universität wien

BAK3: Introduction to Quantitative Methods

Week 5: Univariate Descriptive Statistics II

Leonardo Carella

universität wien

The Plan for today

- Statistics:
  - Measures of Position: **quantiles**.
  - Measures of Spread (or “Dispersion” or “Variability”): the **interquartile range**, the **variance** and the **standard deviation**.
- Coding in R:
  - Measures of Spread in R and Boxplots.

universität wien

Measures of Position

- By ‘position’ (or relative standing) we mean the **location of a data point within a variable by comparing it to other values.**
- We already know one measure of position: the **median**. It’s the mid-point of the variable: 50% of values are smaller, 50% are larger.
- Today: **quantiles**, as a generalisation of the mean.
- Other measure of position (later in the course, briefly): z-score.

universität wien

Quantiles

- The n quantile of a distribution is the point such that $(n \times 100)\%$ of the observations fall below that point.
- In this sense, the median is the **0.5 quantile**. But you can also get the 0.25, the 1/3, the 0.9 quantiles etc…
- Often described as **percentile**: The p^{th} percentile of a distribution is the point such that $p\%$ of observations fall below that point.

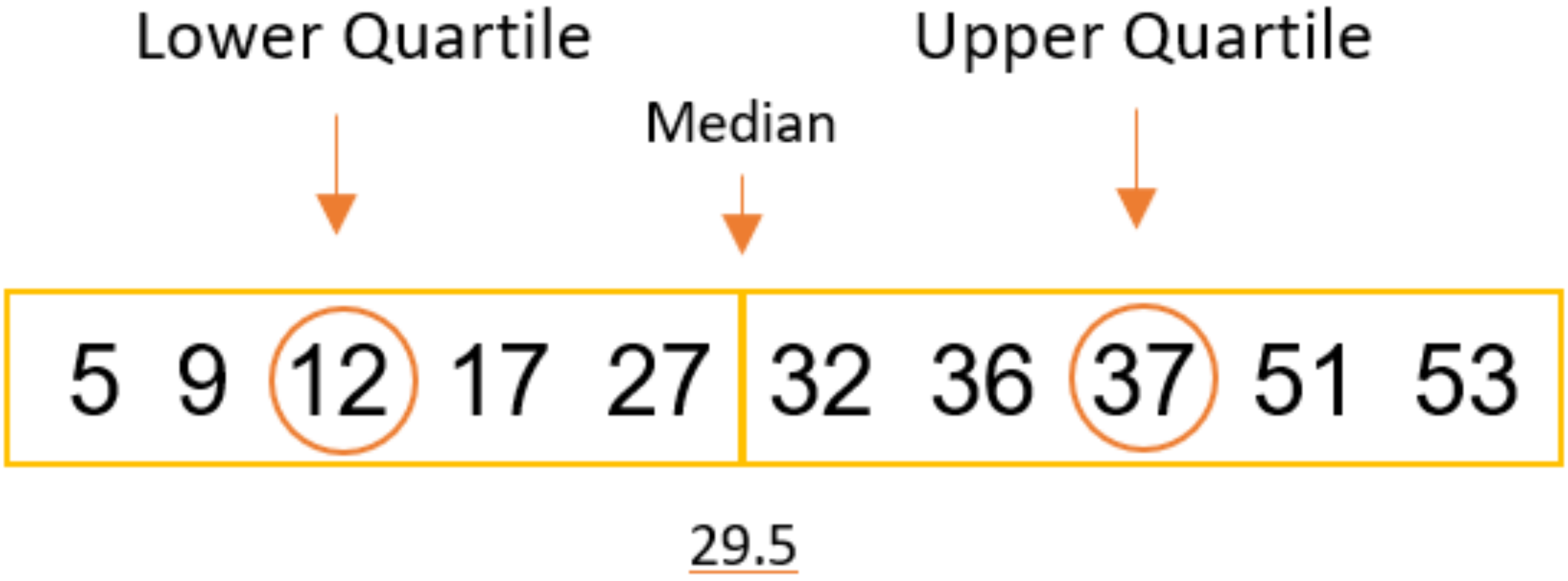
universität wien

Quantiles and Quartiles

- Key quantiles:
- 0.25 quantile, known also as ***first quartile***, or ***lower quartile***, or ***Q1*** or ***25th percentile***. 25% (one-quarter) of observations fall below that point; 75% (three-quarters) of observations fall above.
- 0.75 quantile, known also as ***third quartile***, or ***upper quartile*** or ***Q3*** or ***75th percentile***. 75% (three-quarter) of observations fall below that point; 25% (one-quarter) of observations fall above.

universität wien

Quartiles



universität wien

Interquartile Range

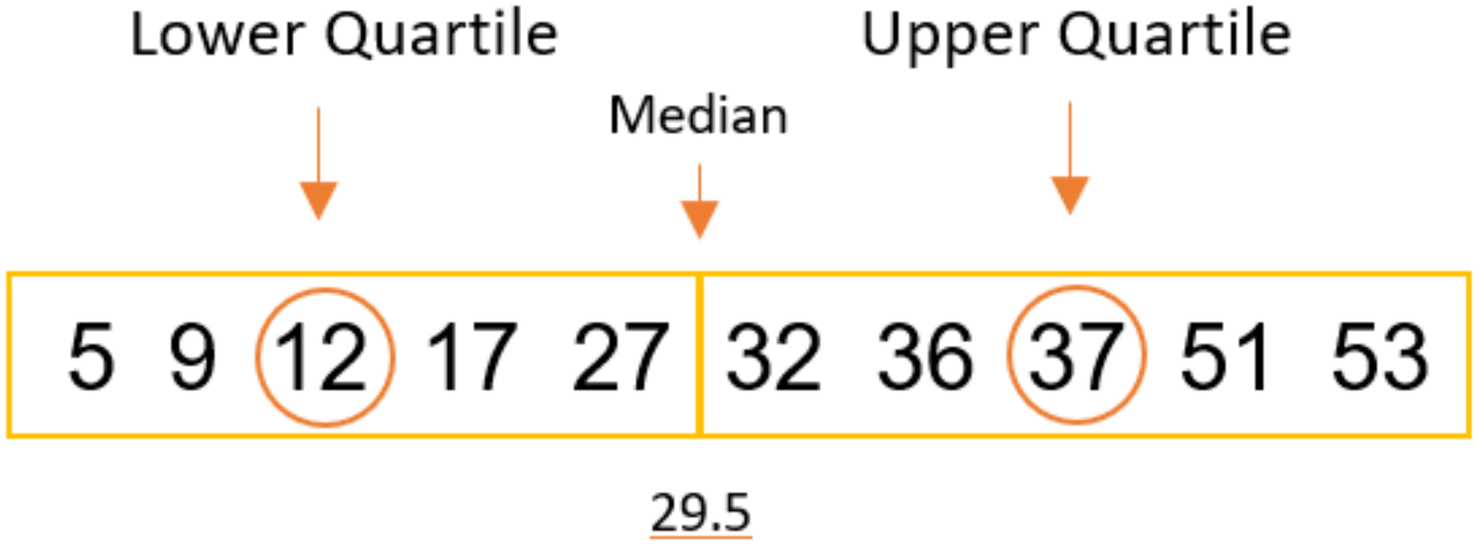
- We like the first and third quartile (Q1 and Q3) because 50% of values fall between these two, giving us a sense of how far from the median we find half of the population distribution.
- So, our first measure of spread is the **interquartile range**:

$$IQR = Q3 - Q1$$

  - The ‘wider’ the IQR, the more variability we have in our variable.

Interquartile Range

- Q3 = 37
- Q1 = 12
- **IQR** = 25



- Why should we care?
  - Definition of ***outliers*** is done with reference to the IQR.
  - Used in a common form of data visualisation: the ***boxplot***.

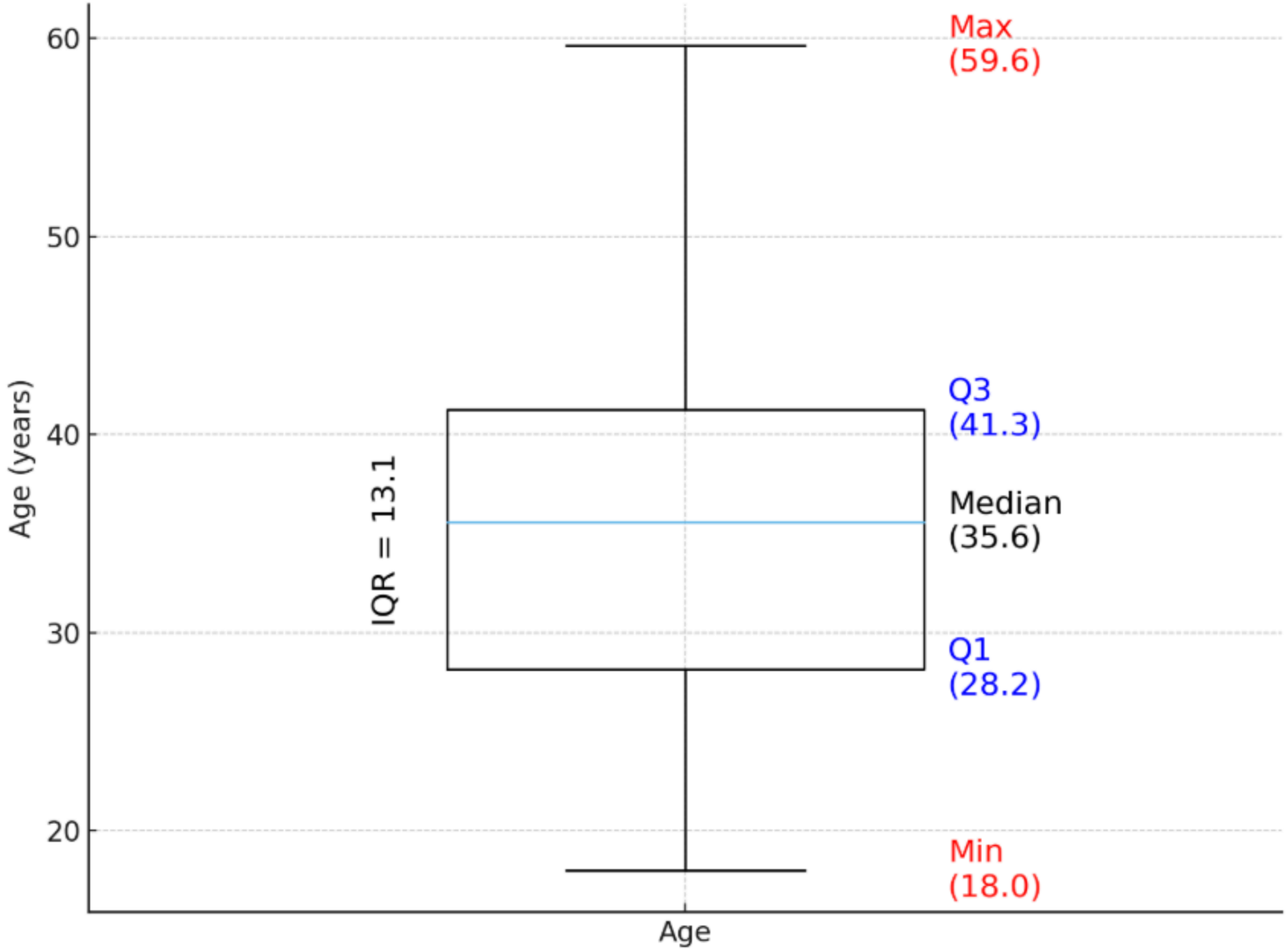
universität wien

Outliers

- An observation is an ***outlier*** if it falls more than 1.5(IQR) above the upper quartile (Q3) or more than 1.5(IQR) below the lower quartile (Q1).
- For example: wages = {1800€, 2035€, 2300€, 2400€, 3200€, 3550€, 3900€, 5500€, 7000€, 500000€}, where **Q1**=2300€ and **Q3**=5500€.
  - **IQR** = 5500-2300 = 3200
- The observation with a wage of 500000€ is an outlier, because…
  - Q3 + 1.5(IQR) = 5500 + 1.5(3200) = 10300; and 500000 > 10300

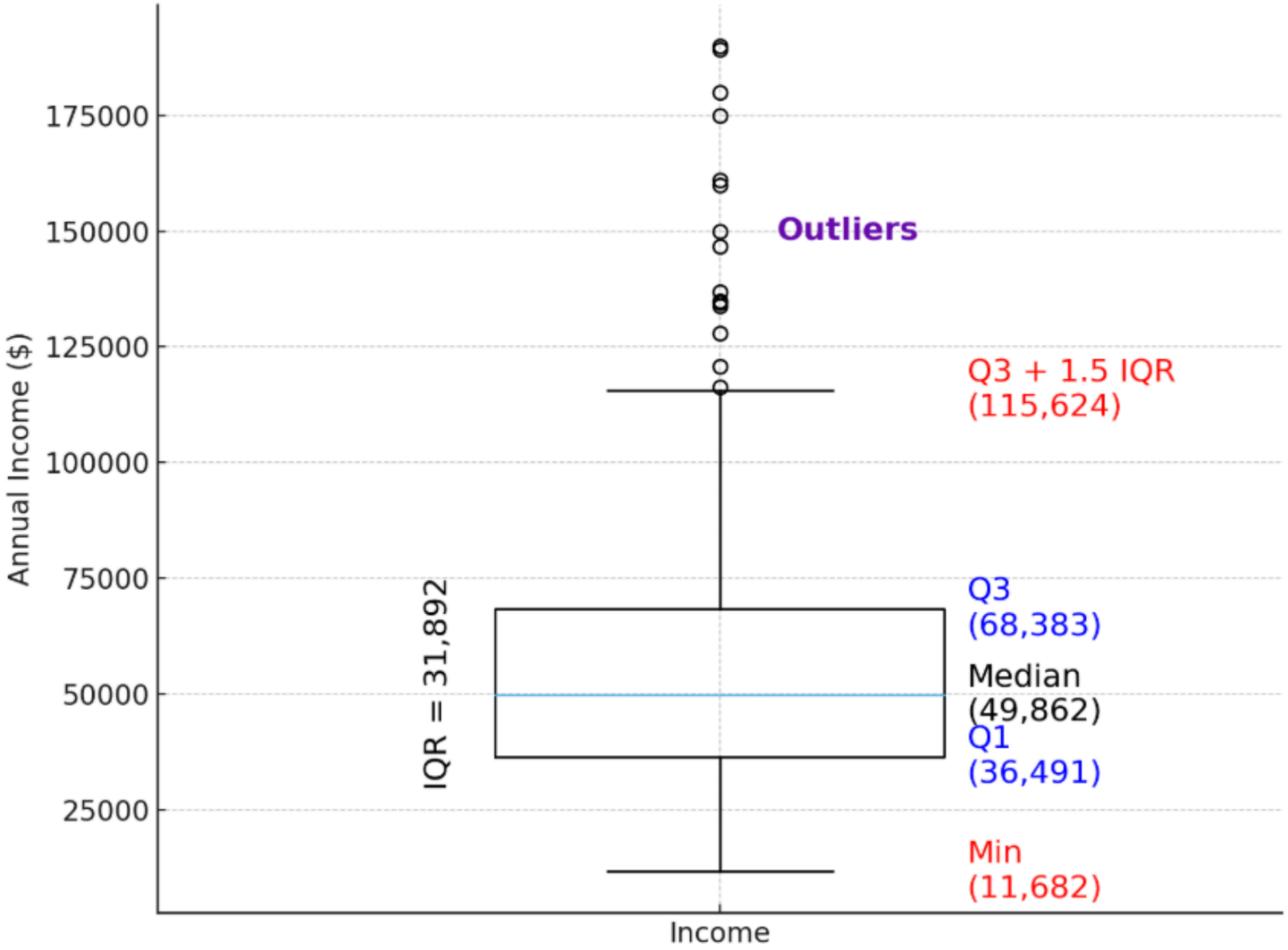
universität wien

Boxplots



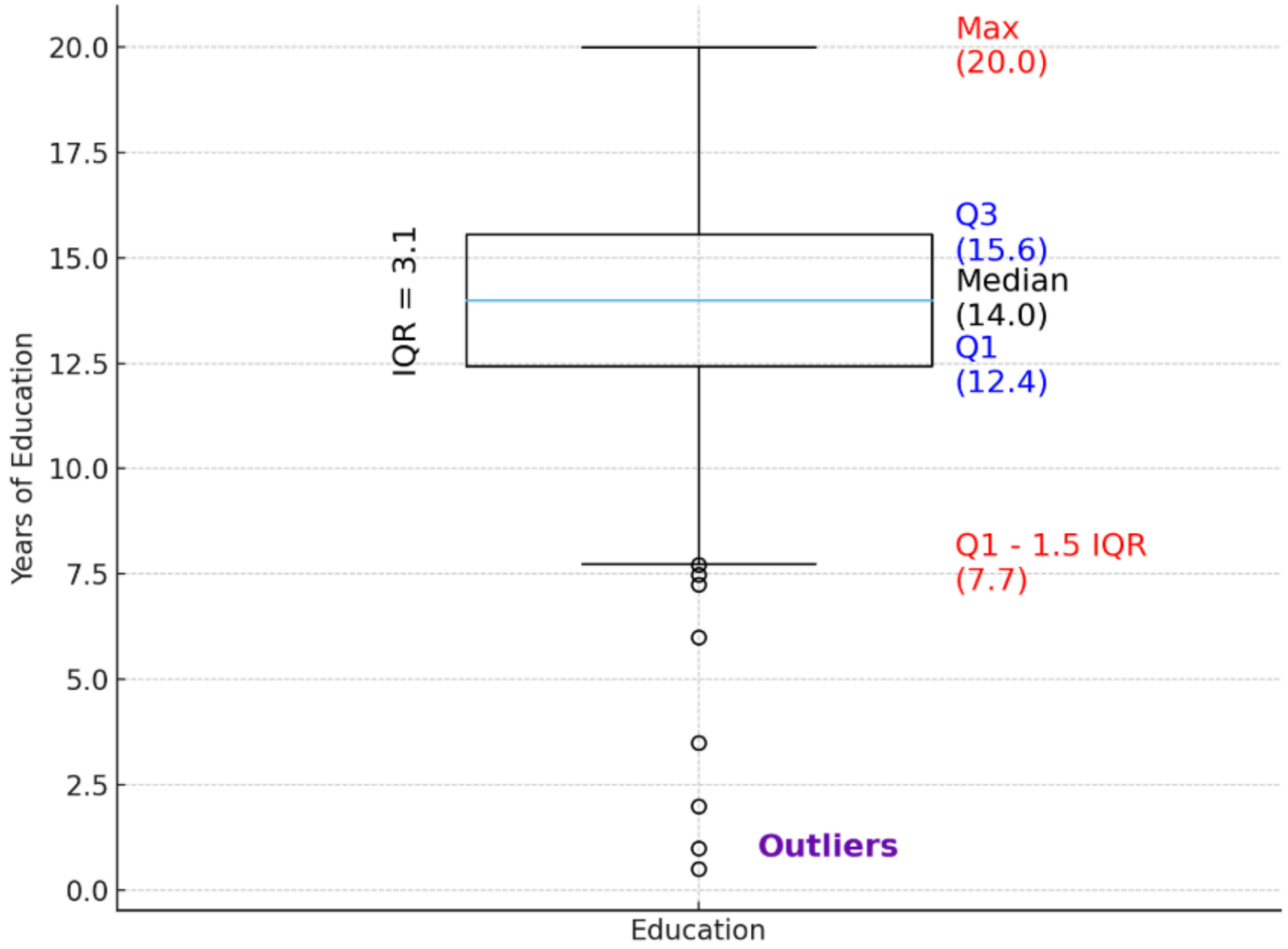
universität wien

Boxplots (with outliers)



universität wien

Boxplots (with outliers)



universität wien

Variance

- The variance is another measure of spread that assesses how different each observation is from the **mean** observation.
- The formula is:

$$\mathrm{Var}(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

universität wien

Variance

$$\mathrm{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Breaking it down:
  - $(x_i - \bar{x})$ means: take the **differences between every single value** $x_1, x_2, x_3 \ldots x_n$ of our variable X, **and the mean** of X, which is $\bar{x}$.
  - Then, we **square all** of them, so all the signs of these ‘deviations’ become positive.
  - The summation sign Σ means: **add up** all these squared deviations.
  - Then, divide by the number of observations **minus one**: $n - 1$.

universität wien

Variance of $X = \{13,16,8,5,8,7\}$

- Step 1: find the mean $\bar{x} = (13 + 16 + 8 + 5 + 8 + 7)/6 = 9.5$
- Step 2: $(x_i - \bar{x}) = \{13 - 9.5; 16 - 9.5; 8 - 9.5; 5 - 9.5; 8 - 9.5; 7 - 9.5\}$
- Step 3: $(x_i - \bar{x})^2 = \{3.5^2; 6.5^2, -1.5^2, -4.5^2; -1.5^2; -2.5^2\}$
- Step 4: $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 12.25 + 42.25 + 2.25 + 20.25 + 2.25 + 6.25 = 85.5$
- Step 5: divide by $n - 1$. Which is to say $85.5/(6 - 1) = 17.1.$

universität wien

Why does this work?

- The variance is a measure of spread because it gets smaller as values ‘cluster’ around the mean, and bigger as observations ‘spread out’.
- Consider these two variables with the **same mean:**
  - ***No variation.*** $X = \{3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3...\}$ with $\bar{x} = 3$.
  - ***Lots of variation.*** $X = \{+2250, -815, -1426...\}$ with $\bar{x} = 3$.

universität wien

Why does this work?

- ***No variation***. $X = \{3,\ 3,\ 3,\ 3,\ 3,\ 3,\ 3...\}$ then $(x_i - \bar{x}) = \{0,\ 0,\ 0,\ 0,\ 0...\}$, then $\mathrm{Var}(X) = 0.$
- ***Lots of variation.*** $X = \{+2250, -856, -1456...\}$, then $(x_i - \bar{x}) = \{\text{Big Number}, \ -\text{Big Number}, \ -\text{Big Number}\ldots\}$, then $(x_i - \bar{x})^2 = \{\text{Huge Number}, \text{Huge Number}, \text{Huge Number}\ldots\}$. Sum them all up and divide by $n - 1$, $\mathrm{Var}(X) =$ **Big Positive Number.**

universität wien

Unsatisfactory Answers to Very Good Questions

- **Why do we take the squares and not the absolute value of deviations?** Mostly for mathematical convenience: squaring has nice algebraic properties that the absolute value does not. Also connected with the ‘least squares’ property of the mean from last week.
- **Why do we divide by $n-1$ and not by n?** You could, and some formulas for the variance do divide by n. The ‘minus one’ is known as **Bessel’s correction**. We divide by $n-1$ when we want to compute the sample variance, we divide by n when we want the population variance. Here, we mostly work with samples, so usually we want $n-1$. But don’t worry about this now.

universität wien

The Standard Deviation

- The standard deviation is simply the **square root of the variance:**

$$s_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- We normally prefer to discuss the standard deviation because it's in the same unit of measurement (cm, percentage points, number of children, euros, years etc.) as the original variable.

universität wien

The Standard Deviation

- For instance, if our variable is people's years = {59, 23, 42, 31, 18, 38}
- The variance is 216.6 "squared years". No immediate sense.
- The std. deviation is 14.7 **years**. For instance, we could say that the 23 year-old is about one standard deviation younger than the mean.

universität wien

The Standard Deviation

- **Properties of the standard deviation:**
  - It is always positive (at most it can be zero).
  - The greater the variability around the mean, the larger it gets.
  - Like the mean (and unlike the IQR and the median), it is affected by all values of the variable, **including outliers**.
- ***Notation***: normally, the standard deviation is notated as σ or as s. The variance is notated in terms of the standard deviation as σ^2 or as s^2.

universität wien

Summing Up…

- **Q1** and **Q3** bracket 50% of observations in the middle of a variable distribution; they are used to compute the **interquartile range**.
- **Boxplot**: visual summary of key points in a variable's distribution: minimum*, Q1, median, Q3, maximum*; outliers if there are any.
- The **standard deviation** is the square root of the **variance**. Both measures summarise the spread of the variable in a single positive number, but the former is more common and easier to interpret.
- And now, let's open RStudio…