Knowledge Discovery & Data Mining
— Data Exploration: Descriptive Statistics—

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Based on the original version by Professor Yizhou Sun

Outline

- Central Tendency
  - Mean, Median, Mode, Midrange
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation

Population vs. Sample

- **A set of data points is a sample from a population:**
  - A **population** is the entire set of objects or events under study.
    - E.g., population can be hypothetical “all students” or all students in this class.
    - E.g., population can be all the houses in a region
  - A **sample** is a “representative” subset of the objects or events under study. Needed because it’s impossible or intractable to obtain or compute with population data.

Basic Statistical Descriptions of Data

An overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

- Central Tendency
- Dispersion of the Data

Central Tendency

Suppose that we have some attribute X, like salary, which has been recorded for a set of objects. Let x1, x2, . . . , xN be the set of N observed values or observations for X. Here, these values may also be referred to as the data set (for X). If we were to plot the observations for salary, where would most of the values fall?

Outline

- Central Tendency
  - Mean, Median, Mode,
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: n is sample size and N is population size.

  - Weighted arithmetic mean:

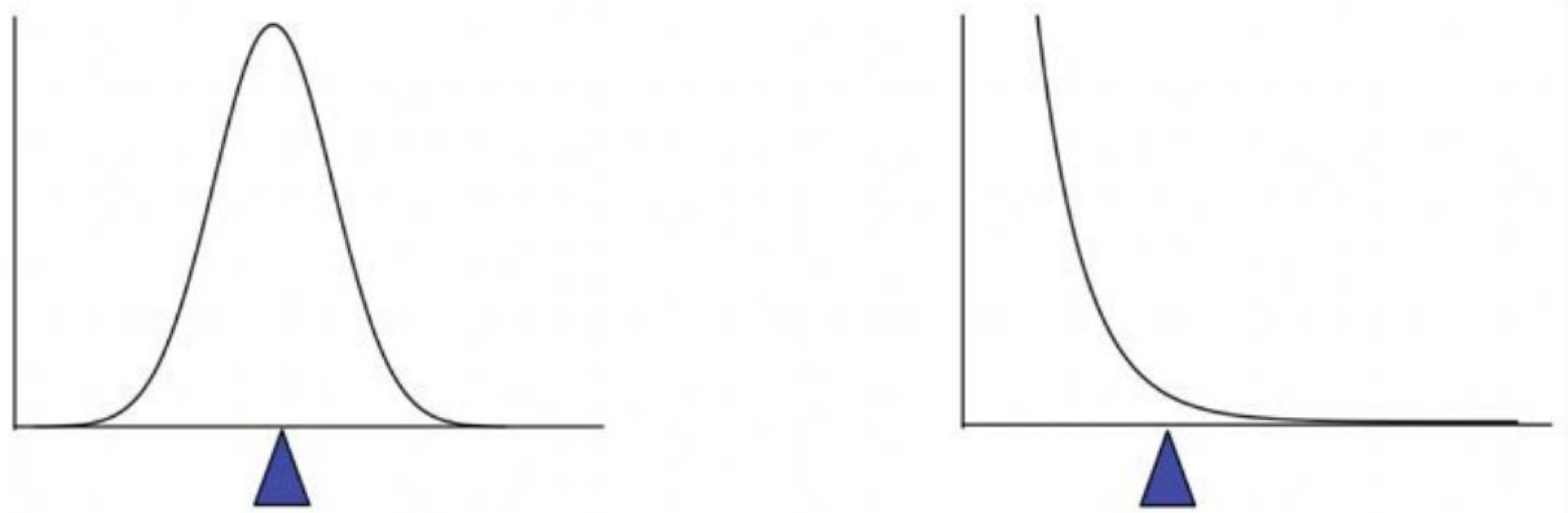
  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

Sample Mean

- The mean of a set of n observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.
- Key theme: there is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample Mean

Example. Suppose we have the following values for salary (in thousands of dollars), shown in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. The we have:

$$\bar{x} =$$

Sample Mean

Example. Suppose we have the following values for salary (in thousands of dollars), shown in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. The we have:

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000.

Sample Mean

The mean is sensitive to extreme (e.g., outlier) values. For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers.

Sample weighted arithmetic mean

We can assign each value x_i in a set X a corresponding weight w_i for $i = 1, \ldots, n$, where the weights reflect the significance, importance, or occurrence frequency associated with their respective values. The weighted average of this set of values is then given by:

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

Sample trimmed mean

Trimmed mean: is the mean obtained after removing the highest and lowest values.

For example, we can sort the values observed for salary and remove the top and bottom 2% before computing the mean.

> **Tip**
>
> We should avoid trimming too large a portion (such as 20%) at both ends, as this can result in the loss of valuable information.

Sample Median

- The median of a set of n number of observations in a sample, **ordered by value**, of a variable is is defined by

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

- Example (already in order):
  - Ages: 17, 19, 21, 22, 23, 23, 23, 38
  - Median = (22+23)/2 = 22.5
- The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

Sample Median

Example. Let's find the median of the data from previous salary example, The data are already sorted in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Then the median should be:

Sample Median

Example. Let's find the median of the data from previous salary example, The data are already sorted in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Then the median should be: **$54,000**

$$\frac{52+56}{2} = \frac{108}{2} = 54$$

Sample Median

We have a large number of observations. It is too expensive to compute the median.

Sample Median

We can easily approximate the median.

Assume that data are grouped in intervals according to their x_i data values and that the frequency (i.e., number of data values) of each interval is known. For example, employees may be grouped according to their annual salary in intervals such as \$10,001–20,000, \$20,001–50,000, and so on. We can approximate the median of the entire data set (e.g., the median salary) by interpolation using:

$$median \approx L_1 + \left(\frac{N/2 - \left(\sum freq\right)_l}{freq_{median}} \right) \times width,$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $\left(\sum freq\right)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Sample Median

$$median \approx L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) \times width,$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Exercise: Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as:

age	frequency
1-12	300
12-20	450
21-50	1500
51-80	700
81-110	44

Compute an approximate median value for the data.

Sample Median

$$median \approx L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) \times width,$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Exercise: Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as:

- median interval: ?, L1 = ?, N = ?
- $(\sum freq)_l$ = ?, $freq_{median}$ = ?, width = ?
- **median ≈ ?**

age	frequency
1-12	300
12-20	450
21-50	1500
51-80	700
81-110	50

Sample Median

$$median \approx L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) \times width,$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Exercise: Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as:

- median interval: **21 - 50**, L1 = **21**, N = **3000**
- $(\sum freq)_l$ = **750**, $freq_{median}$ = **1500**, width = **30**
- **median ≈ 36**

age	frequency
1-12	300
12-20	450
21-50	1500
51-80	700
81-110	50

Mean vs. Median

- The mean is sensitive to extreme values (outliers)

Mode

Mode: The mode for a set of data is the value that occurs most frequently compared to all neighboring values in the set. It is possible to have more than one mode.

- unimodal, bimodal, trimodal, multimodal…

Example. The data from 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. are ______. The modes are:

Mode

Mode: The mode for a set of data is the value that occurs most frequently compared to all neighboring values in the set. It is possible to have more than one mode.

- unimodal, bimodal, trimodal, multimodal…

Example. The data from 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. are **bimodal**. The modes are: **$52,000 and $70,000**.

Mode

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical formula:

$$mean - mode \approx 3 \times (mean - median).$$

Midrange

Midrange: is the average of the largest and smallest values in the set.

Example. The midrange of the data of 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 is

Midrange

Midrange: is the average of the largest and smallest values in the set.

Example. The midrange of the data of 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 is

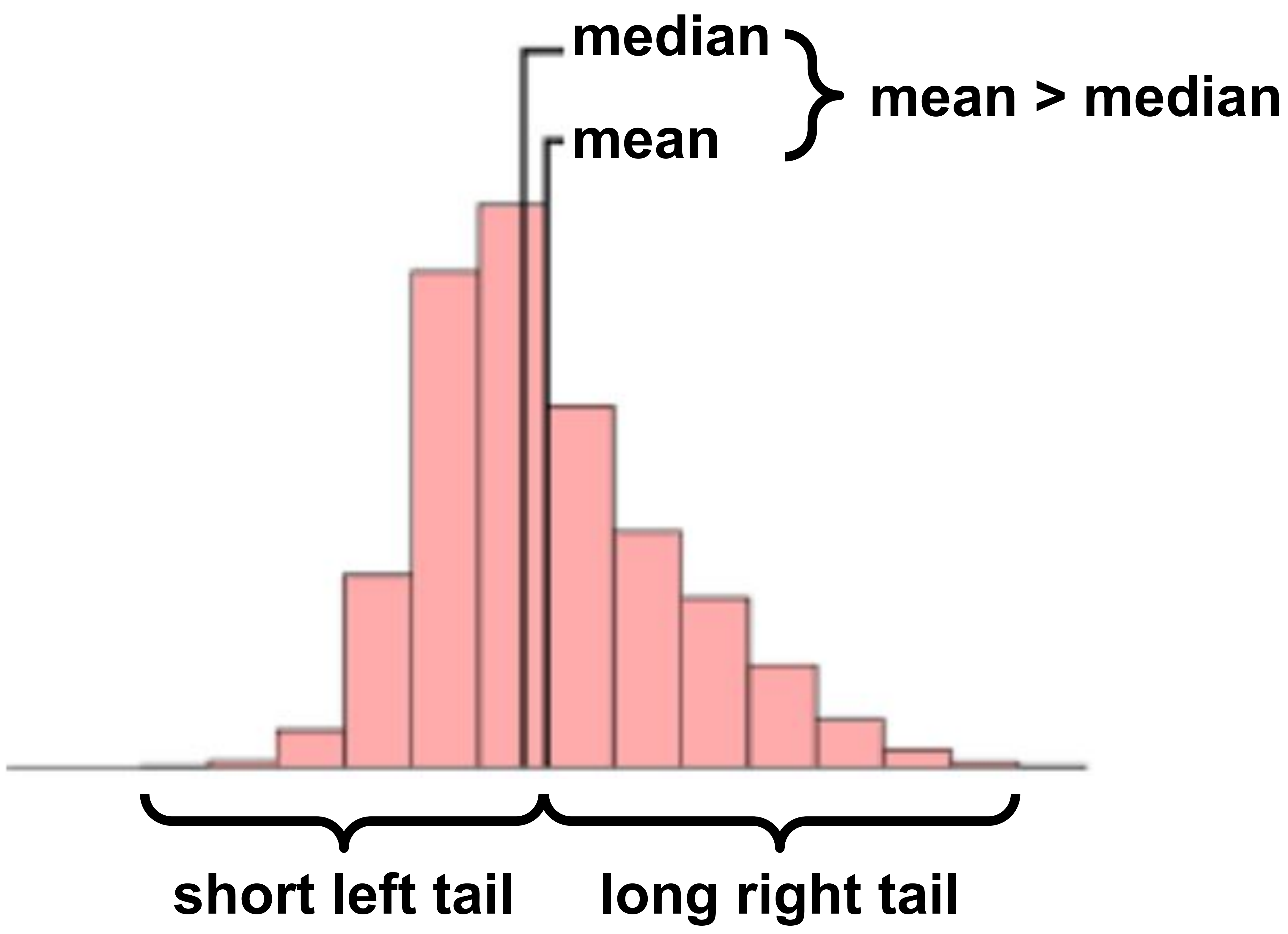
$$(30 + 110) / 2 = 70$$

Outline

- Central Tendency
  - Mean, Median, Mode, Midrange
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation

Mean, median, and skewness

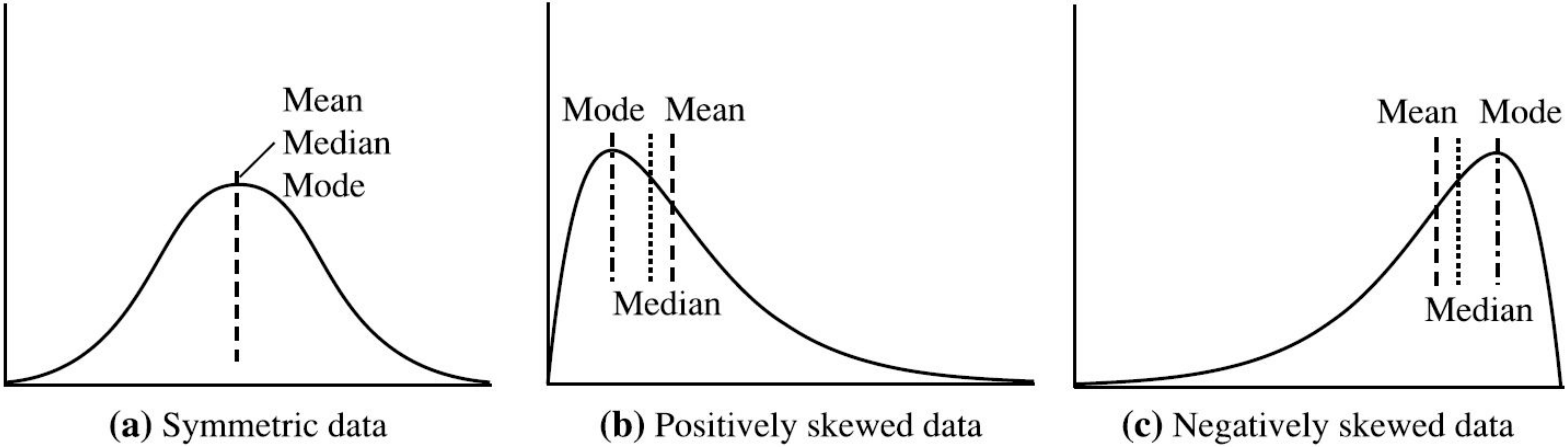
- The following distribution is called right-skewed since the mean is greater than the median. Note: skewness often “follows the longer tail”



Symmetric vs. Skewed Data

- In a unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value.
- Data in most real applications are not symmetric. positively skewed, or negatively skewed.

Median, mean and mode of symmetric, positively and negatively skewed data

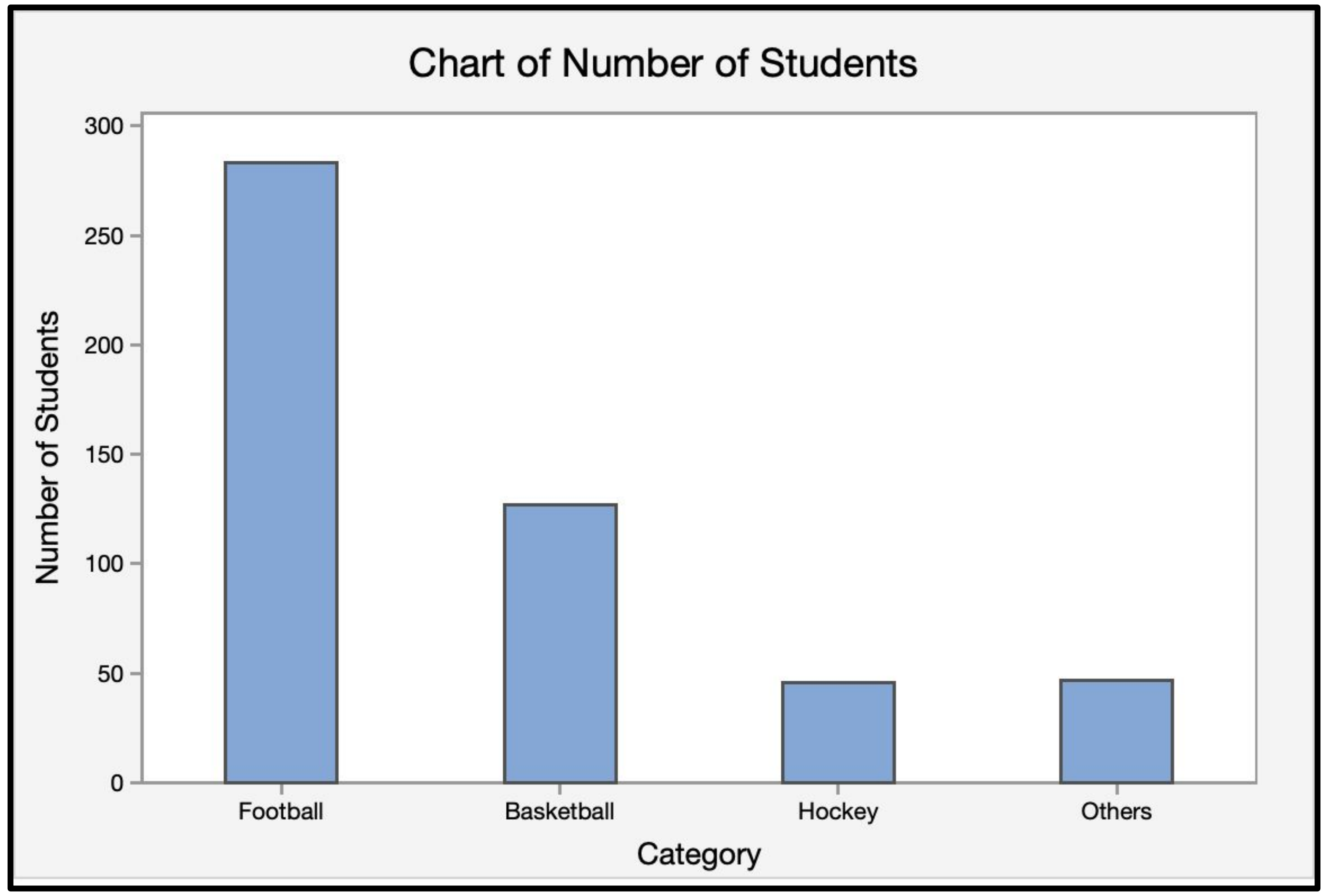


Question

Is income positively or negatively skewed?

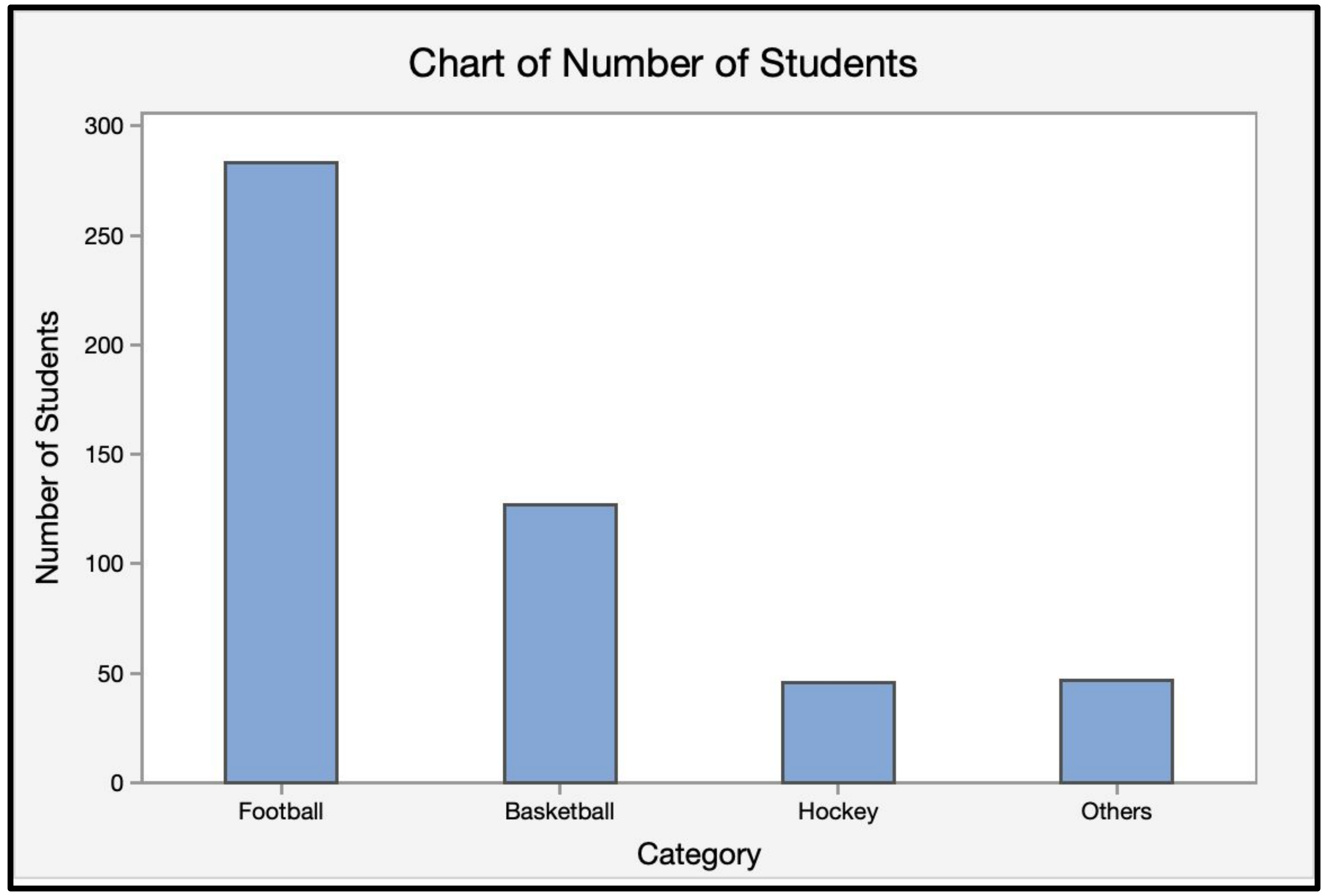
Regarding Categorical Variables...

- **For categorical variables, neither mean or median make sense. Why?**



Regarding Categorical Variables...

- **For categorical variables, neither mean or median make sense. Why?**



- The **mode** might be a better way to find the most "representative" value.

Outline

- Central Tendency
  - Mean, Median, Mode,
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation

Measures of Spread: Range

- The spread of a sample of observations measures how well the mean or median describes the sample.
- One way to measure spread of a sample of observations is via the **range**.

Range = Maximum Value - Minimum Value

Measures of Spread: Quantiles

Suppose that the data for attribute **X** are sorted in ascending numeric order. Imagine that we can pick certain data points so as to split the data distribution into contiguous sets of equal size, These data points are called **quantiles**.

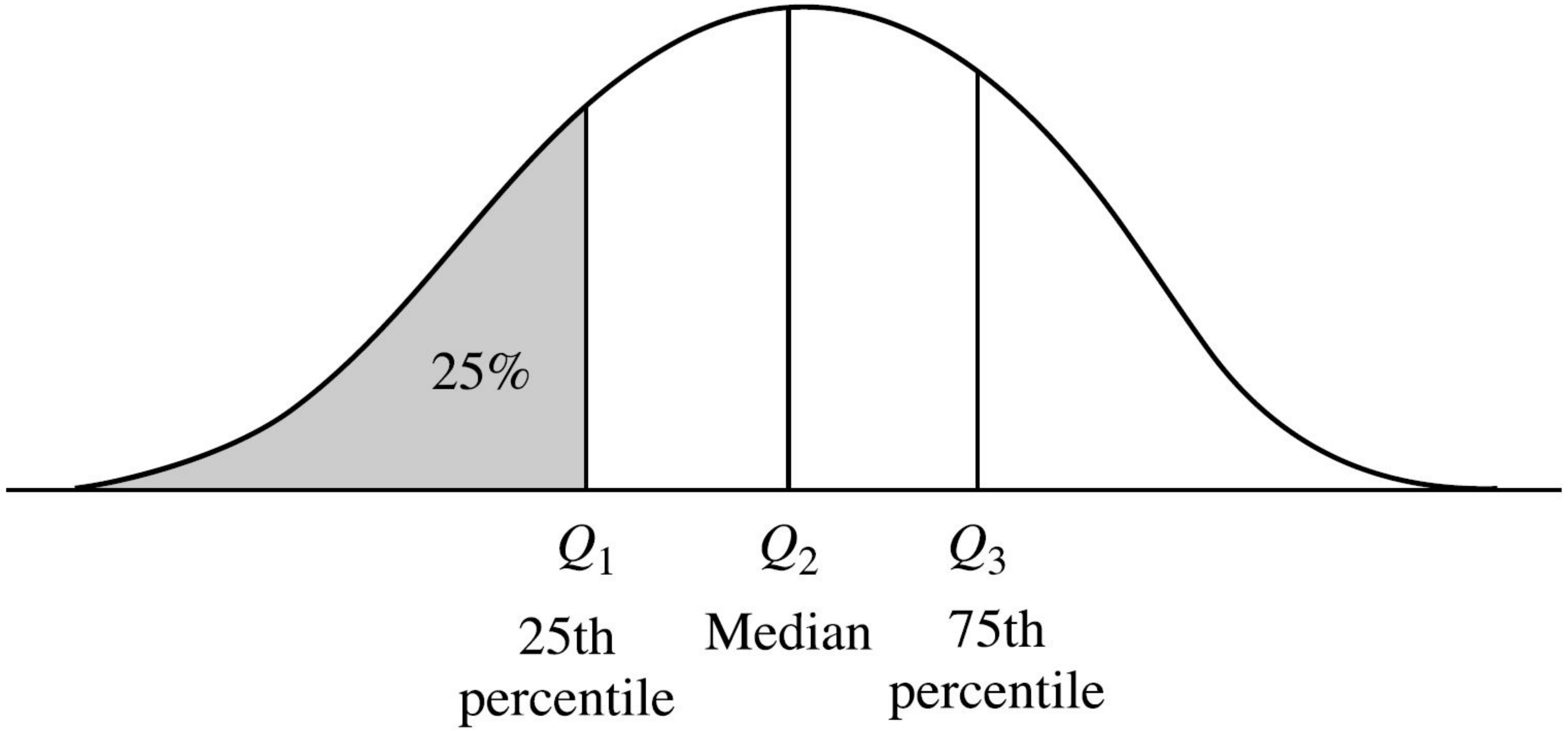
The 2-quantile is the data point dividing the lower and upper halves of the data distribution.

What is the 2-quantile also called?

median

Measures of Spread: Quartiles

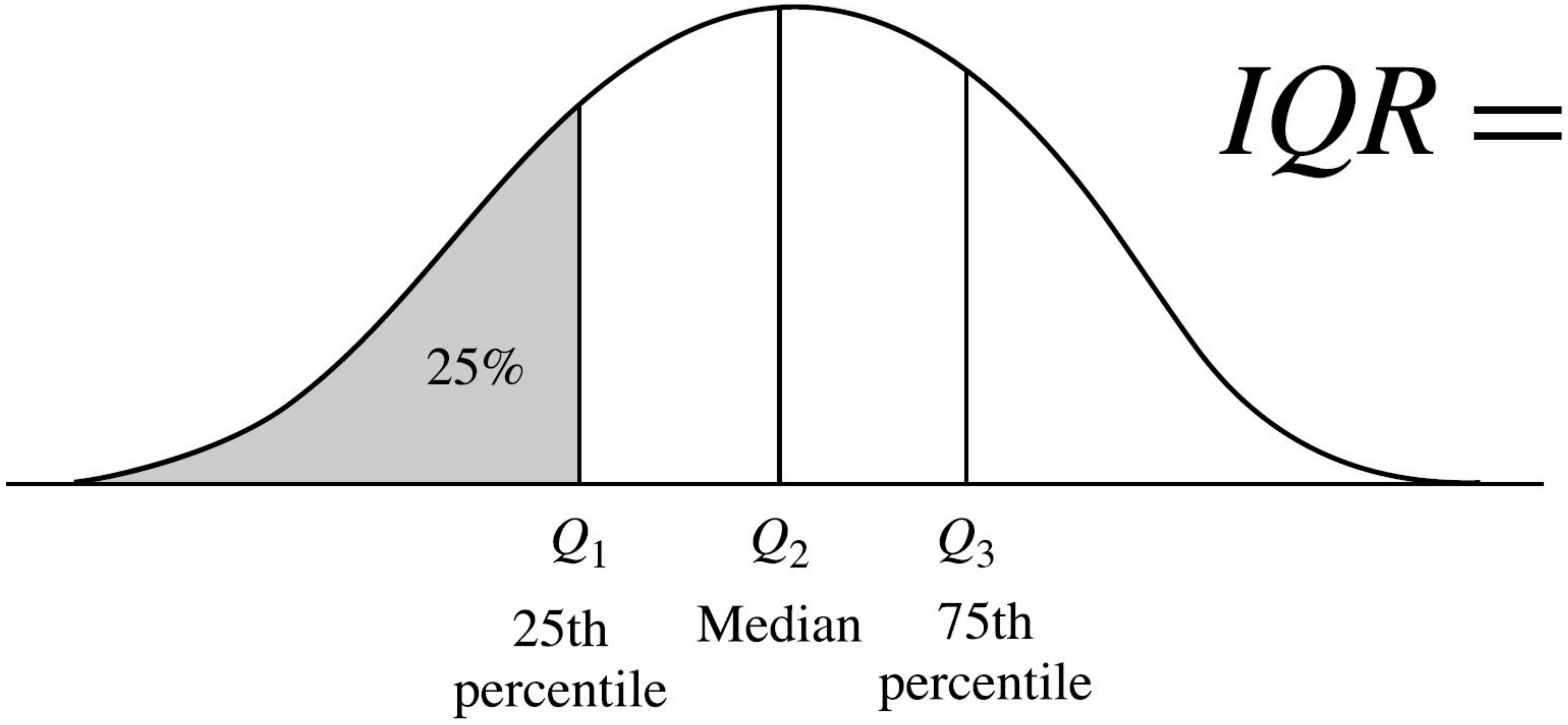
The **4-quantiles** are the **three data points** that split the data distribution into **four equal parts**; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**



The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size contiguous subsets. The second quartile corresponds to the median.

Measures of Spread: IQR

The distance between the **first** and **third** quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range (IQR)**.



$$IQR = Q_3 - Q_1$$

Measures of Spread: IQR

Example. Let's calculate the Q1, Q2, Q3 and IQR of the salary data, The data are already sorted in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Then:

- Q1 = **($47,000 + $50,000)/2 = $48,500**
- Q2 = **($52,000 + $56,000)/2 = $54,000**
- Q3 = **($63,000 + $70,000)/2 = $66,500**
- IQR = **$66,500 − $48,500 = $18,000**

$$IQR = Q_3 - Q_1$$

Outline

- Central Tendency
  - Mean, Median, Mode, Midrange
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation

Measures of Spread: Variance

- The (sample) variance, measures how much on average the sample values deviate from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|^2$$

- Note: the term $|x_i - \bar{x}|$ measures the amount by which each x_i deviates from the mean $\bar{x}$. Squaring these deviations means that **variance is sensitive to extreme values (outliers)**.
- What does a variance of 1,008 mean? Or 0.0001?

Measures of Spread: Standard Deviation

- The (sample) standard deviation, denoted s, is the square root of the variance

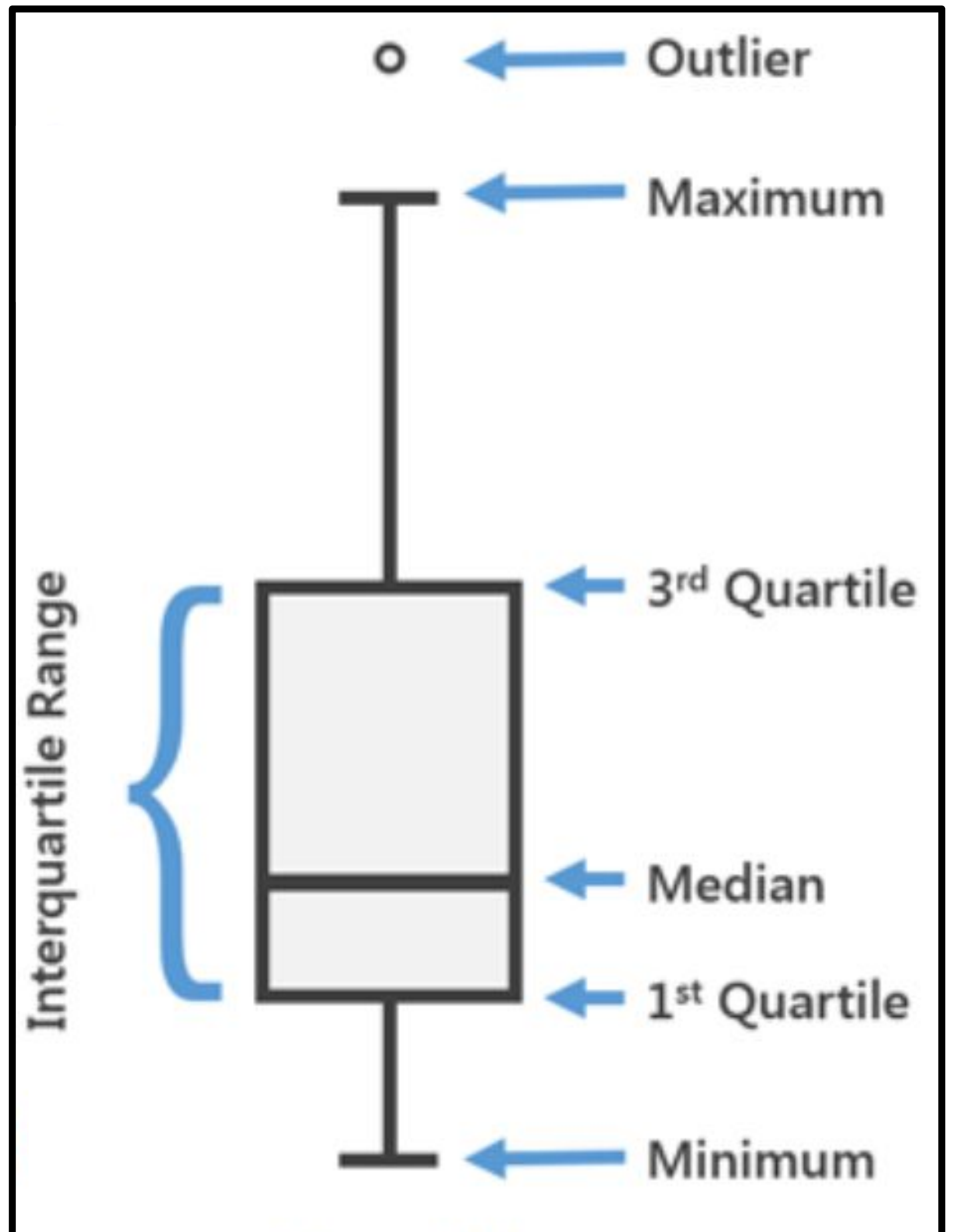
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|^2}$$

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: IQR = Q3 –Q1
  - Five number summary: **min, Q1, median,Q3, max**
  - **Outlier: usually, a value higher/lower than 1.5 x IQR of Q3 or Q1**
- Variance and standard deviation
  - Standard deviation s (or σ)

    is the square root of variance

Population Variance	**Sample Variance**
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
$\sigma^2 =$ population variance $x_i =$ value of i^{th} element $\mu =$ population mean $N =$ population size	$s^2 =$ sample variance $x_i =$ value of i^{th} element $\bar{x} =$ sample mean $n =$ sample size



Box Plot

Summary

- Central Tendency
  - Mean, Median, Mode, Midrange
  - Symmetric vs. Skewed Data
- Dispersion of Data
  - Range, Quantiles, Quartiles, Interquartile Range(IQR),
  - Variance, Standard Deviation