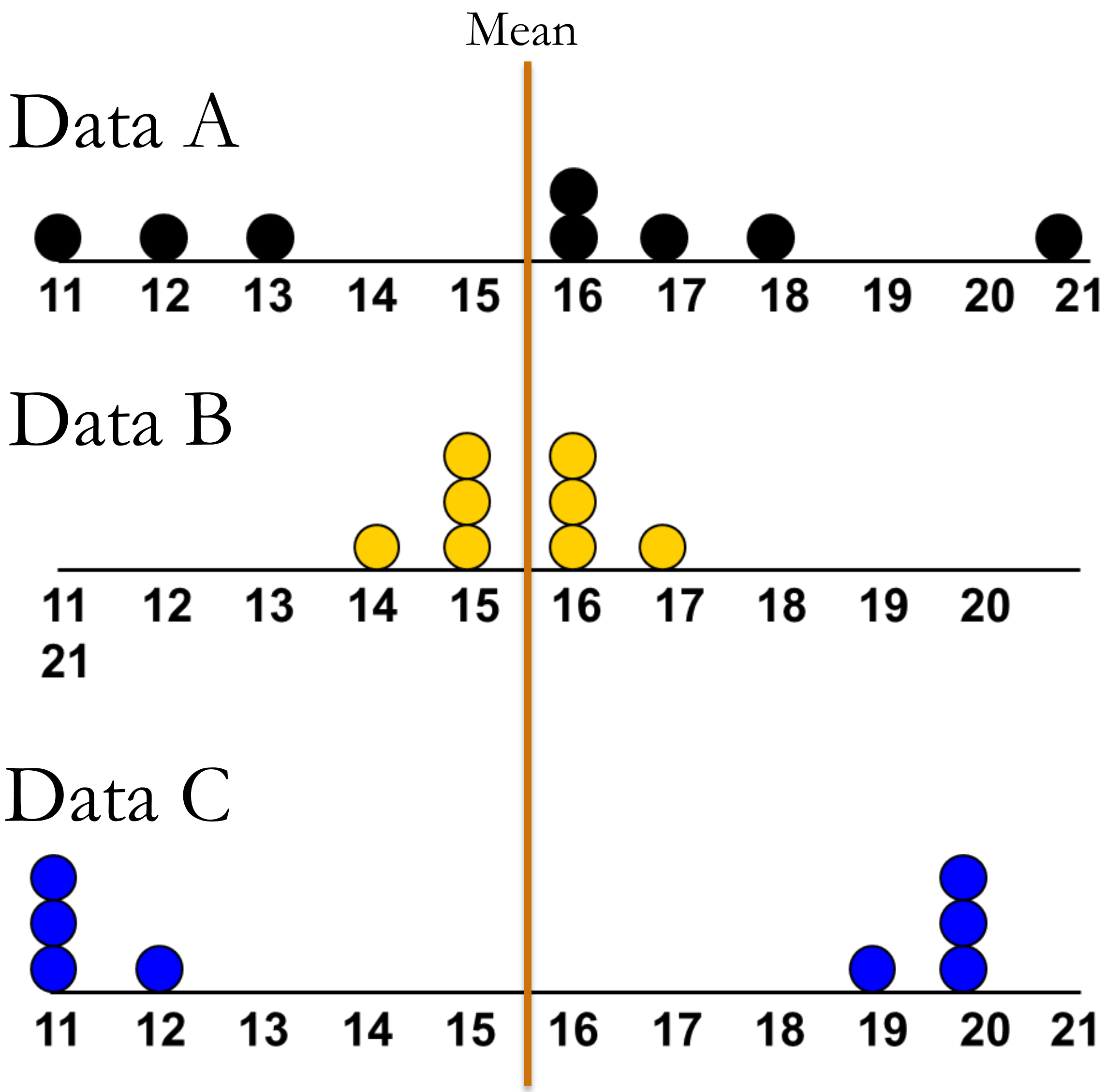
4. **Divide by the Number of Data Points Minus One** (for a sample variance):

$$s^2 = \frac{250}{5 - 1} = 62.5$$

Standard Deviation is the square root of the variance, so $\sqrt{62.5} = 7.9$

Taking the square root of 62.5 gives us approximately 7.9. The standard deviation, about 7.9, tells us that, on average, the test scores deviate from the mean score of 80 by about 7.9 points.

Measures of Variation: Comparing Standard Deviations



Mean = 15.5

$S = 3.338$

Mean = 15.5

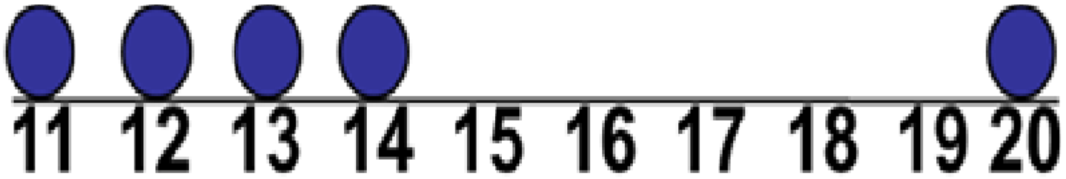
$S = 0.926$

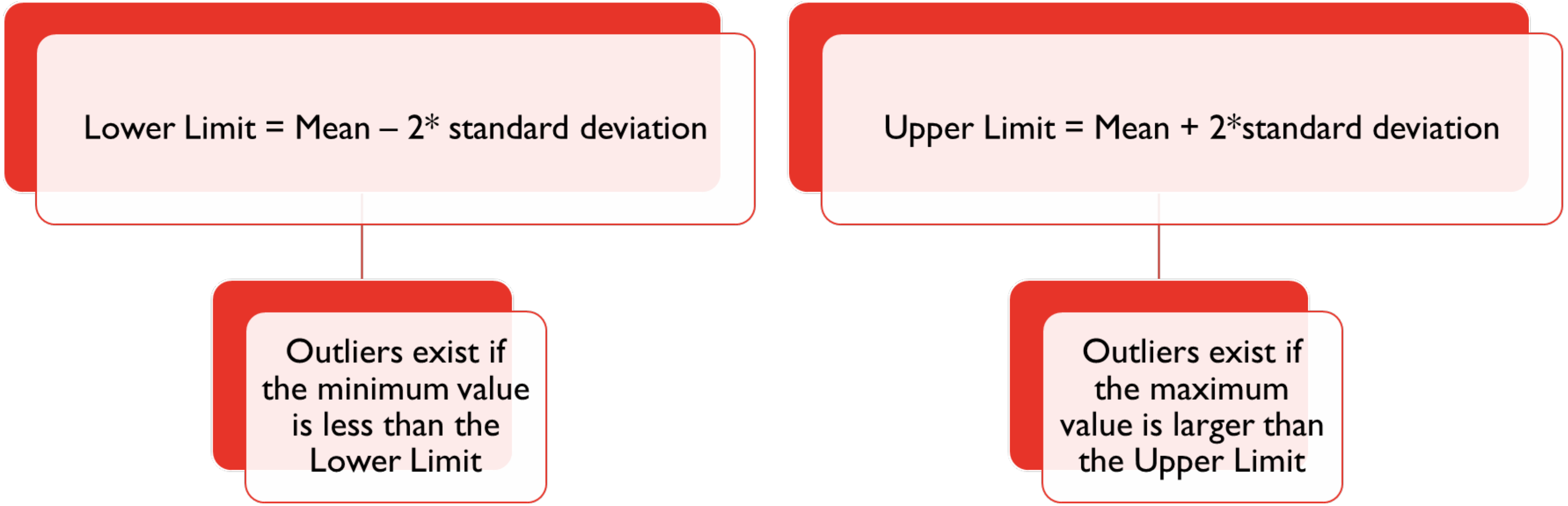
Mean = 15.5

$S = 4.567$

Outliers

An outlier is a data point that differs significantly from other observations.





Lower Limit = Mean – 2* standard deviation

Outliers exist if the minimum value is less than the Lower Limit

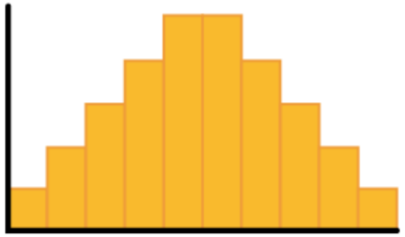
Upper Limit = Mean + 2*standard deviation

Outliers exist if the maximum value is larger than the Upper Limit

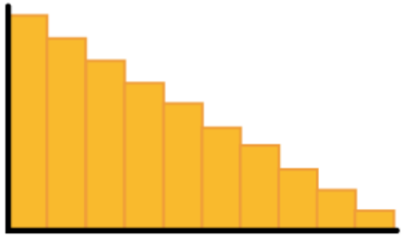
Shape of a Distribution

- Describes how data are distributed.

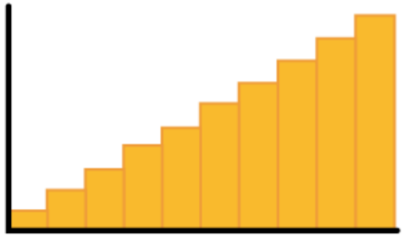
Symmetric (normal) vs skewed and uniform distriutions



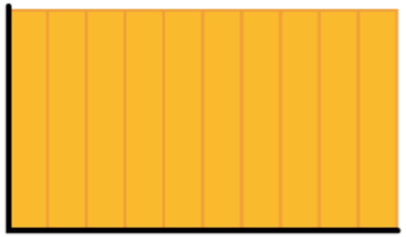
Normal distribution
(unimodal, symmetric,
the “bell curve”)



Right-skewed distribution
(Positively-skewed)

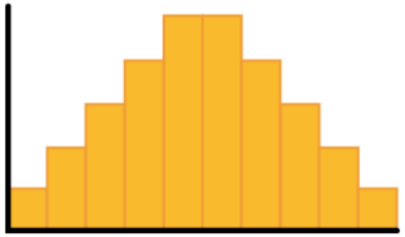


Left-skewed distribution
(Negatively-skewed)

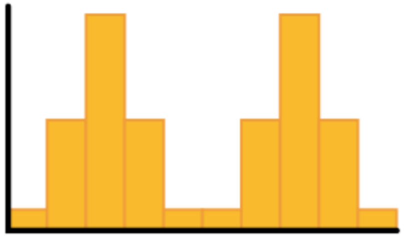


Uniform distribution
(equal spread,
no peaks)

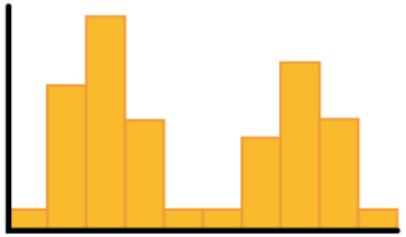
Unimodal vs bimodal distributions



Normal distribution
(unimodal, symmetric,
the “bell curve”)



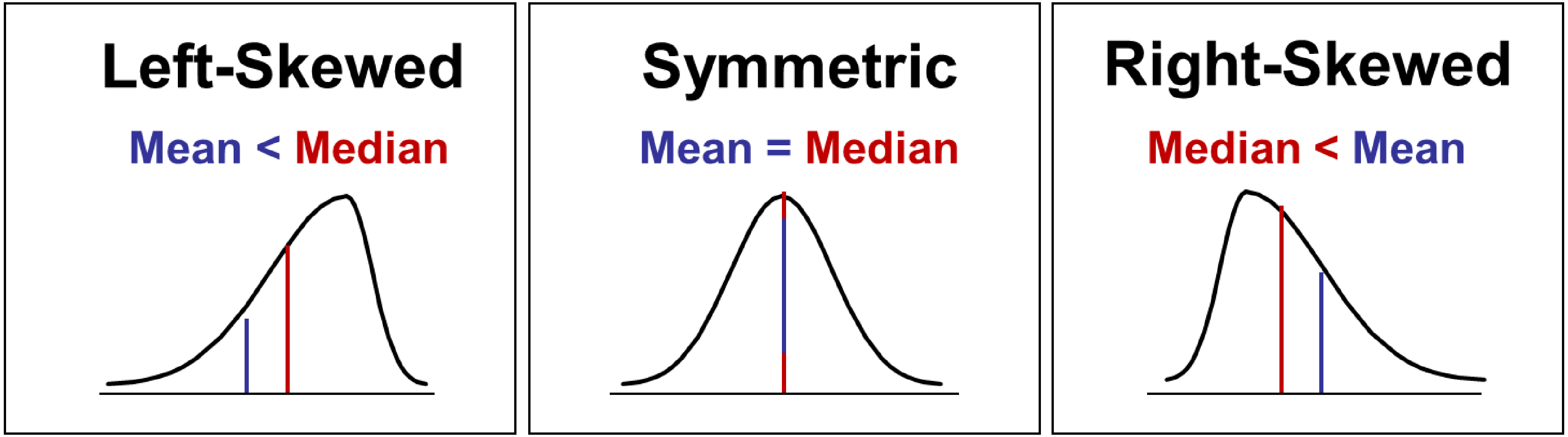
Symmetric bimodal distribution
(two modes)



Non-symmetric bimodal distribution
(two modes)

Shape of a Distribution -- Skewness

- Measures the extent to which data is not symmetrical



Skewness Statistic < 0

Negatively Skewed

0

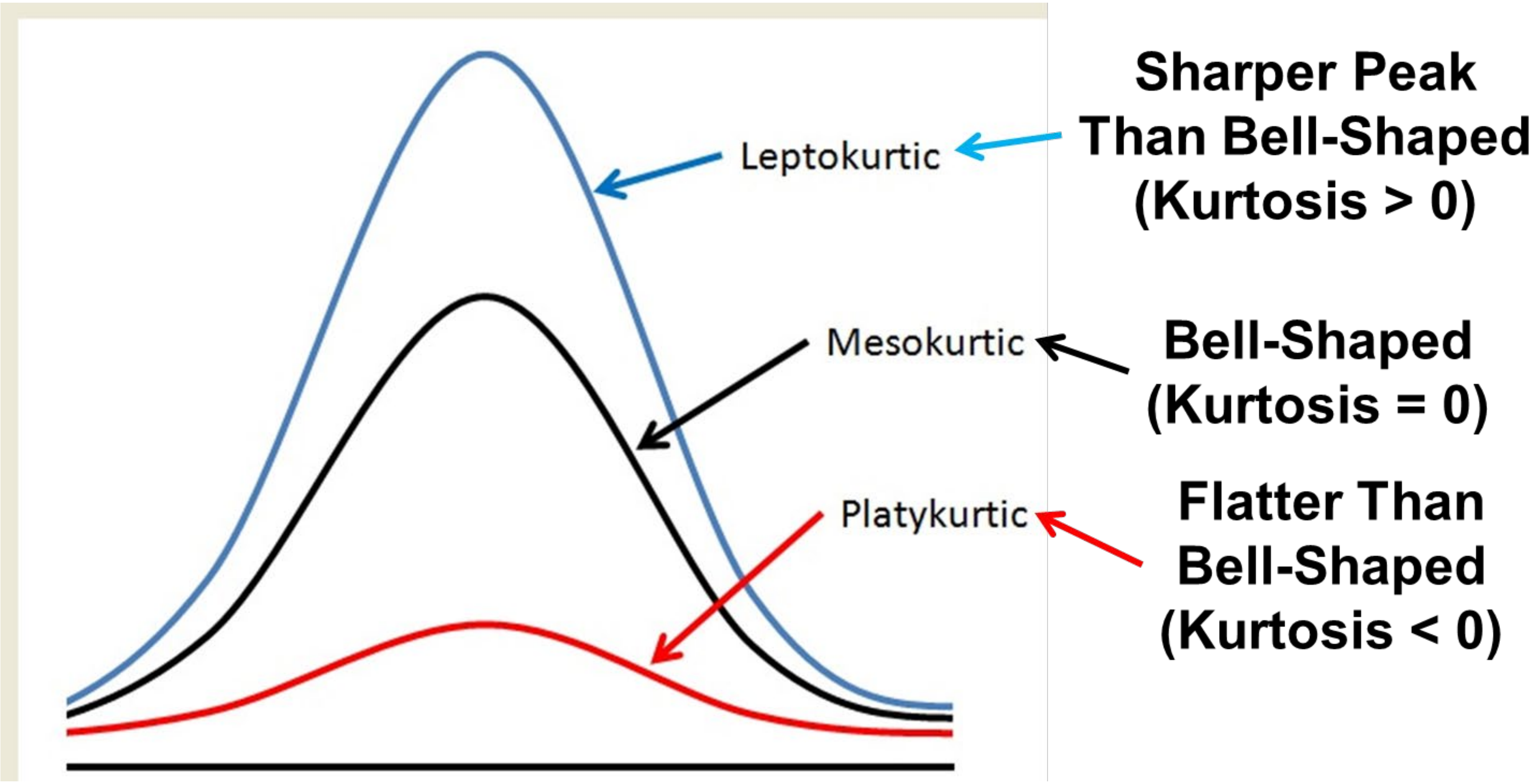
Symmetric

>0

Positively Skewed

Shape of a Distribution -- Kurtosis

- Measures peakedness - how sharply the curve rises approaching the center of the distribution



Sharper Peak Than Bell-Shaped (Kurtosis > 0)

Bell-Shaped (Kurtosis = 0)

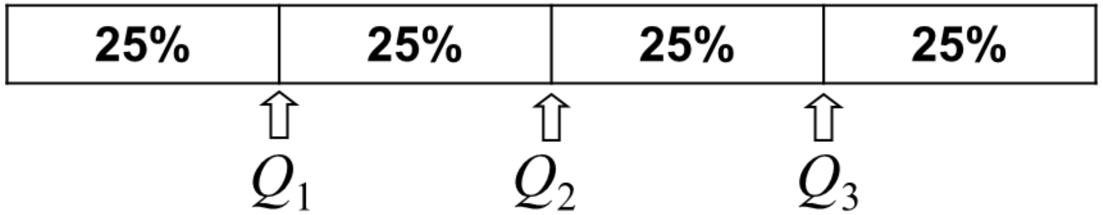
Flatter Than Bell-Shaped (Kurtosis < 0)

Exploring Numerical Data Using Quartiles

- We can visualize the distribution of the values for a numerical variable by computing:
  - The quartiles
  - The five-number summary
  - Constructing a boxplot

Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment.



- The first quartile, Q_1, is the value for which 25% of the values are smaller and 75% are larger.
- Q_2 is the same as the median (50% of the values are smaller and 50% are larger).
- Only 25% of the values are greater than the third quartile.

Quartile Measures: Locating Quartiles

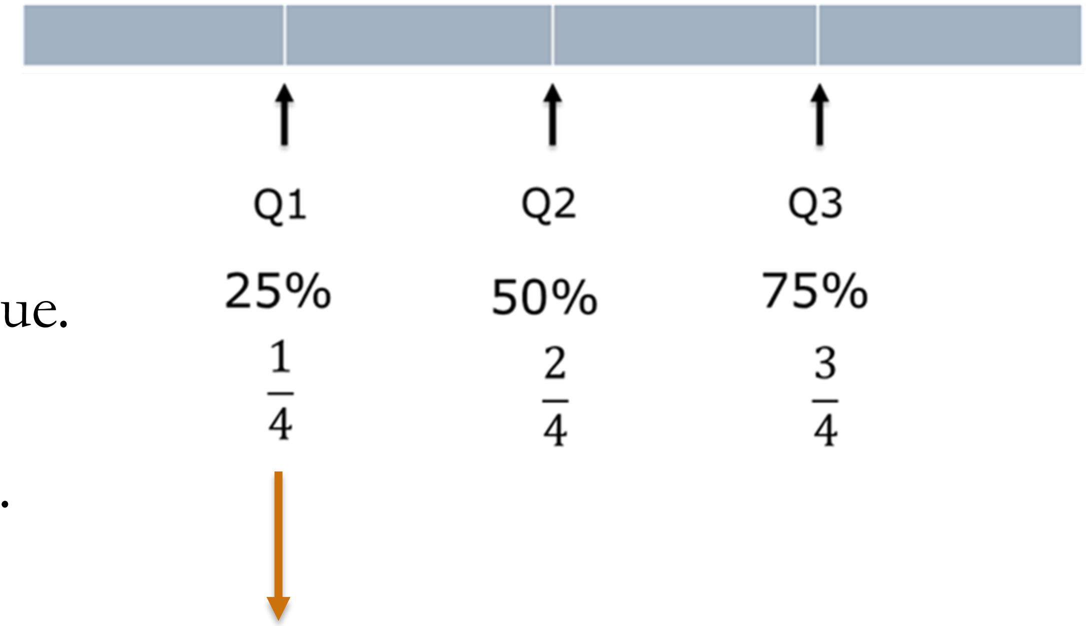
Find a quartile by determining the value in the appropriate position in the ranked data, where:

First quartile position: $Q_1 = \frac{(n+1)}{4}$ ranked value.

Second quartile position: $Q_2 = \frac{(n+1)}{2}$ ranked value.

Third quartile position: $Q_3 = \frac{3(n+1)}{4}$ ranked value.

where n is the number of observed values.



Q1 Q2 Q3

25% 50% 75%

$\frac{1}{4}$ $\frac{2}{4}$ $\frac{3}{4}$

□ Q1 = $\frac{1}{4} * (n + 1)$

Example Finding Quartiles

- **Question**: Find Q1 of the following data set of observations:

108,103,252,121,93,57,40,53,22,116,98

Start with arranging scores in ascending order and determine the number of observations.

22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252

n represent the number of observations ⟶ $\boldsymbol{n} = 11$

Formula ⟶ $Q1 = \frac{1}{4} * (n + 1)$

$$Q1 = \frac{1}{4} * (11 + 1)$$

$$Q1 = \frac{12}{4}$$

$$Q1 = 3$$

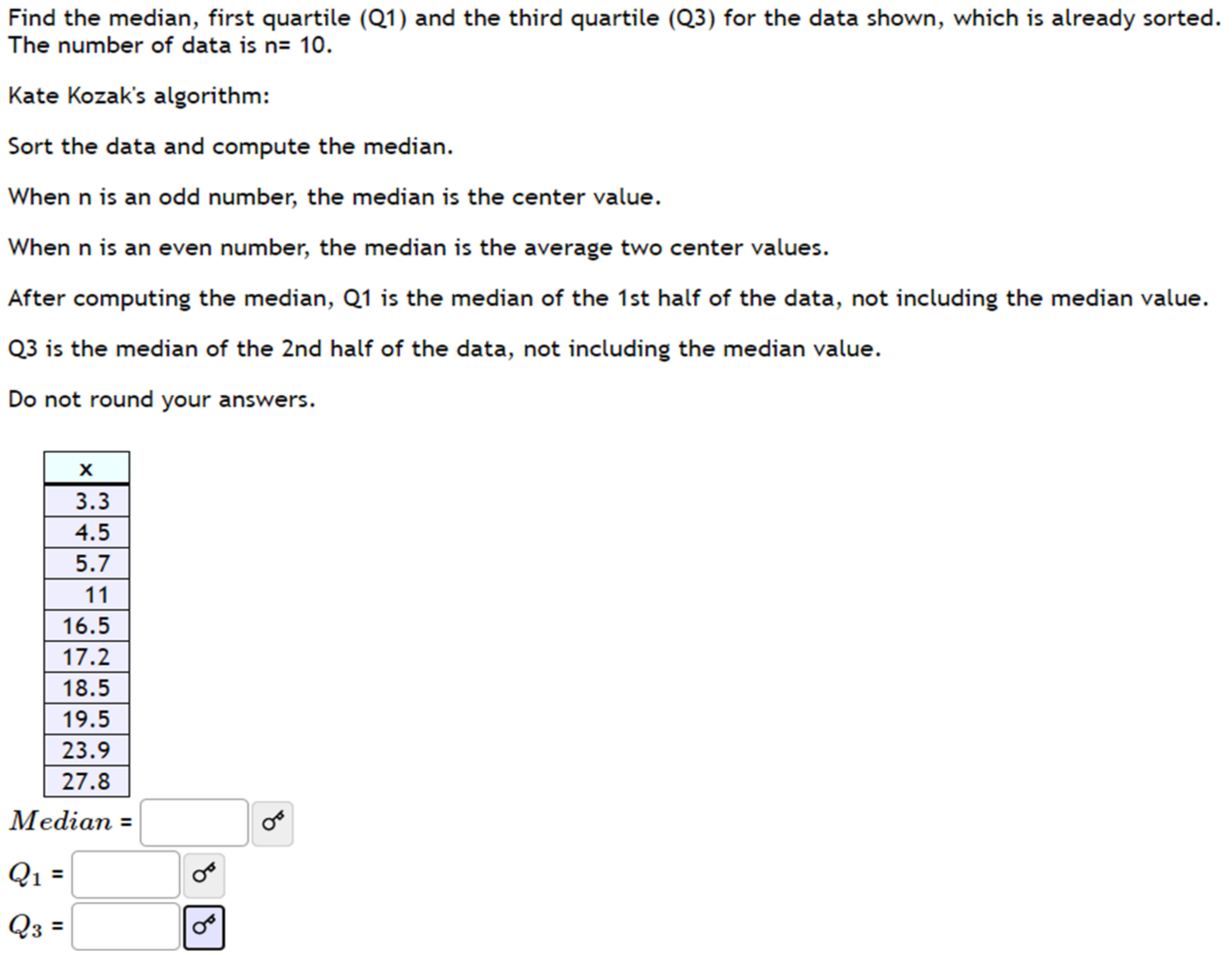
Location ⟶ Q1 = 3rd score

- We found Q1 = 3rd observation

$22,40,\boxed{53},57,93,98,103,108,116,121,252$

$\downarrow$

Q1



Find the median, first quartile (Q1) and the third quartile (Q3) for the data shown, which is already sorted.
The number of data is n= 10.

Kate Kozak's algorithm:

Sort the data and compute the median.

When n is an odd number, the median is the center value.

When n is an even number, the median is the average two center values.

After computing the median, Q1 is the median of the 1st half of the data, not including the median value.

Q3 is the median of the 2nd half of the data, not including the median value.

Do not round your answers.

x
3.3
4.5
5.7
11
16.5
17.2
18.5
19.5
23.9
27.8

$Median$ = □

Q_1 = □

Q_3 = □

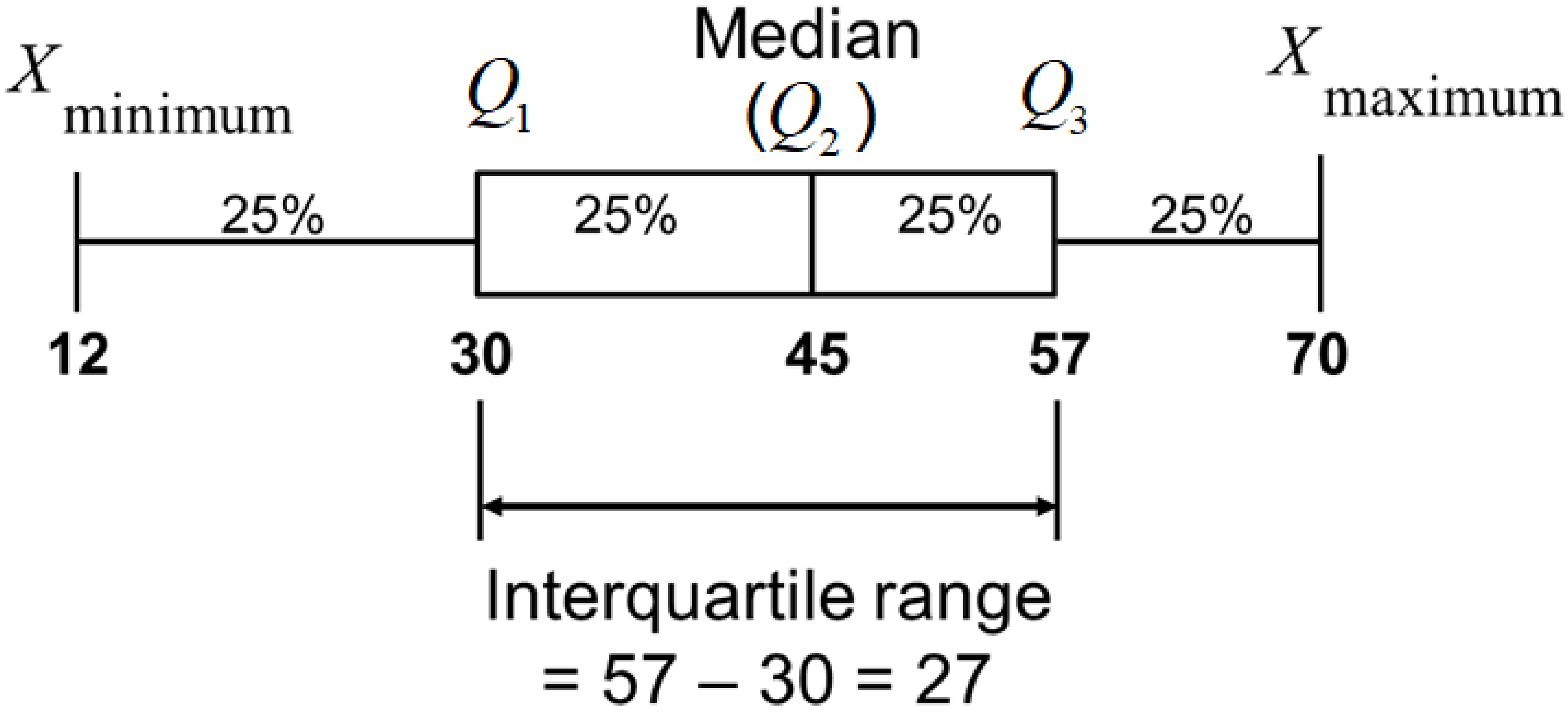
n = 10
The median is the average of the two center values.

(16.5 + 17.2)/2 = 16.85

Q1 Take the median of the lower 5 values
Q1 = 5.7

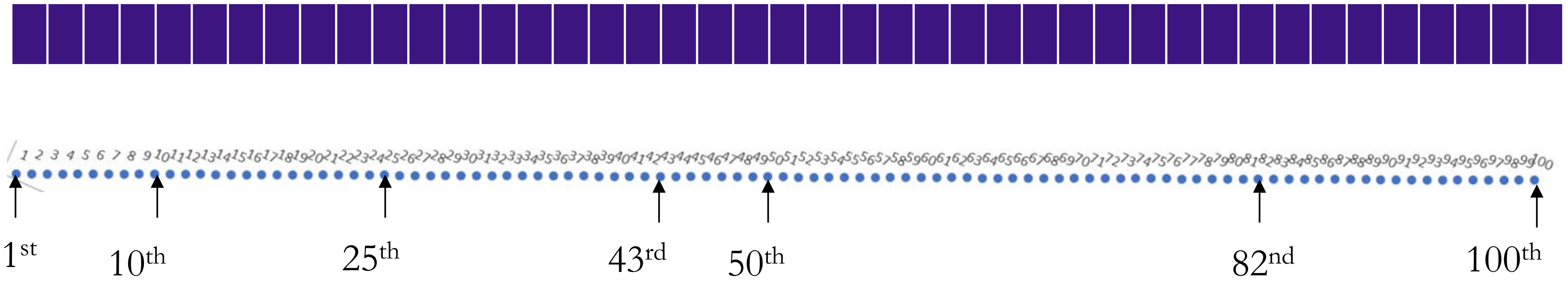
Q3 Takes the median of the upper 5 values
Q3 = 19.5

Five Number Summary and Interquartile Range



Percentiles

- Divides ordered data into 100 equal parts

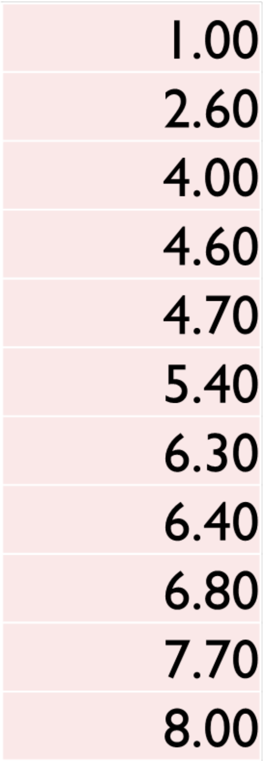


Formula for Percentile Rank Calculation

$$\text{Percentile Rank} = \frac{Percentile}{100} * (n+1)$$

n represents the number of observations

The formula does not give you the value, it gives you the place (location) as percentile rank.



1.00
2.60
4.00
4.60
4.70
5.40
6.30
6.40
6.80
7.70
8.00

Sort the data

Percentile Rank = k(n+1)/100

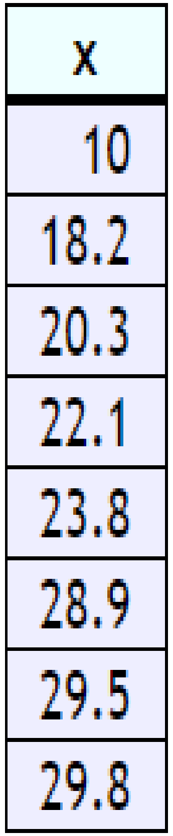
k = percentile

n = sample size

- If i is an integer, then the kth percentile is the data value in the ith position.
- If i is not an integer, then take the average of the ith and (i+1)th data value.

Example Percentile Calculation when the Percentile Rank is not an integer

Calculate the 45th percentile of the data shown using the (n+1)*k/100 method.



x
10
18.2
20.3
22.1
23.8
28.9
29.5
29.8

Make sure the data is sorted.

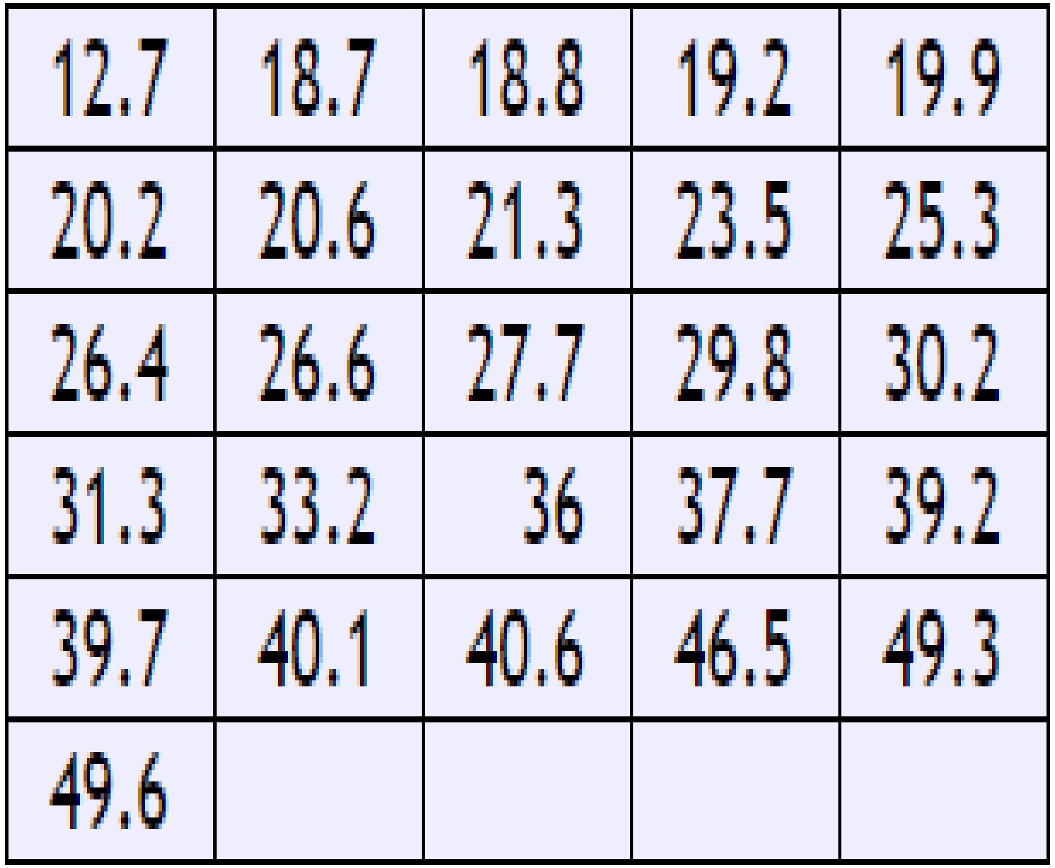
The sample size n = 8, and k = 45 Apply the formula

Percentile Rank = 45(8+1)/100 = 4.05 this value is not integer. So, take the average of the 4th and 5th value

45th percentile =(22.1 + 23.8)/2 = 22.95

Example Percentile Calculation when i is not an integer

Find the 50th percentile from the following sorted list of 26 numbers. Use the (n+1)*k/100 method of calculating the percentage.

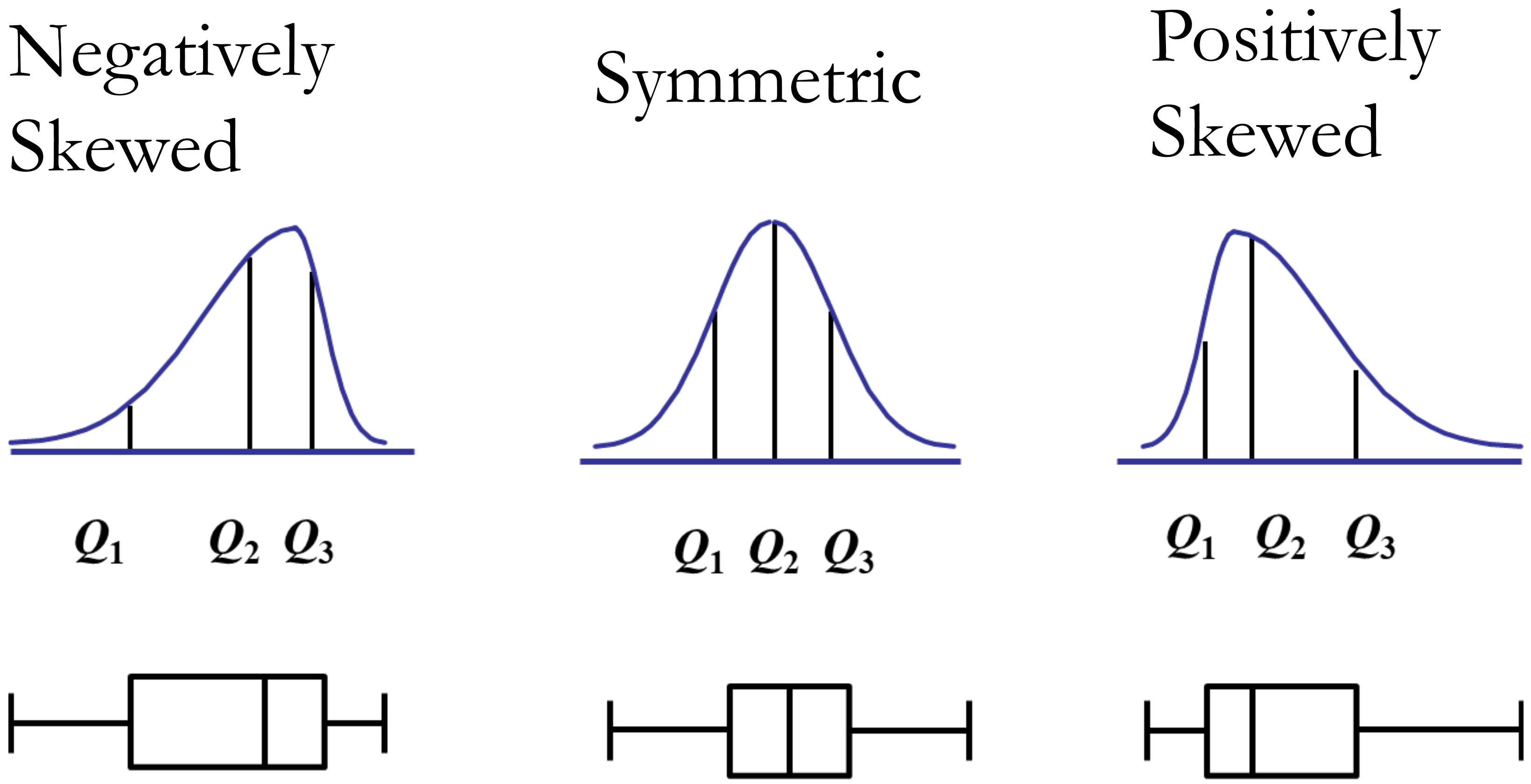


12.7	18.7	18.8	19.2	19.9
20.2	20.6	21.3	23.5	25.3
26.4	26.6	27.7	29.8	30.2
31.3	33.2	36	37.7	39.2
39.7	40.1	40.6	46.5	49.3
49.6				

Note the data is already sorted.
The 50th percentile for n= 26 is
percentile rank = 50(26+1)/100 = 13.5
Take the average of the 13th and 14th data value.

50th percentile = (27.7 + 29.8)/2 =28.75

Distribution Shape and the Boxplot



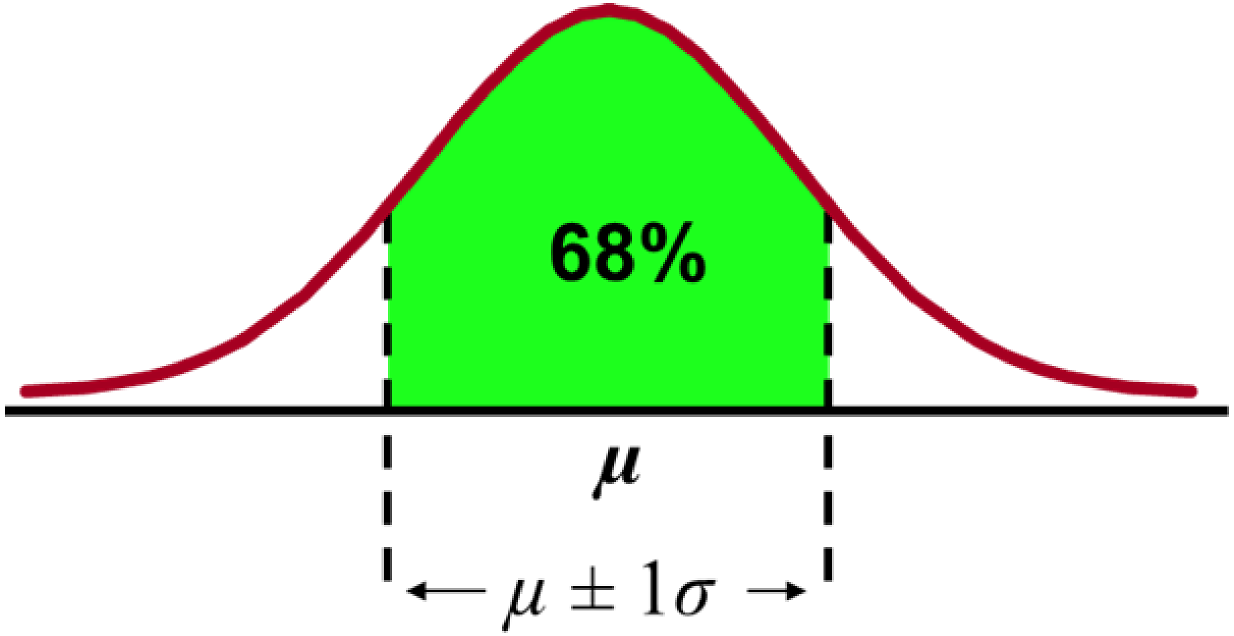
Negatively Skewed

Symmetric

Positively Skewed

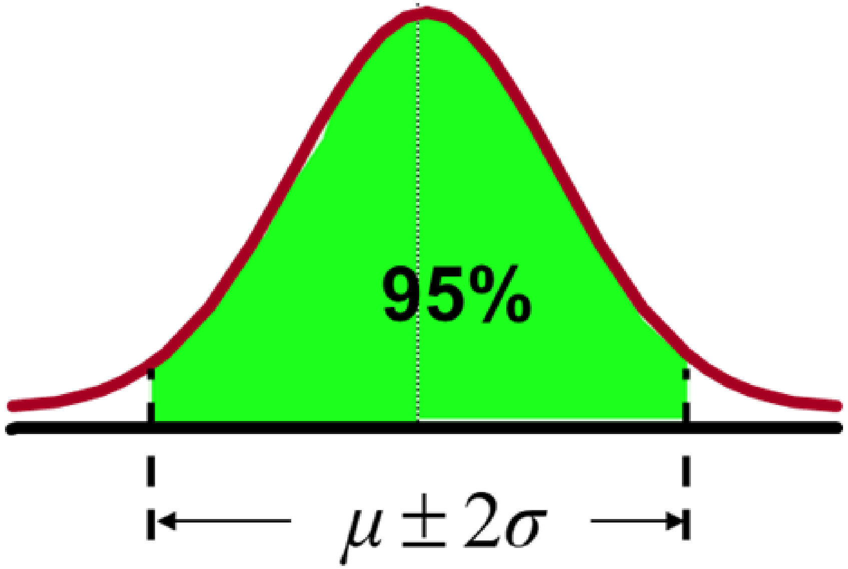
The Empirical Rule

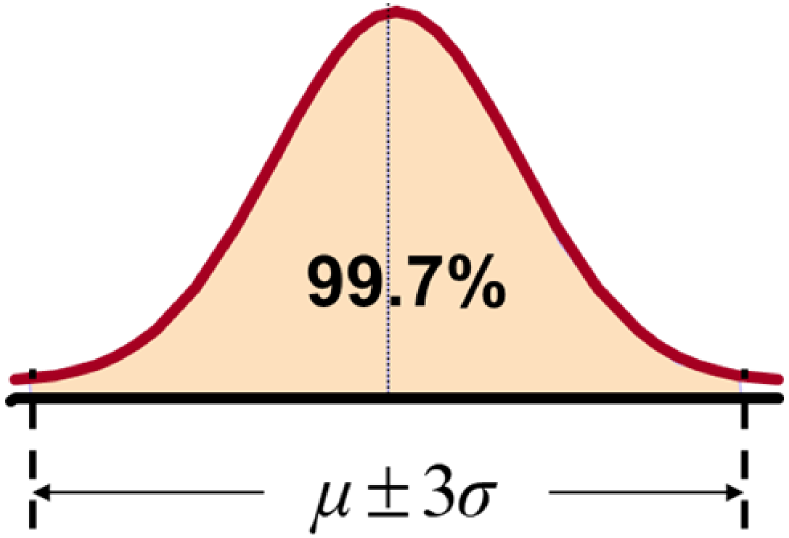
- The empirical rule approximates the variation of data in a symmetric mound-shaped distribution.
- Approximately 68% of the data in a symmetric mound shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$.



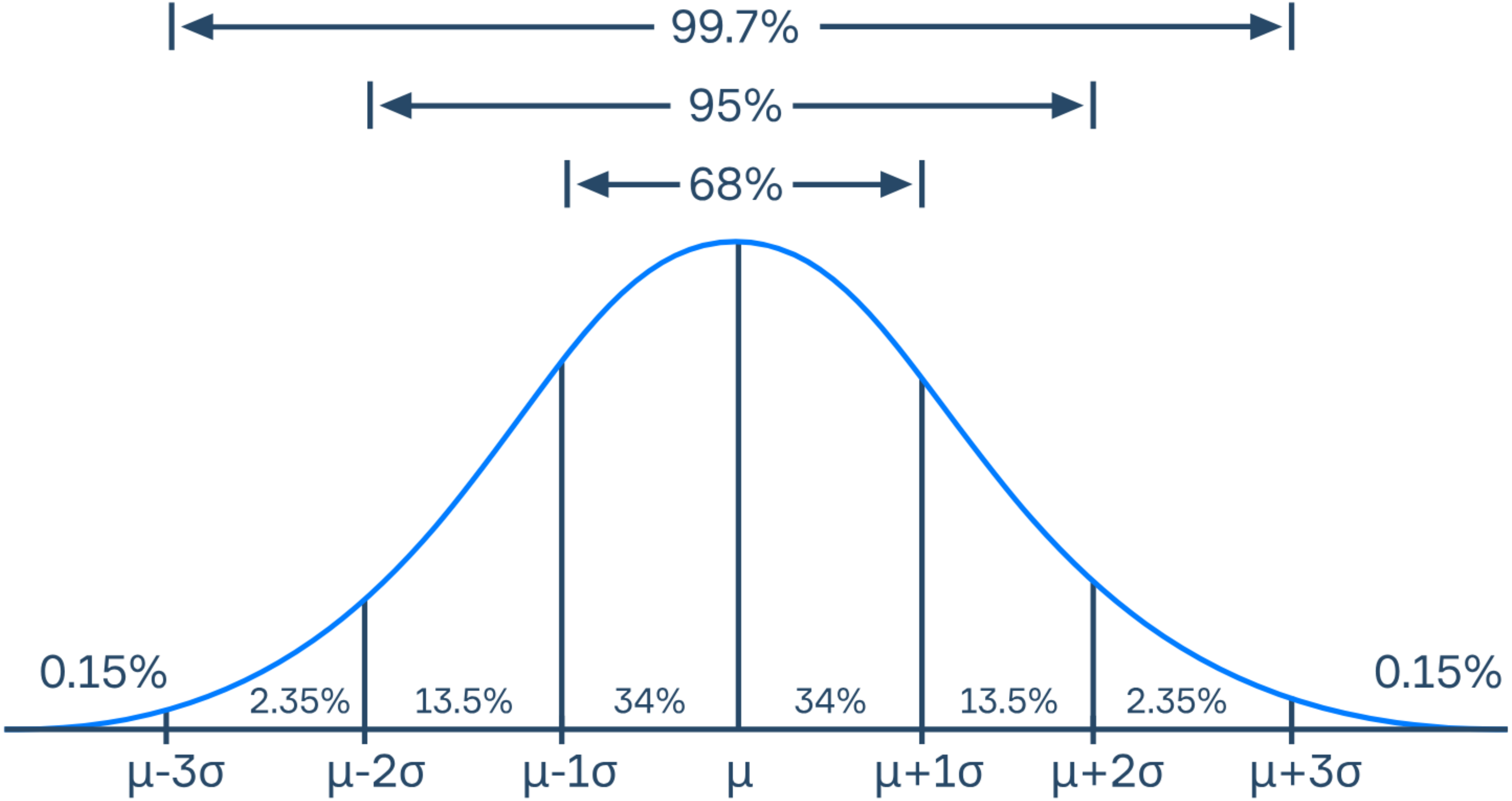
The Empirical Rule

- Approximately 95% of the data in a symmetric mound-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$.
- Approximately 99.7% of the data symmetric mound-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$.





Empirical Rule



Z scores

$$Z = \frac{X \ - Mean}{Standard\ deviation}$$

- It is a measure that tells us how many standard deviations a value is from the mean.

  If Z < 0, the X value is below the mean.

  If Z > 0, the X value is above the mean.

The Covariance

- The covariance measures the strength of the **linear relationship** between two numerical variables (X and Y).

  - The sample covariance:

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n-1}$$

Important:

- Only concerned with the strength of the relationship.
- No causal effect is implied.

Interpreting Covariance

- **Covariance** between two variables:

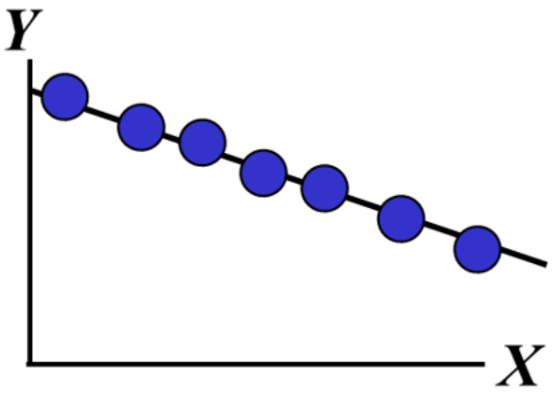
$\text{cov}(X,Y) > 0 \longrightarrow X$ and Y tend to move in the same direction.

$\text{cov}(X,Y) < 0 \longrightarrow X$ and Y tend to move in opposite directions.

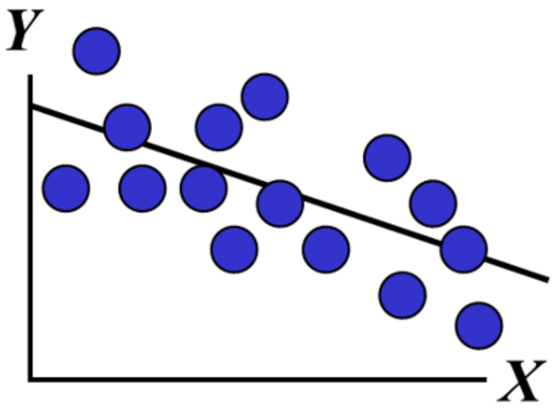
$\text{cov}(X,Y) = 0 \longrightarrow X$ and Y are independent.

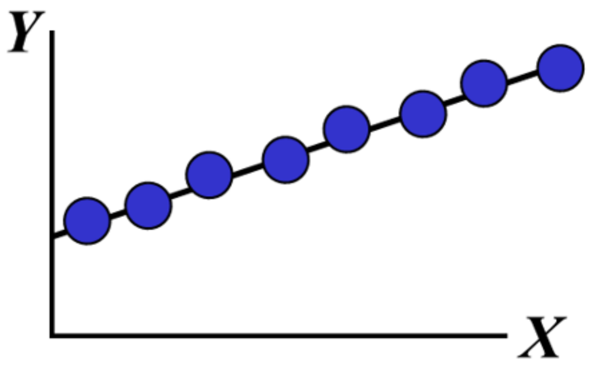
- The covariance has a major flaw:
  - It is not possible to determine the relative strength of the relationship from the size of the covariance.

Scatter Plots of Sample Data to Visualize the Covariance

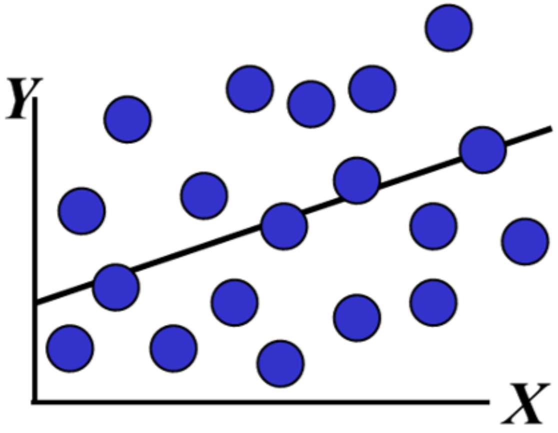


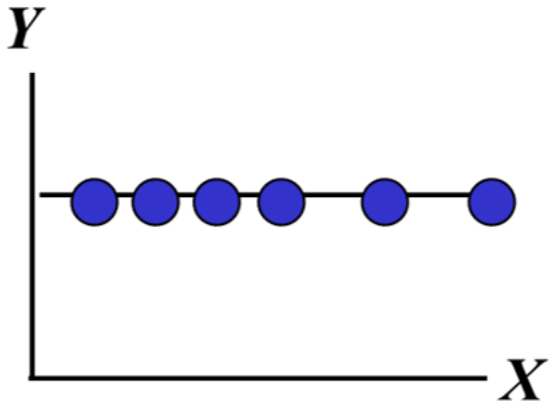
X and Y tend to move in the opposite directions





X and Y tend to move in the same direction





X and Y are independent

Coefficient of Correlation

- Measures the **relative strength of the linear relationship** between two numerical variables.

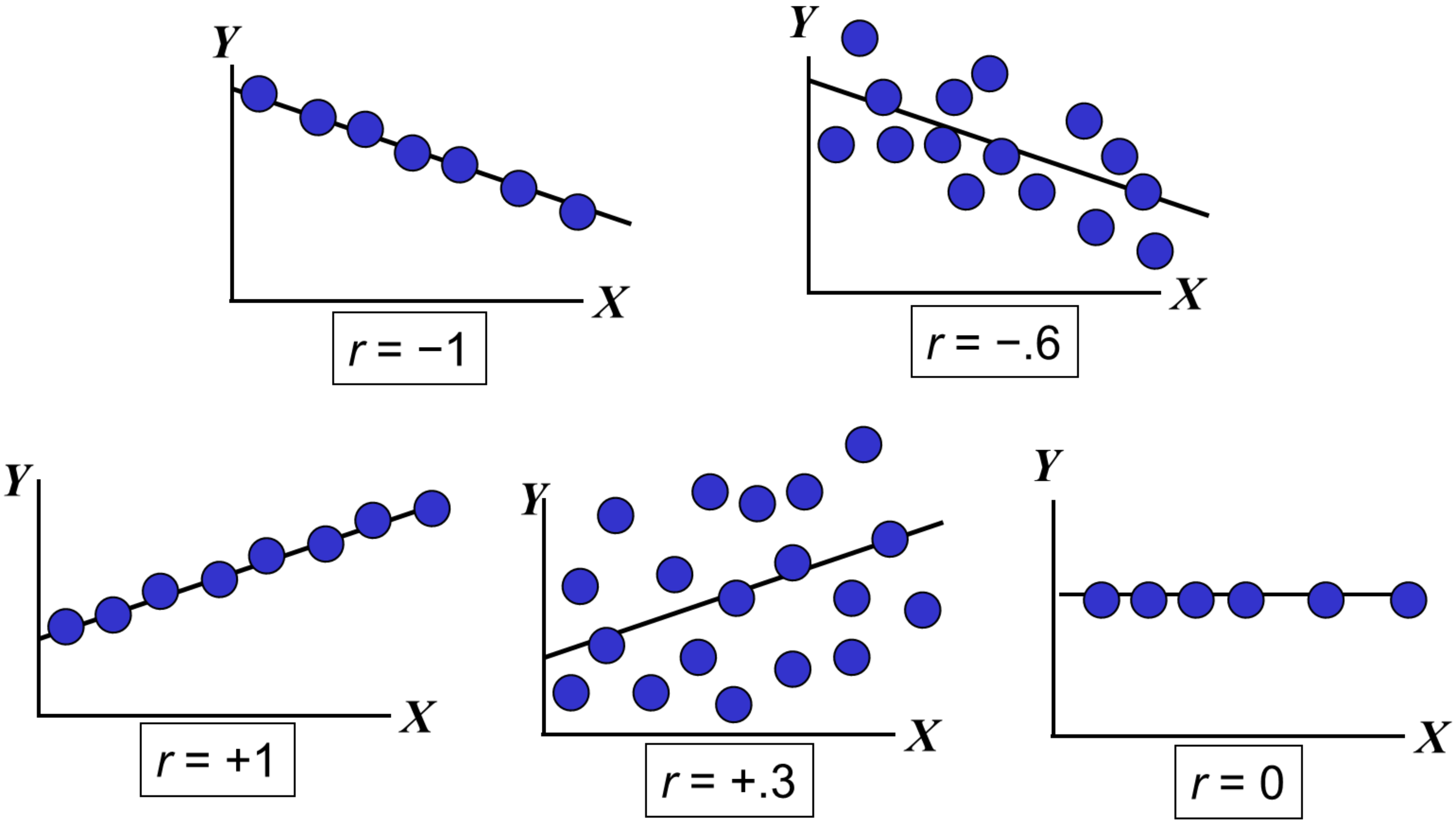
- Sample coefficient of correlation: $r = \dfrac{\operatorname{cov}(X,Y)}{S_X S_Y}$

Features of the Coefficient of Correlation

The Coefficient of correlation has the range between -1 and 1.

- The closer to -1, the stronger the negative linear relationship.
- The closer to 1, the stronger the positive linear relationship.
- The closer to 0, the weaker the linear relationship.

Scatter Plots of Sample Data with Various Coefficients of Correlation



r = −1

r = −.6

r = +1

r = +.3

r = 0