universität wien

BAK3: Introduction to Quantitative Methods

Week 4: Univariate Descriptive Statistics I

Leonardo Carella

universität wien

The Plan for today

- Statistics:
  - Descriptive Statistics: Measures of Central Tendency (mode, mean and median).
- Coding in R:
  - Data Visualisation with ‘ggplot’: bar-plots and histograms.

universität wien

Week 1: Statistics

- **Description**: summarising large quantities of information (the data) into smaller, more manageable pieces of information (an average, a graph, a table, a correlation etc.)
- **Inference**: making predictions, using what we know about the data to *infer* what's likely to be 'true' more generally. Involves dealing with uncertainty.

universität wien

Week 3: Measurement

	Scale Level	Meaningful order?	Meaningful distance?	Meaningful zero point?	
Categorical	Nominal	-	-	-	Discrete
	Ordinal	✓	-	-	
Numerical	Interval	✓	✓	-	Discrete or Continuous
	Ratio	✓	✓	✓	

universität wien

Describing Variables

- **Descriptive Statistics**: values that tell us something important about a variable.
  - Measures of central tendency: ‘typical’ value of a variable.
- **Data Visualisation:** graphical representation of the variable’s distribution (in today’s case) via size, shape, length, color etc.

universität wien

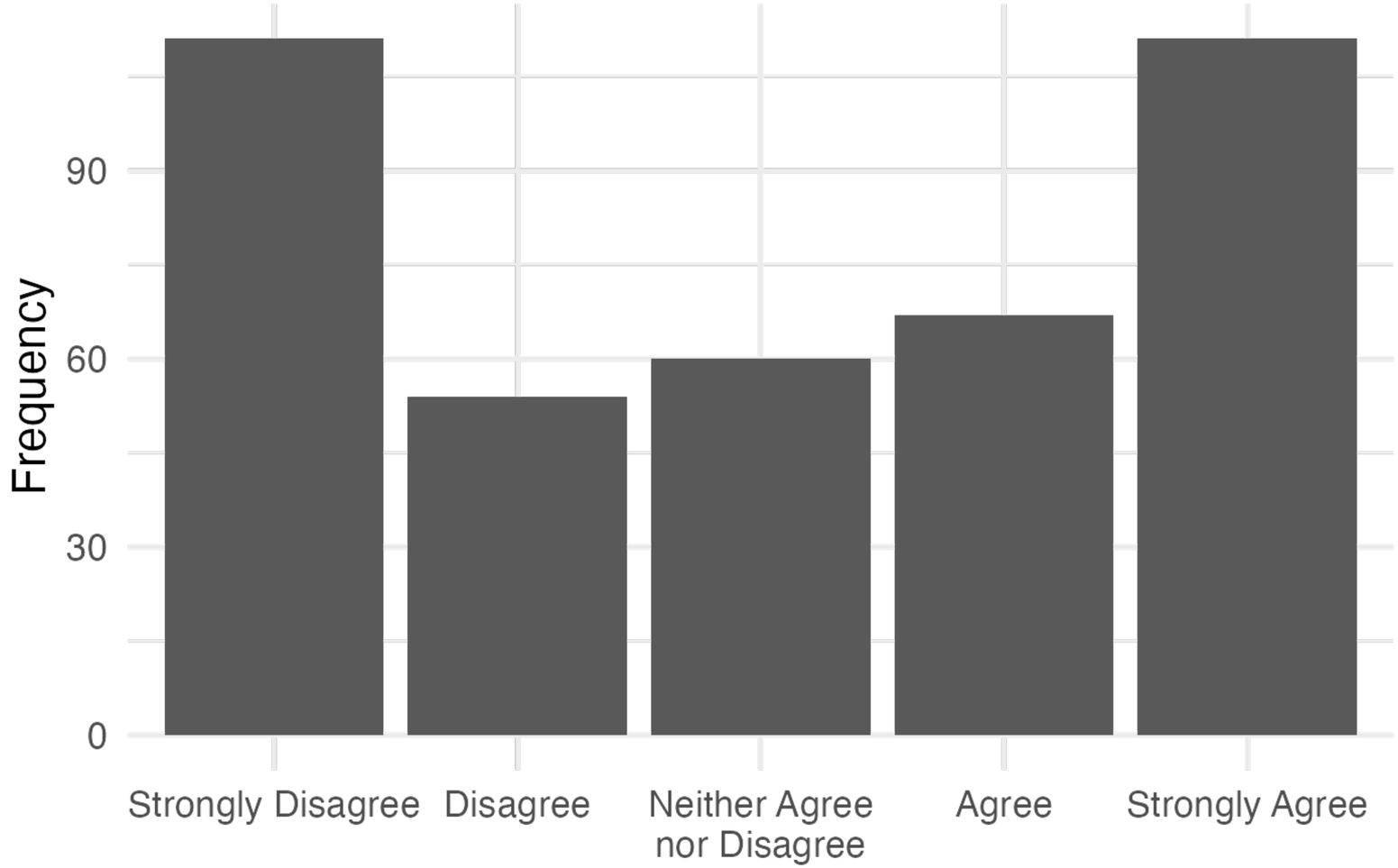
The Mode

- **Most Frequent value of a variable.**
- Useful for **nominal** and **ordinal** variables.
  - The mode for the “party voted” variable is “FPÖ”.
  - The modal response is “Disagree”.
- Makes sense with **discrete** numerical variables.
  - The modal number of siblings is 2.
- Rarely useful with highly **continuous** numerical variables.

The Mode

- A variable can have more than one mode. In this case, it is known as ‘bimodal’ (or multimodal).

- Bar plot: plot showing the distribution of frequencies of values in each category



universität wien

The Median ($\tilde{x}$)

- The median is the observation that falls in the middle of an **ordered** variable.
- When the number of observations is odd, a single observation occurs in the middle — the value of that observation is the median.
- When the number of observations is even, two middle observations occur, and the median is the midpoint between the two.
- Age of students: {18, 19, 19, 20, 21, 21, 22, 22, 23}
- Party share of the vote: {2.5%, 7.5%, 8.4%, 10.2%, 22.0%, 49.4%}

(8.4%+10.2%)/2 = 9.3%

universität wien

The Median ($\tilde{x}$)

- Formally, given a variable x of length n, ordered in ascending order so that $x_1 \leq x_2 \leq x_3 \leq \ldots x_n$, the median $\tilde{x}$ is…
  - $\tilde{x} = x_{\frac{n+1}{2}}$ if n is odd.
  - $\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$ if n is even.

universität wien

The Median ($\tilde{x}$)

- Properties of the median:
  - 50% of the variable's values fall above (below) the median.
  - As a result, the sum of the **absolute deviations** between each value and the median cannot be larger the sum of the **absolute deviations** between each value and any other number.
  - if $x = \{3,\ 17,\ 11,\ 28,\ 12\}$, then $\tilde{x} = 12$, then
  - $|3 - y| + |17 - y| + |11 - y| + |28 - y| + |12 - y|$ is ***smallest*** when $y = \tilde{x} = 12$.

universität wien

The Median ($\tilde{x}$)

- Uses of the median:
- Ordinal variables, **but not nominal variables**.
  - The median of {Strongly Disagree, Strongly Disagree, Strongly Disagree, Disagree, Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree, Strongly Agree} is “Disagree”.
- All numerical variables.

universität wien

The Mean ($\bar{x}$)

- The mean is **the sum of the observations divided by the number of observations**. It is also called the average or the expected value.
- E.g. height of patients: {1.78, 1.65, 1.77, 1.95, 1.57, 1.64}
- $$\bar{x} = \frac{1.78 + 1.65 + 1.77 + 1.95 + 1.57 + 1.64}{6} = \frac{10.36}{6} \approx 1.73$$

universität wien

The Mean ($\bar{x}$)

- Formally, for a variable x of length n, $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots x_n}{n}$
- Or, even fancier: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Where the summation symbol $\sum$ tells us that we need to add all the values of x from x_1 to x_n . Then, we divide by n.

universität wien

The Mean ($\bar{x}$)

- Properties of the mean:
  - "**zero-sum property**": the sum of the differences between every value and the mean is 0.
- if $x = \{3,\ 17,\ 11,\ 28,\ 12\}$, then $\bar{x} = 14.2$, then
- $(3 - 14.2) + (17 - 14.2) + (11 - 14.2) + (28 - 14.2) + (12 - 14.2) = 0$

universität wien

The Mean ($\bar{x}$)

- Properties of the mean:
  - "**least-squares property**": the sum of the **squared differences** between each value and the mean is smaller than the sum of the **squared differences** between each value and any other number.
- if $x = \{3,\ 17,\ 11,\ 28,\ 12\}$, then $\bar{x} = 14.2$, then
- $(3 - y)^2 + (17 - y)^2 + (11 - y)^2 + (28 - y)^2 + (12 - y)^2$ is ***smallest*** when $y = \bar{x} = 14.2$.

universität wien

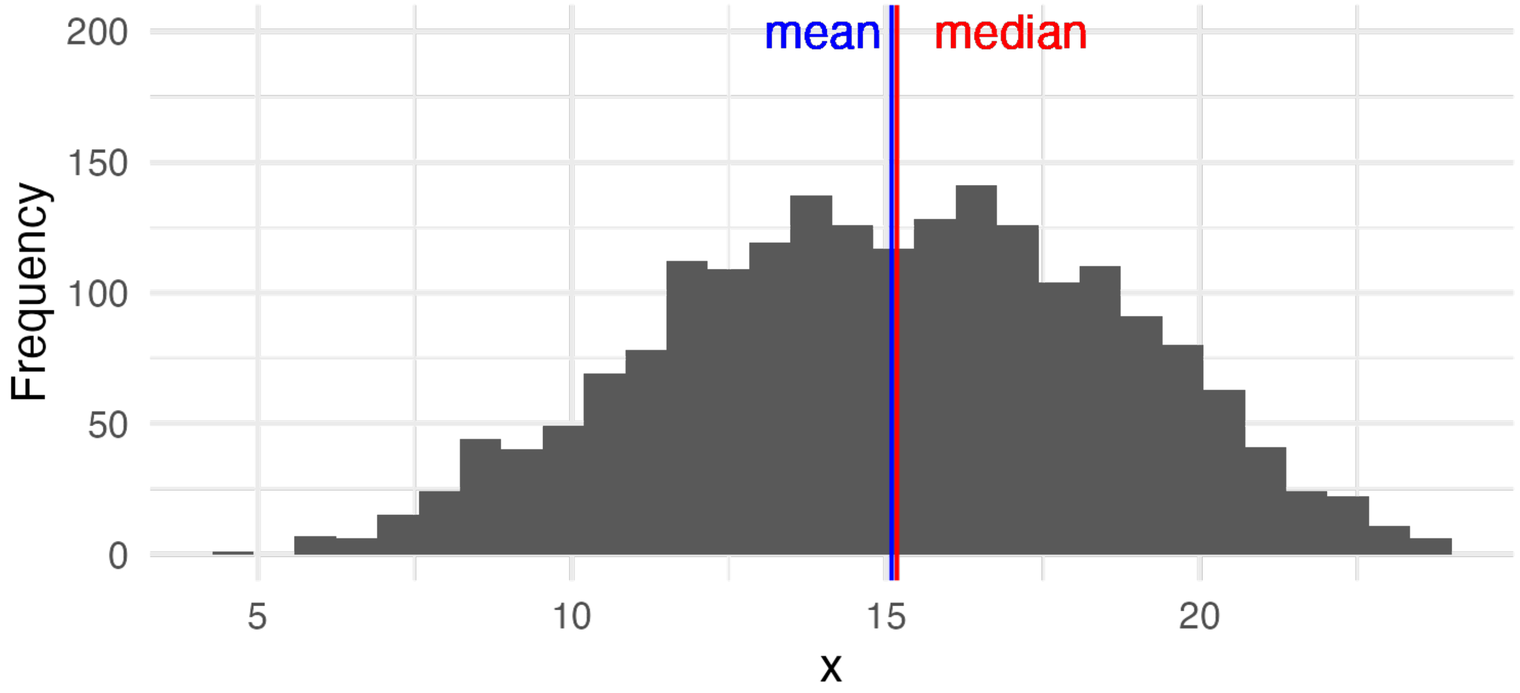
The Mean ($\bar{x}$)

- Uses of the mean:
- Numerical variables, ***but not nominal or ordinal variables***.
- One exception: binary variables, **where the mean represents the proportion of 1s in the variable**.
  - If $x = \{0, 0, 1, 1, 0\}$, then $\bar{x} = 2/5 = 0.4 = 40\,\%$.

Mean and Median Compared

- When the data are symmetrically distributed, the mean and the median tend to be similar…

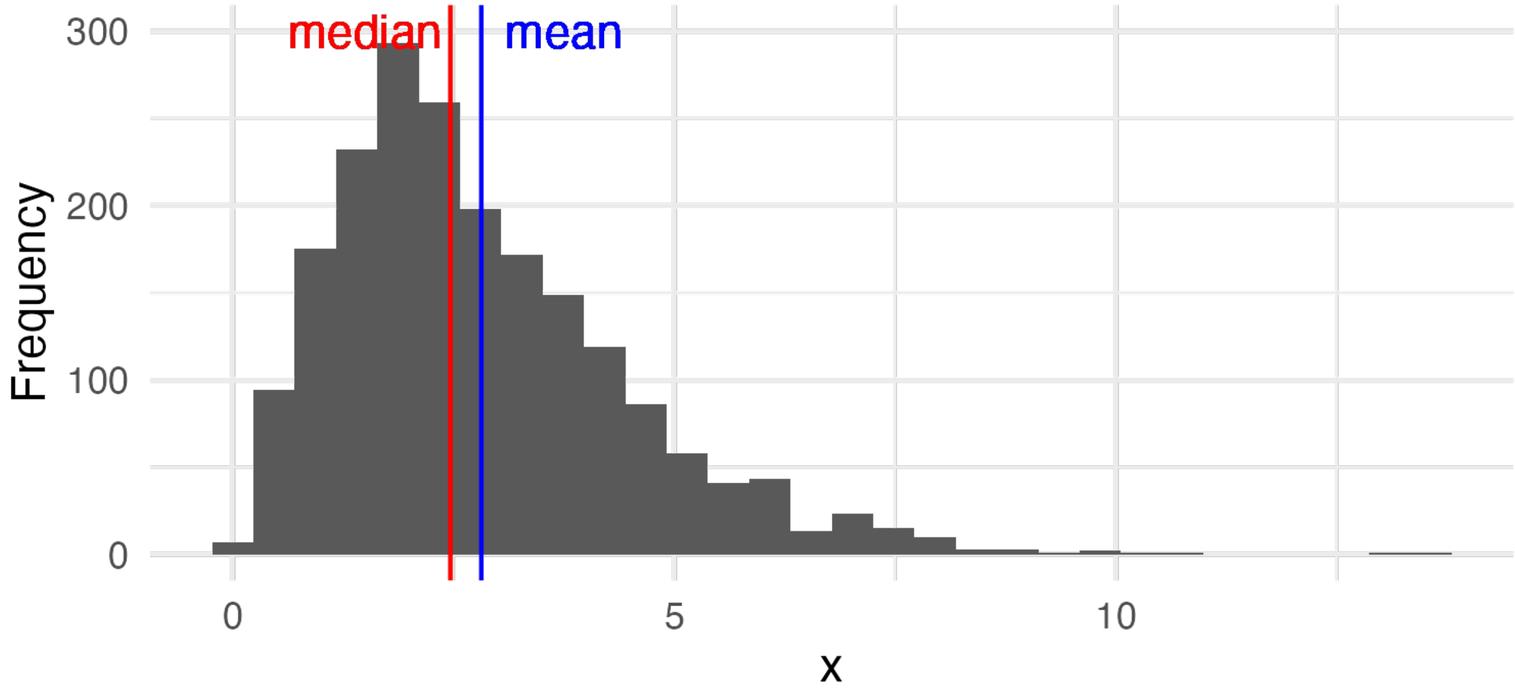
- Histogram: plot showing the distribution of frequencies of values within intervals (‘bins’)



Mean and Median Compared

- When the data are ***skewed***, the mean is pulled in the direction of the longer tail.

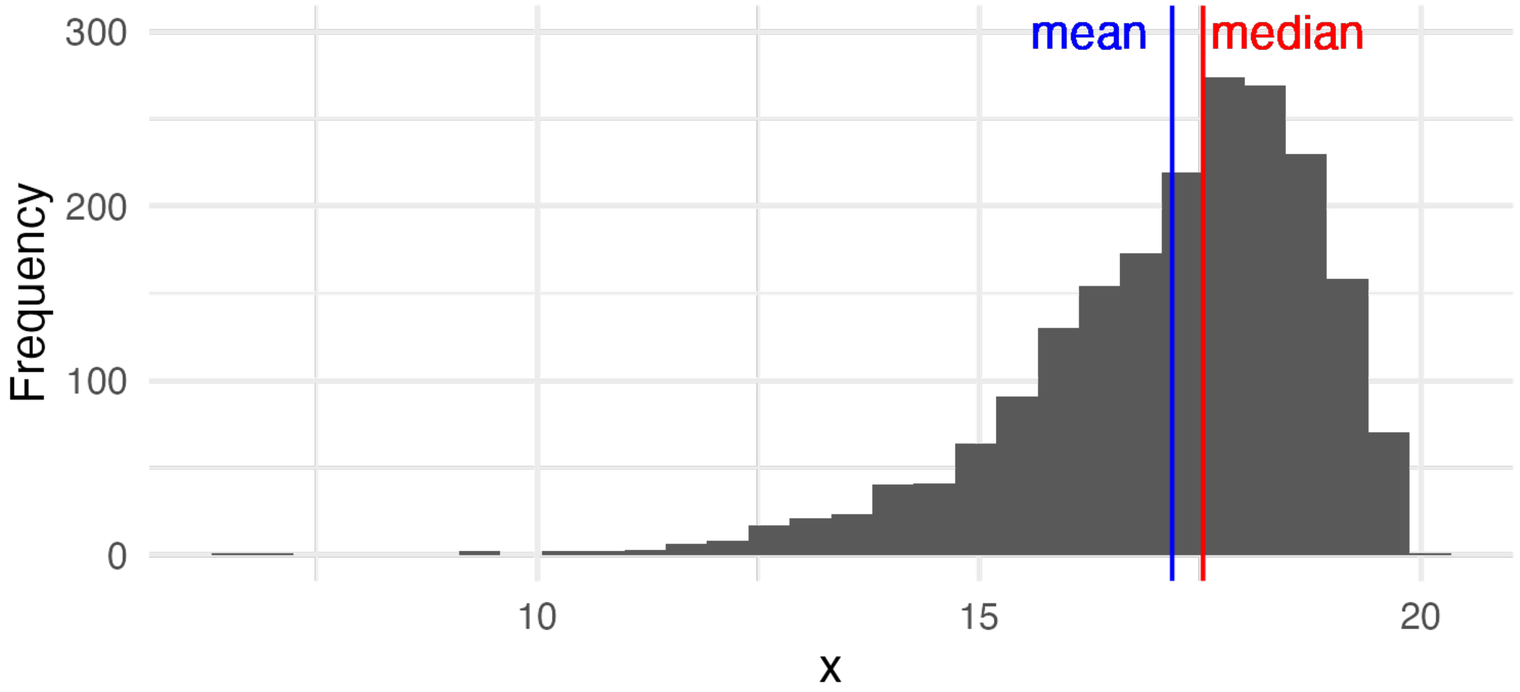
- Histogram: plot showing the distribution of frequencies of values within intervals (‘bins’)



Mean and Median Compared

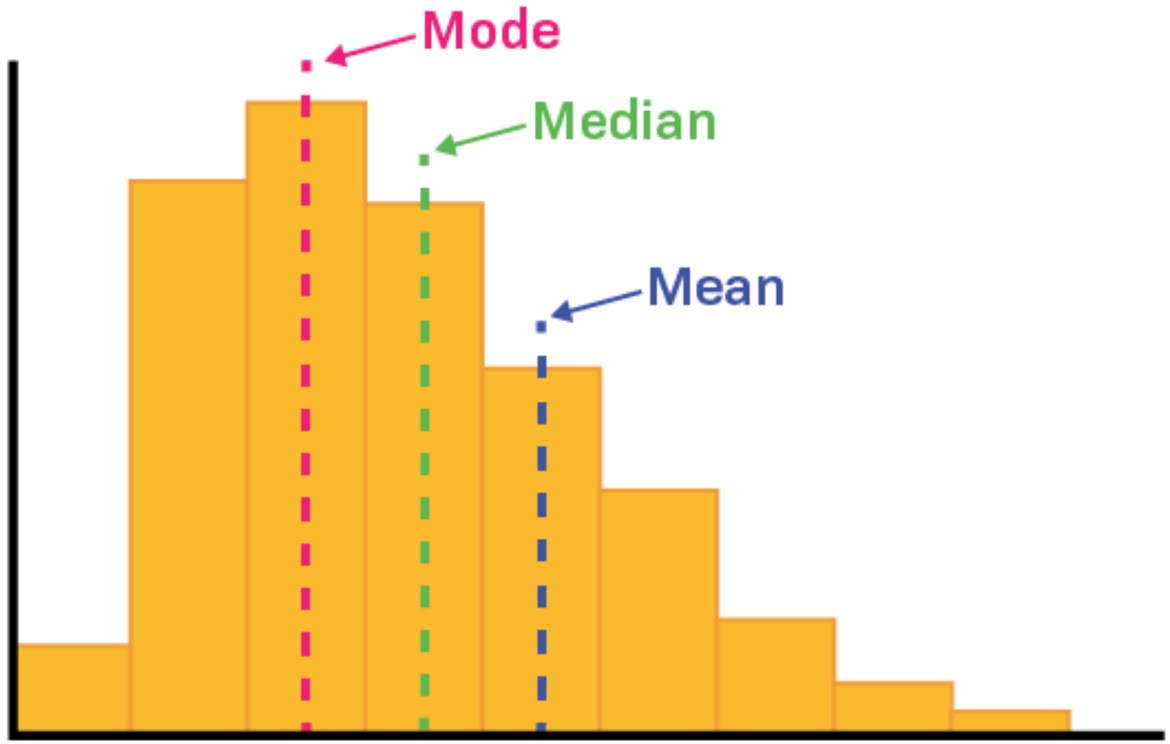
- When the data are ***skewed***, the mean is pulled in the direction of the longer tail.

- Histogram: plot showing the distribution of frequencies of values within intervals (‘bins’)



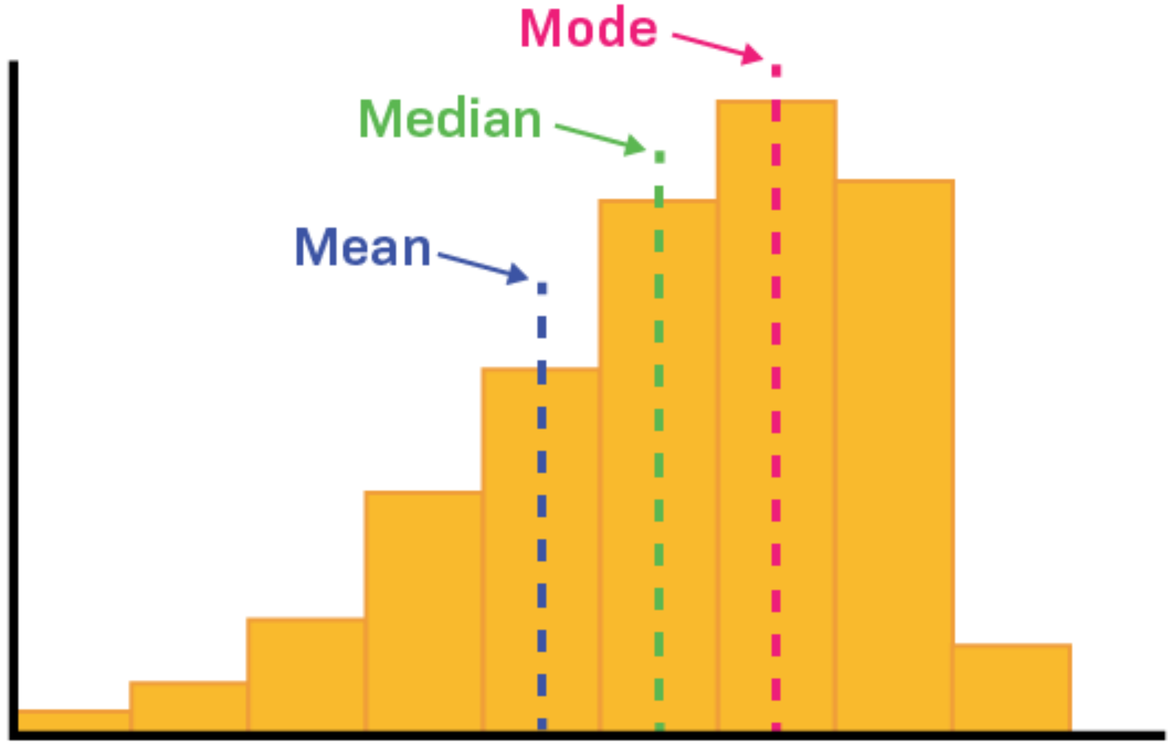
Mean and Median Compared

B. Right-skewed (or Positive-skewed)



- E.g. income

C. Left-skewed (or Negative-skewed)



- E.g. age of retirement

Mean and Median Compared

- As a result, the median can be more representative of a ‘typical’ value of a variable in presence of **outliers.**

- x (wage) = {1800€, 2300€, 3200€, 3900€, 5500€, 7000€, 500000€}

$\bar{x}$ = 74814.29€ $\tilde{x}$ = 3900€

universität wien

And now, let's open RStudio…