SOC2069 Researching Social Life 1

Quantitative data and descriptive statistics

Dr. Chris Moreh

Outline

1. Variables

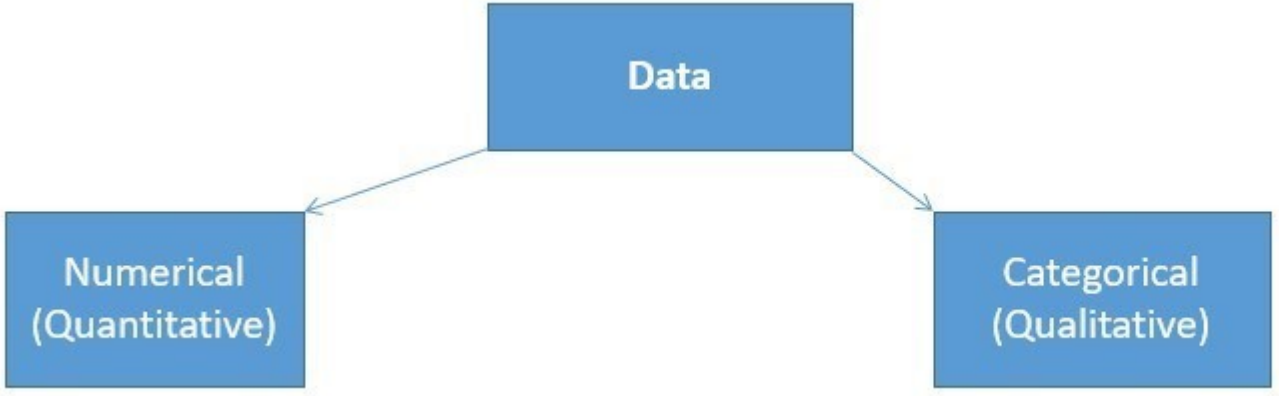
2. Descriptive statistics

Variables

What is a *variable*?

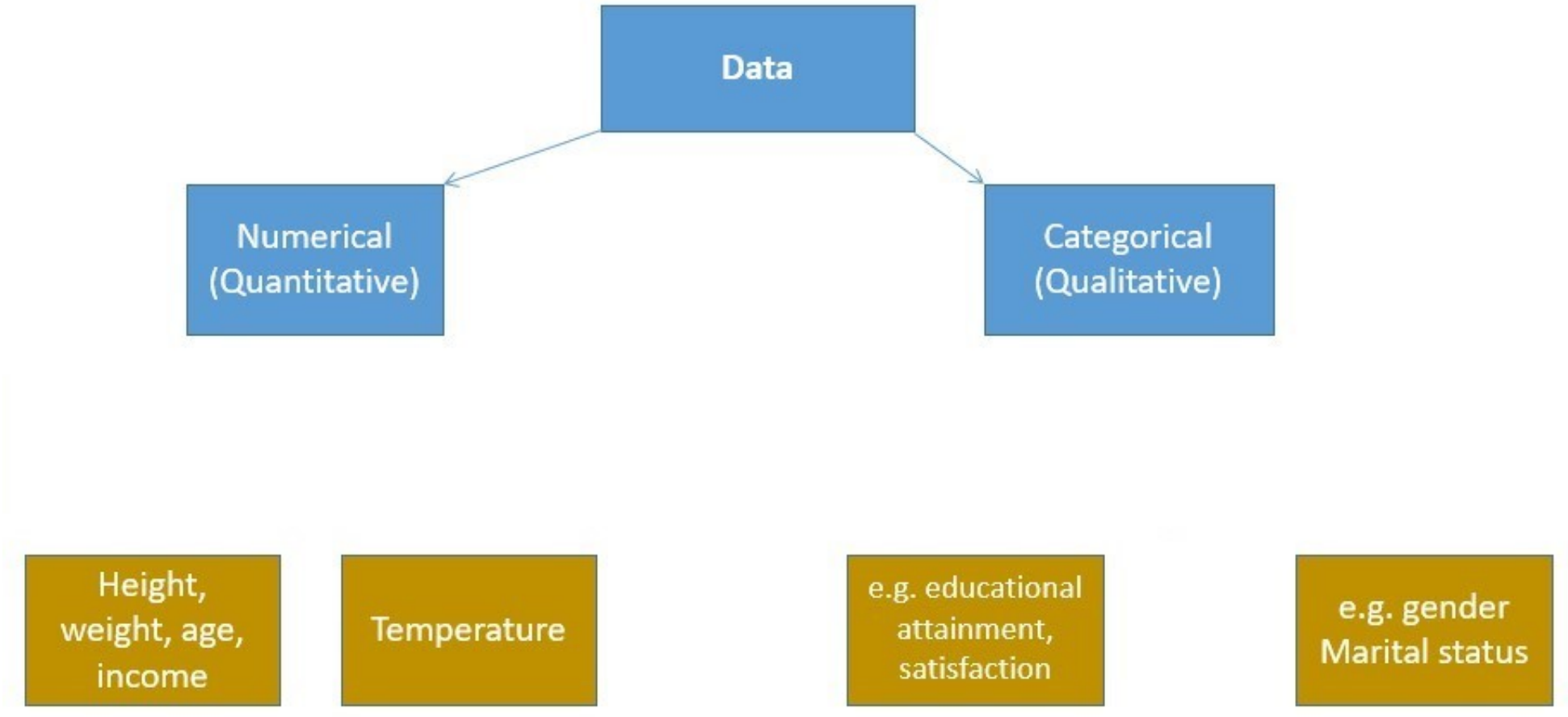
- Statistical methods help us determine the factors that explain **variability** among subjects/respondents
- For instance, variation occurs from student to student in their grades. What factors are responsible for that variability?
- Any characteristic that we can measure for each subject is called a **variable**
- Variable are characteristics that can *vary* in value among subjects in a *sample* or *population*
- Examples of variables are income last year, number of children or siblings, whether employed, gender, how much one likes ice-cream on a scale of 1 to 10, etc.
- The values the variable can take form the **measurement scale**
- For gender, for instance, the measurement scale consists of the two (or more) labels, (female, male, other). For number of children/siblings, it would be (0, 1, 2, 3, 4, …)

Measurement scales

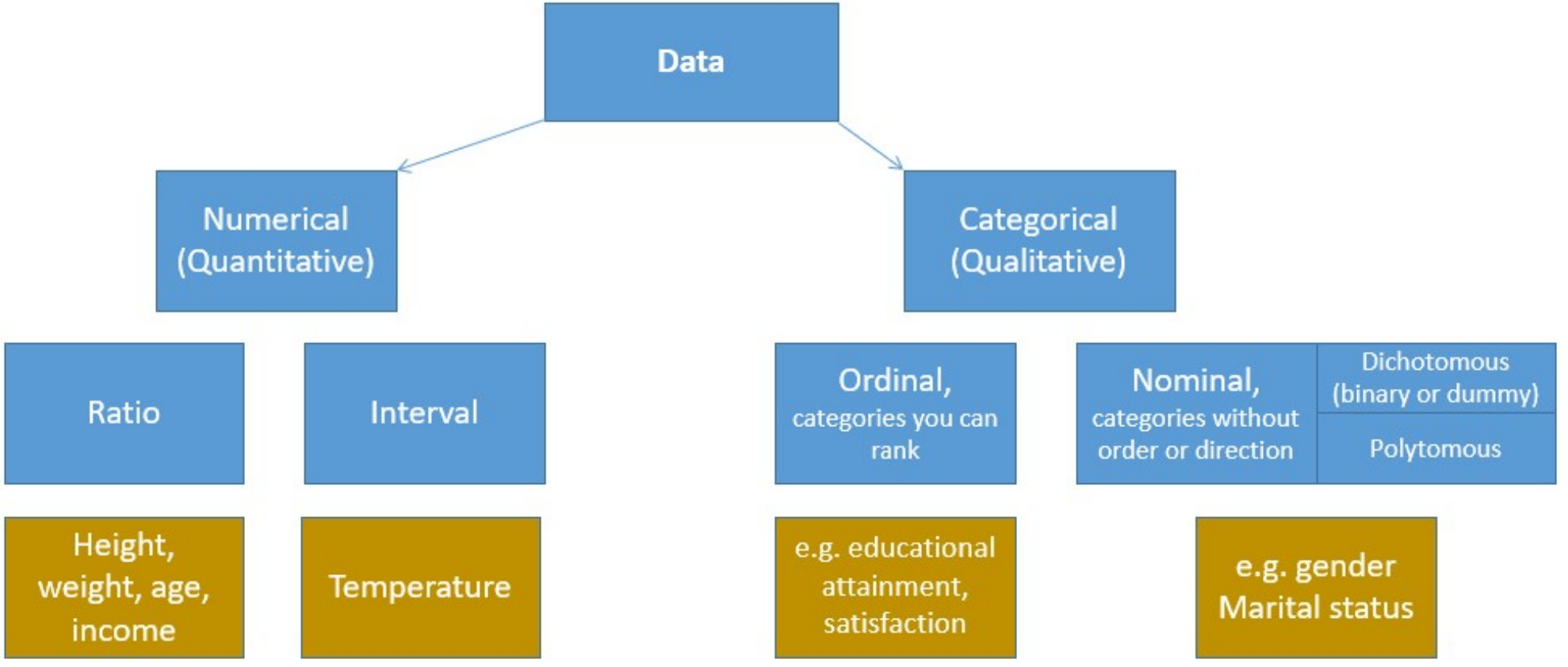


- A variable is called **quantitative** when the measurement scale has **numerical** values that represent different magnitudes of the variable
- A variable is called **categorical** when the measurement scale is a set of categories
- For categorical variables, distinct categories differ in *quality*, not in numerical magnitude. For this reason, categorical variables are often called **qualitative** (but we won't call them as such, to avoid confusion with the type of qualitative data we covered in the first half of the module)

Measurement scales



Measurement scales



The position of ordinal scales on the quantitative–qualitative classification is fuzzy. Because their scale is a set of categories, they are often analyzed using the same methods as nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: *each level has a greater or smaller magnitude than another level*

Measurement scales

Values can also be:

Discrete

Continuous

Even when the ***variables*** that contain them are continuous:

e.g. think of: number of cars on the road

e.g. height, weight, age income

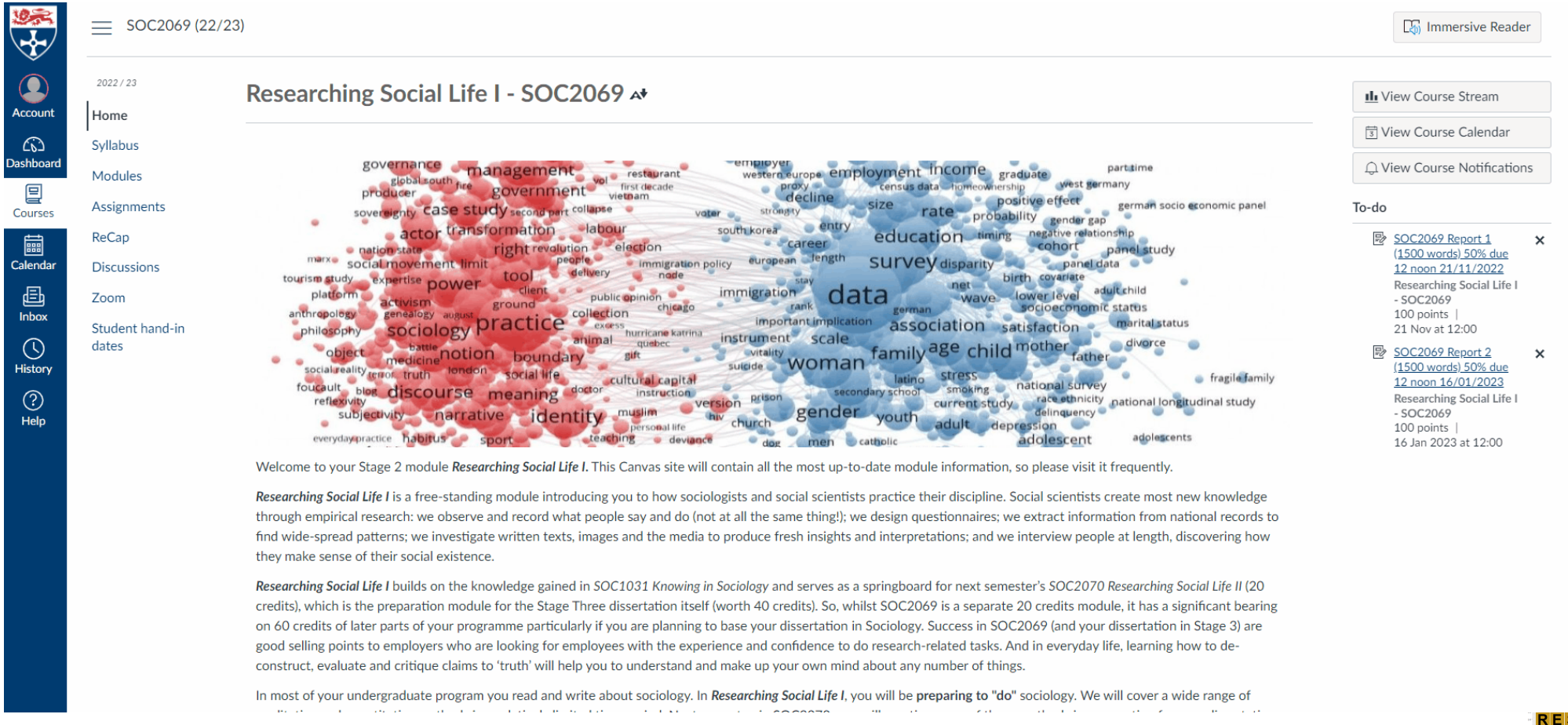
A variable's values are **discrete** if its possible values form a set of separate numbers, such as (0, 1, 2, 3, . . .).
They are **continuous** if it can take an infinite continuum of possible real number values.

Measurement scales

Scale	Values	Examples
Nominal	Order values: No	Yes/no questions
	Same distance: No	Gender
	Absolute zero point: Not applicable	Ethnicity
Ordinal	Order values: Yes	Attitude questions
	Same distance: No	Self-rated health
	Absolute zero point: Not applicable	Educational level
Ratio	Order values: Yes	Age
	Same distance: Yes	Income
	Absolute zero point: Yes	School marks
Interval	Order values: Yes	Temperature (Celsius)
	Same distance: Yes	
	Absolute zero point: No	

Where do *variables* come from?

- Data collection: Observation, interviewing, experiments…
- The data we use in this module comes from **Wave 8** of the **UK Household Longitudinal Study (Understanding Society) Main Survey**:



Descriptive statistics

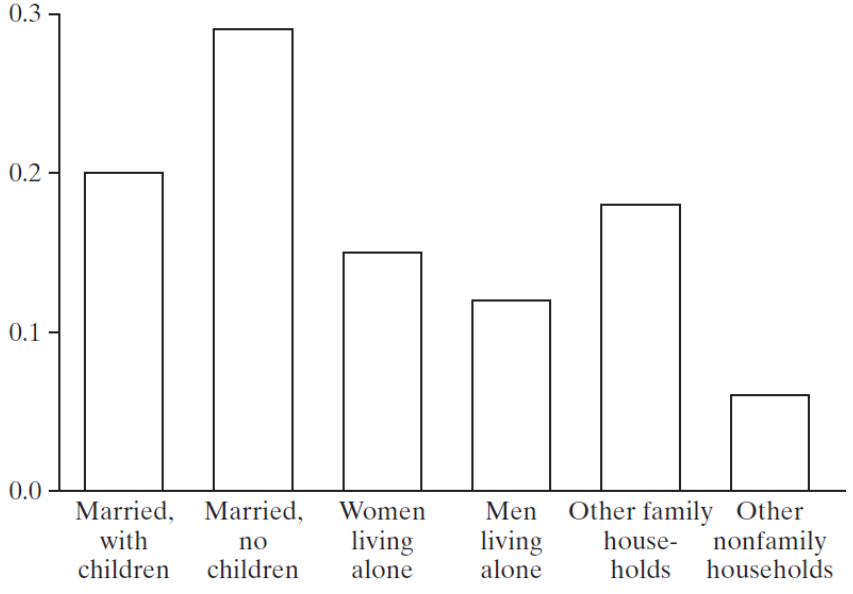
Describing categorical variables

- Categorical data are characterized by a **frequency distribution**
- A frequency table is a listing of possible values for a variable, together with the **number of observations** (n) at each value
- When the table shows the **proportions** or **percentages** instead of the numbers, it is called a - **relative- frequency distribution**
- Frequency distributions can also be visualised with a **bar graph**

Type of Family	Number (millions)	Proportion	Percentage (1970)
Married couple with children	23.3	0.20	20 (40)
Married couple, no children	33.7	0.29	29 (30)
Women living alone	17.4	0.15	15 (11)
Men living alone	14.0	0.12	12 (6)
Other family households	20.9	0.18	18 (11)
Other nonfamily households	7.0	0.06	6 (2)
Total	116.3	1.00	100 (100)

Describing *categorical* variables

- Categorical data are characterized by a **frequency distribution**
- A frequency table is a listing of possible values for a variable, together with the **number of observations** (n) at each value
- When the table shows the **proportions** or **percentages** instead of the numbers, it is called a - **relative- frequency distribution**
- Frequency distributions can also be visualised with a **bar graph**



Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Central tendency

Measure	Definition
Mean	The average value
Median	The value in the absolute middle
Mode	The most frequently occurring value

Example 1: Mean

1.1 kilos 0.8 kilos 1.1 kilos 1.0 kilos

Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Central tendency

Measure	Definition
Mean	The average value
Median	The value in the absolute middle
Mode	The most frequently occurring value

Example 1: Mean

1.1 kilos + 0.8 kilos + 1.1 kilos + 1.0 kilos = 4

4/(1+1+1+1) = 1 -> Mean = 1 kilo

1. Add values together
2. Divide sum by number of values

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

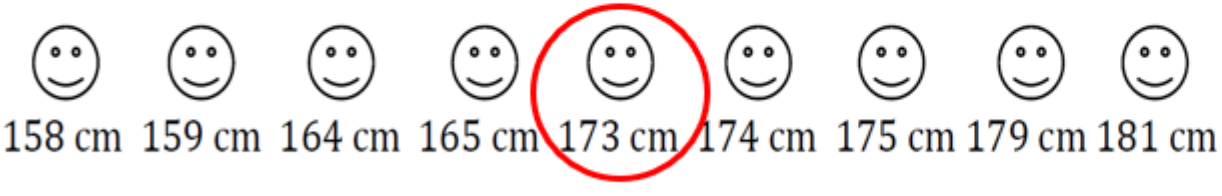
Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Central tendency

Measure	Definition
Mean	The average value
Median	The value in the absolute middle
Mode	The most frequently occurring value

Example 2: Median



Sort values from low to high and then identifying the value in the middle

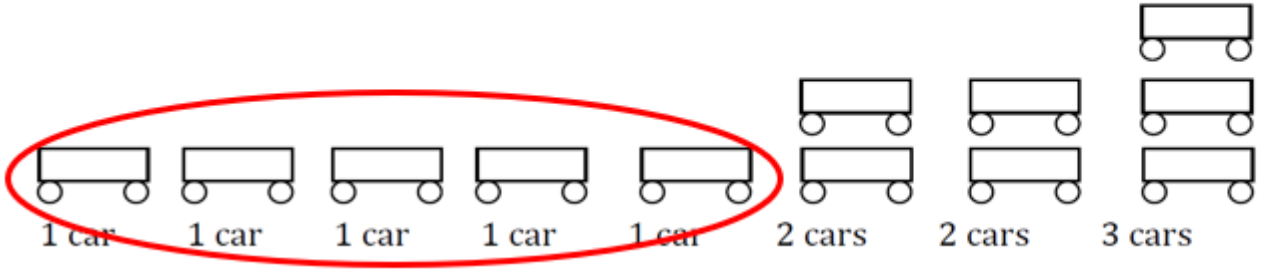
Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Central tendency

Measure	Definition
Mean	The average value
Median	The value in the absolute middle
Mode	The most frequently occurring value

Example 3: Mode



It's the most frequently occurring value in a distribution.

The **mode** also applies to *categorical* variables - it's more useful for describing the category with the highest frequency

Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Variation (spread)

Measure	Definition
Min	The lowest value
Max	The highest value
Range	The difference between the lowest and highest value
Standard deviation	The dispersion of values from the mean

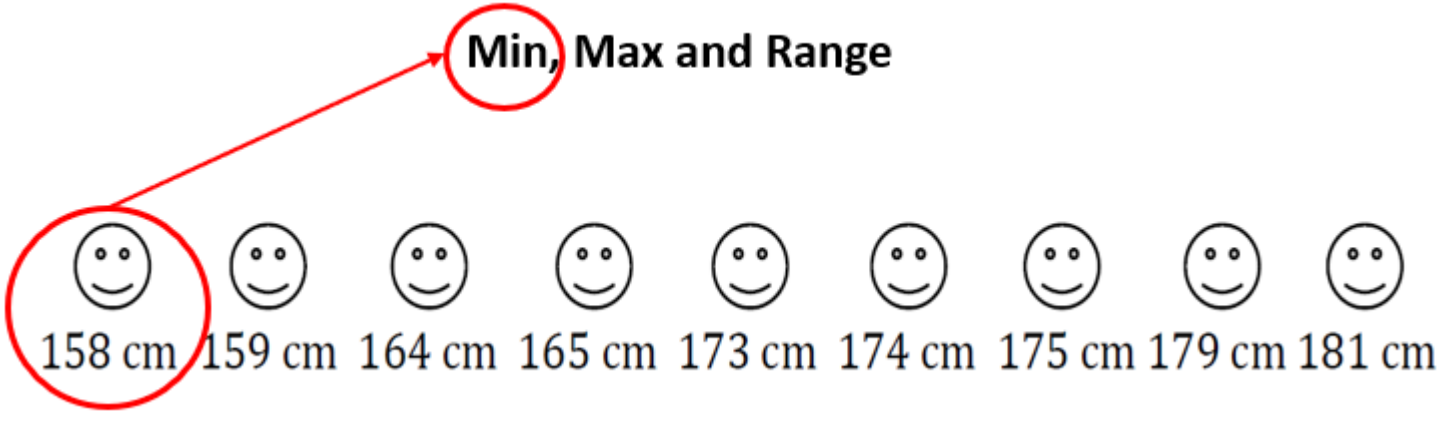
Min, Max and Range

Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Variation (spread)

Measure	Definition
Min	The lowest value
Max	The highest value
Range	The difference between the lowest and highest value
Standard deviation	The dispersion of values from the mean

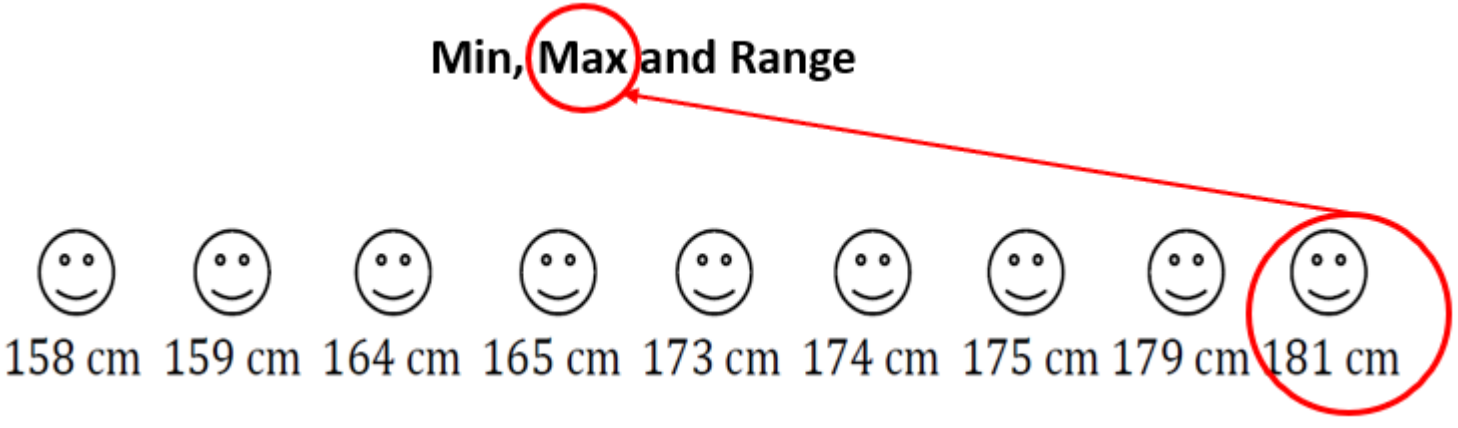


Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Variation (spread)

Measure	Definition
Min	The lowest value
Max	The highest value
Range	The difference between the lowest and highest value
Standard deviation	The dispersion of values from the mean

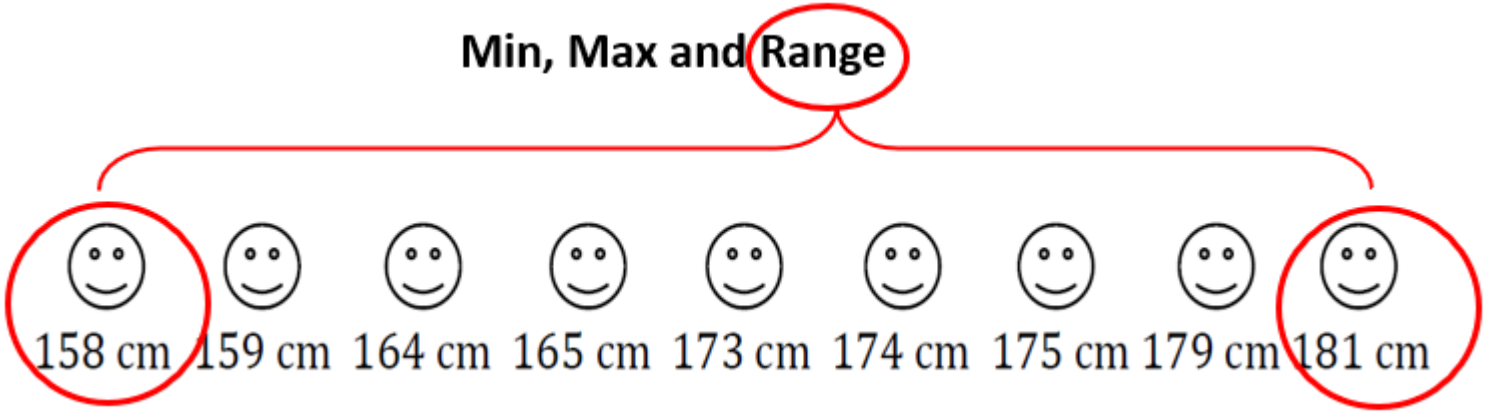


Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Variation (spread)

Measure	Definition
Min	The lowest value
Max	The highest value
Range	The difference between the lowest and highest value
Standard deviation	The dispersion of values from the mean



Describing *numeric* variables

Quantitative variables can be summarised by measures of **central tendency** and **variation** (spread)

Variation (spread)

Measure	Definition
Min	The lowest value
Max	The highest value
Range	The difference between the lowest and highest value
Standard deviation	The dispersion of values from the mean

The ***standard deviation*** s of n observations is

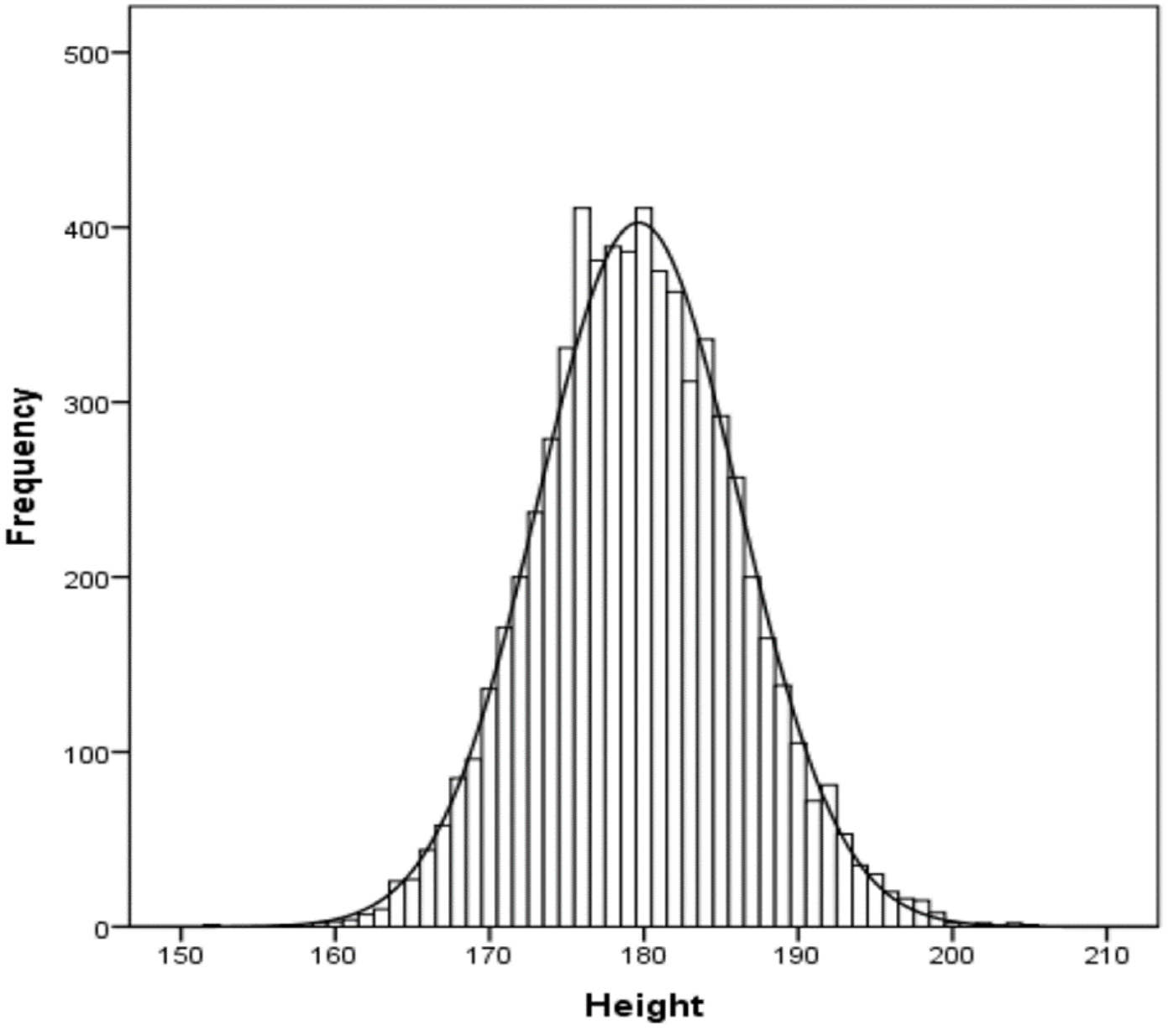
$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size } -1}}.$$

This is the positive square root of the ***variance*** s^2, which is

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}.$$

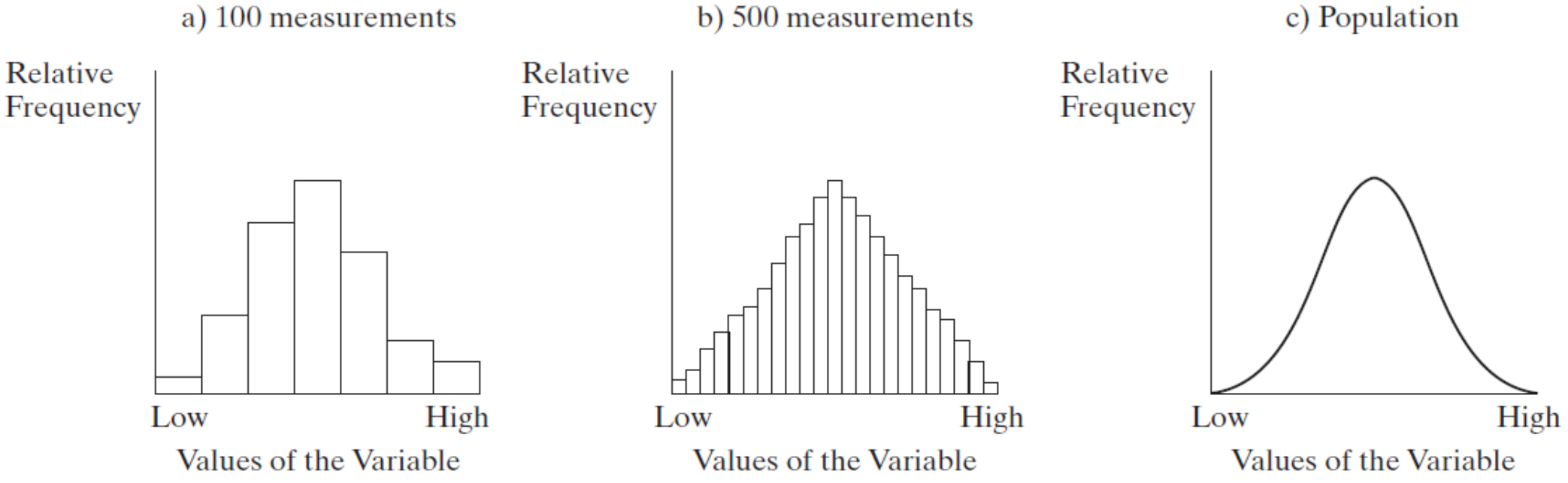
Describing *numeric* variables

Quantitative variables can be visualised with a **histogram** (a special *frequency distribution* with grouped numeric values)

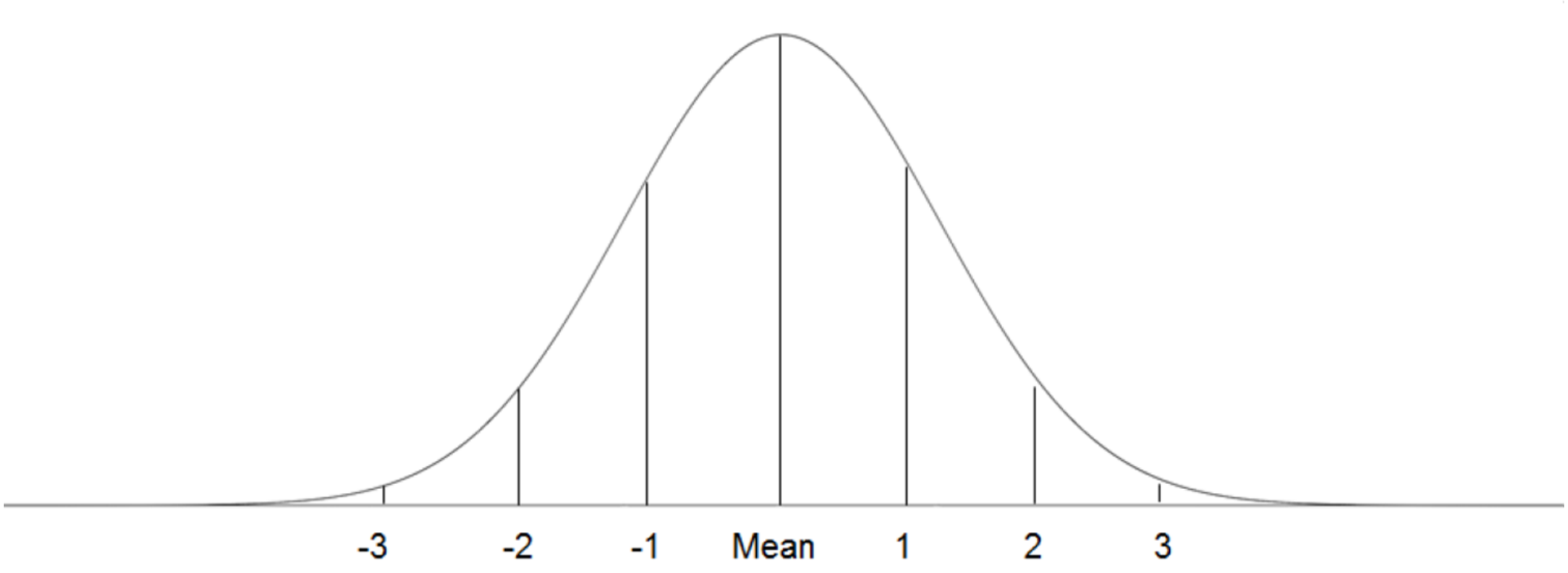

500
400
300
200
100
0
Frequency
150
160
170
180
190
200
210
Height

Describing *numeric* variables

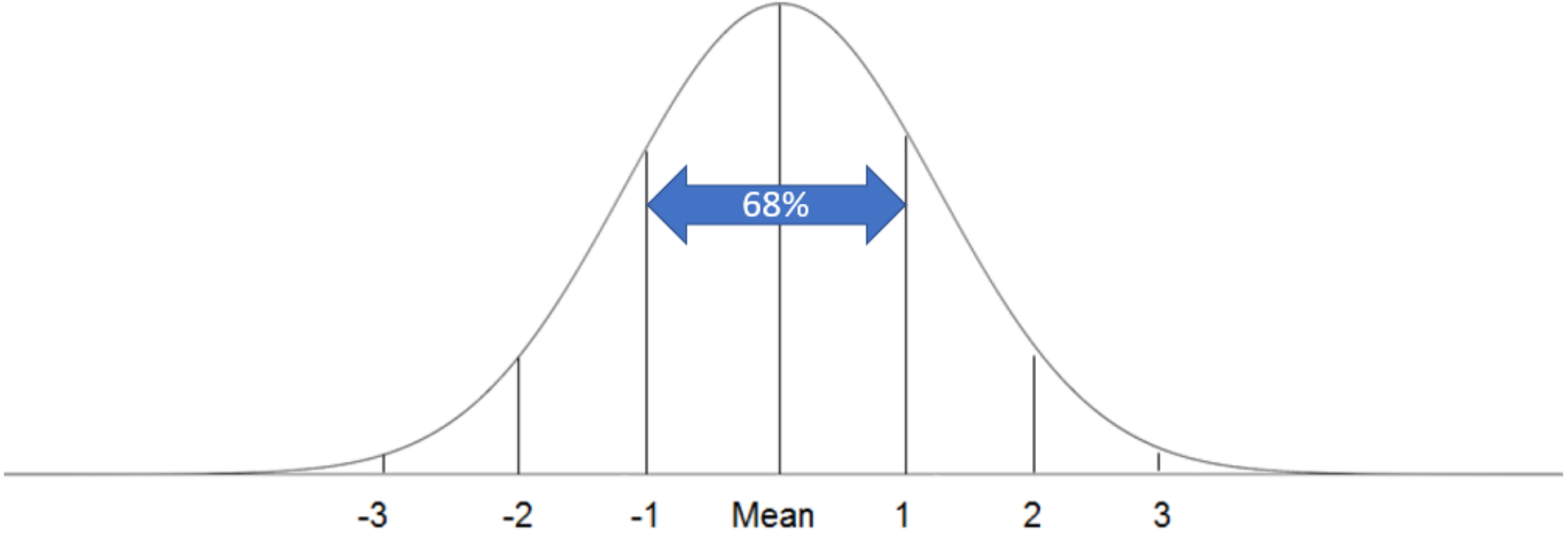
Quantitative variables can be visualised with a **histogram** (a special *frequency distribution* with grouped numeric values)



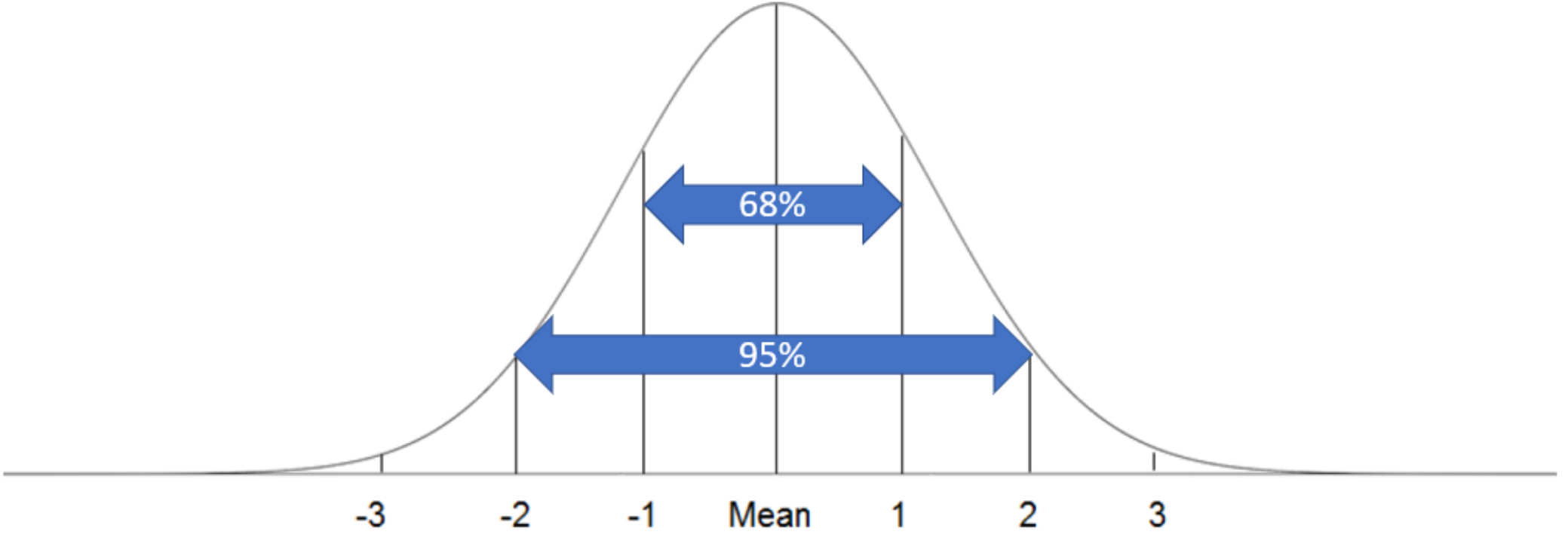
The *normal* distribution



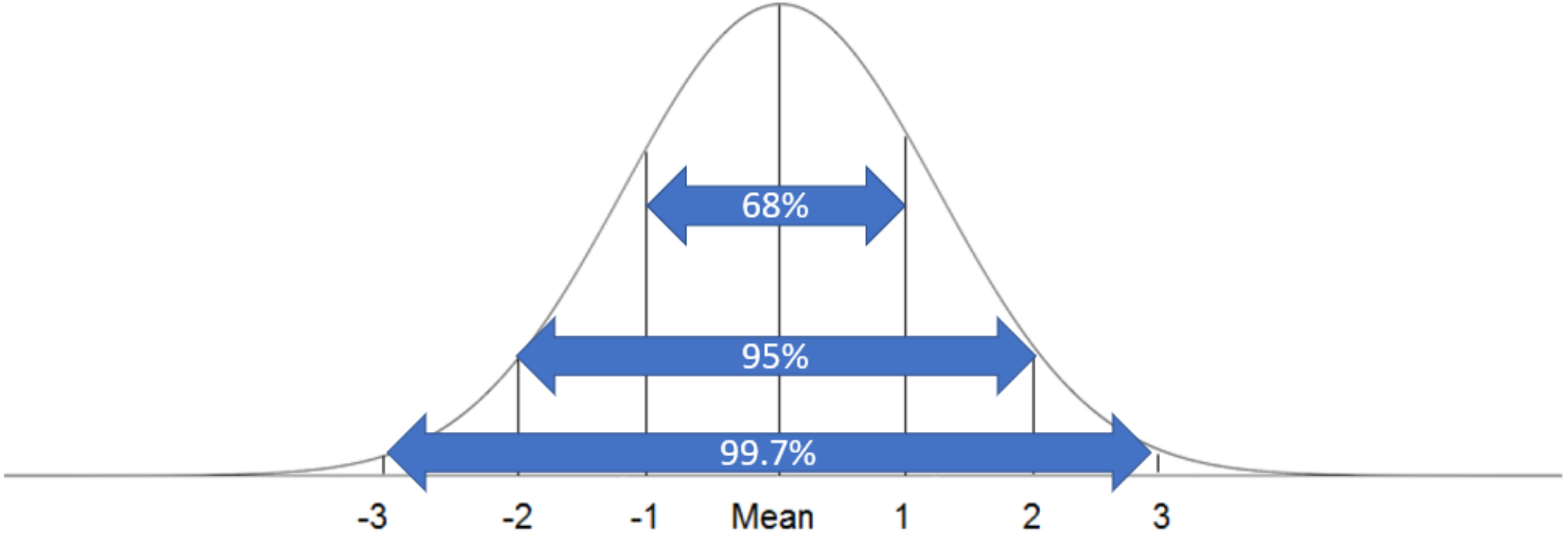
The *normal* distribution



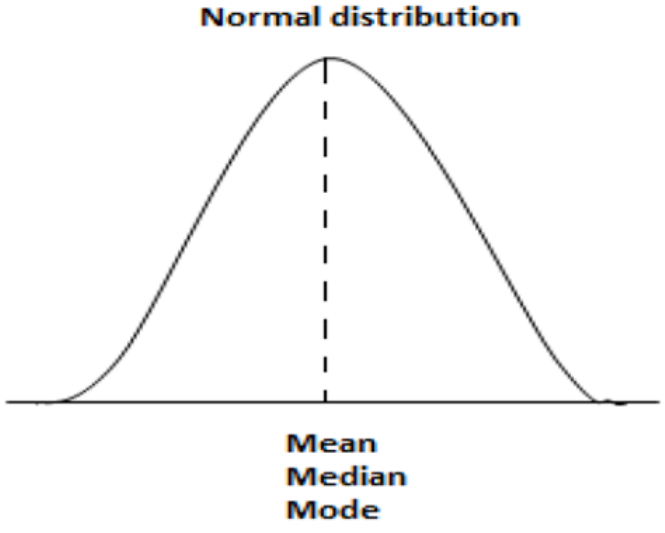
The *normal* distribution

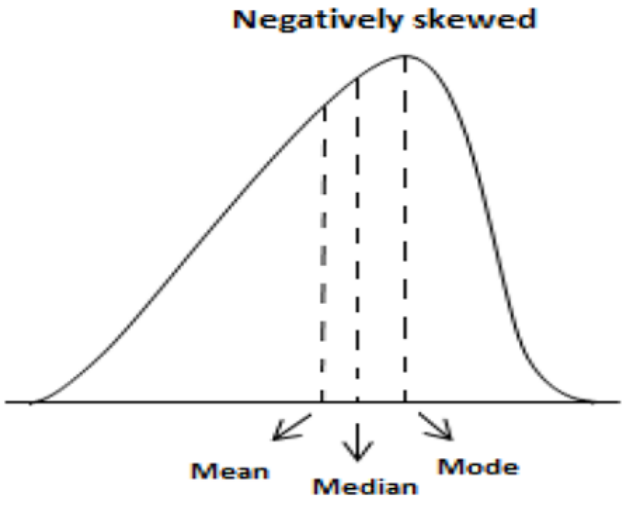


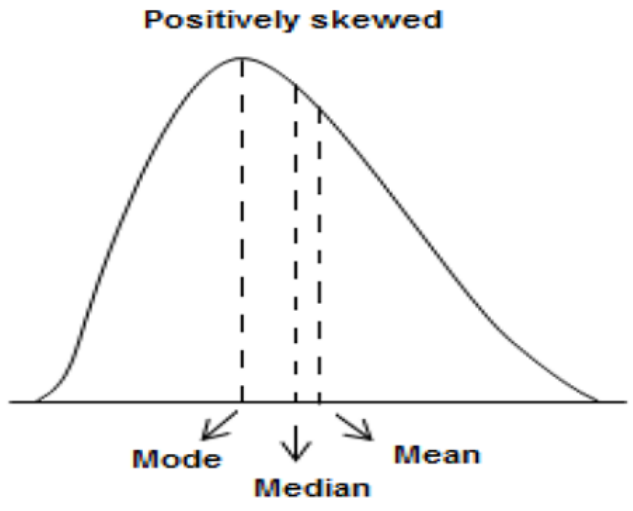
The *normal* distribution



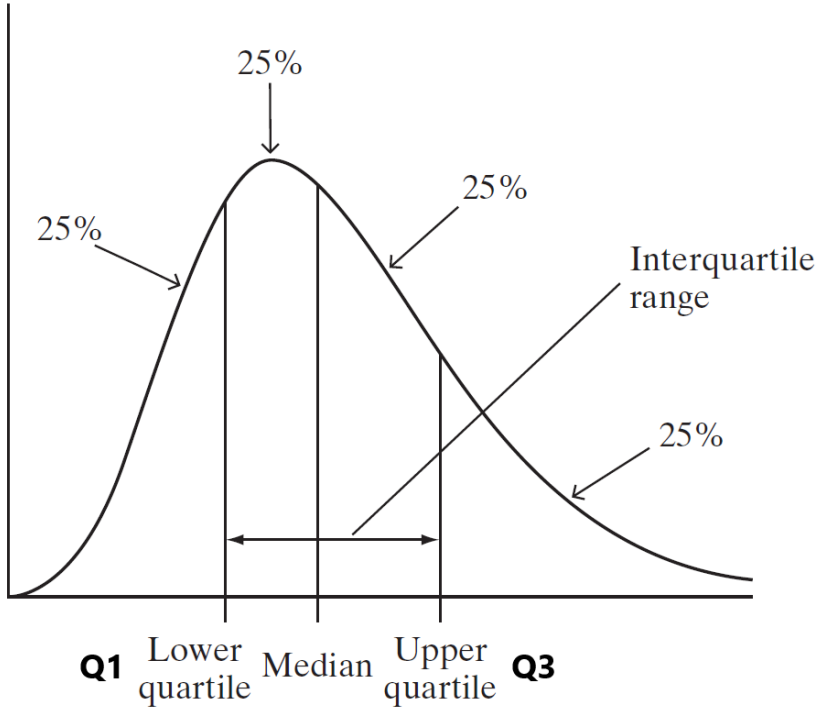
Skewed distribution







Quartiles and outliers



Boxplot: