Basics of quantitative methodology

E2040 Week 4

2024

Study objectives

At the end of this lesson, student will be able to:

1. Understand the basics of quantitative methodology
2. List basic types of variables
3. Know basic types of data visualization
4. Know basic types of summary statistics

Types of variables

Categorical

- **Nominal**
- **Binary/dichotomous**
- **Ordinal**

Continuous

Categorical variables

Nominal – qualitative values, no ordering

- sex, ethnicity, birth month

Binary/dichotomous – two categories

- No/yes, dead/alive, case/control

Ordinal – have several ordered categories

- Level of education (elementary, high school, college), Likert scale ([1]strongly disagree – strongly agree[5])

Continuous variables

Can take any value within a range

- Theoretically infinite values

Examples: blood pressure, height, temperature, liquid volume

KvIS 1

Distribution

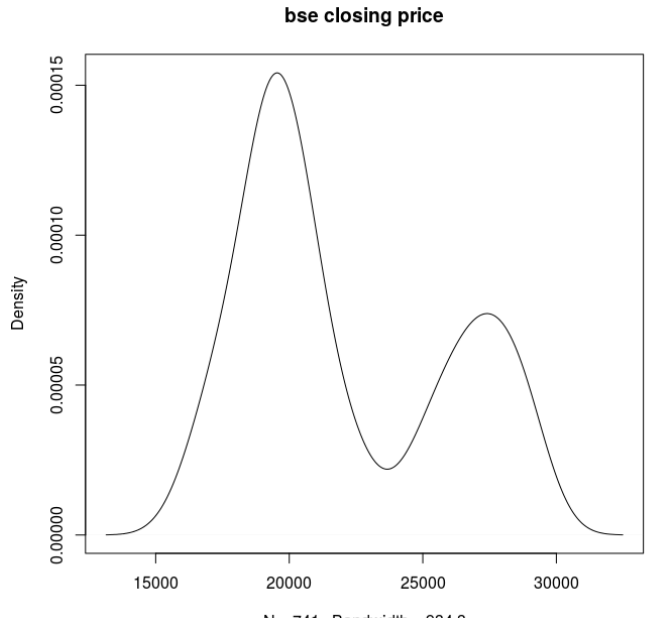
How data values are spread across different values

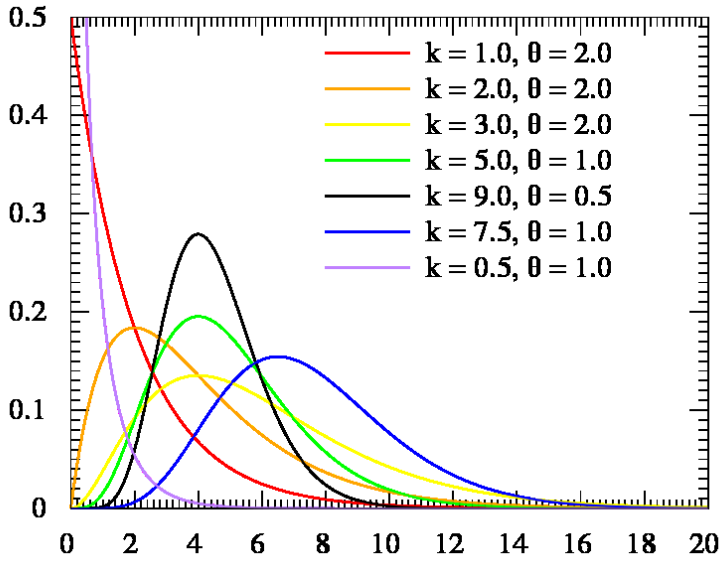
Normal distribution – Gaussian, symmetrical

Skewed distribution

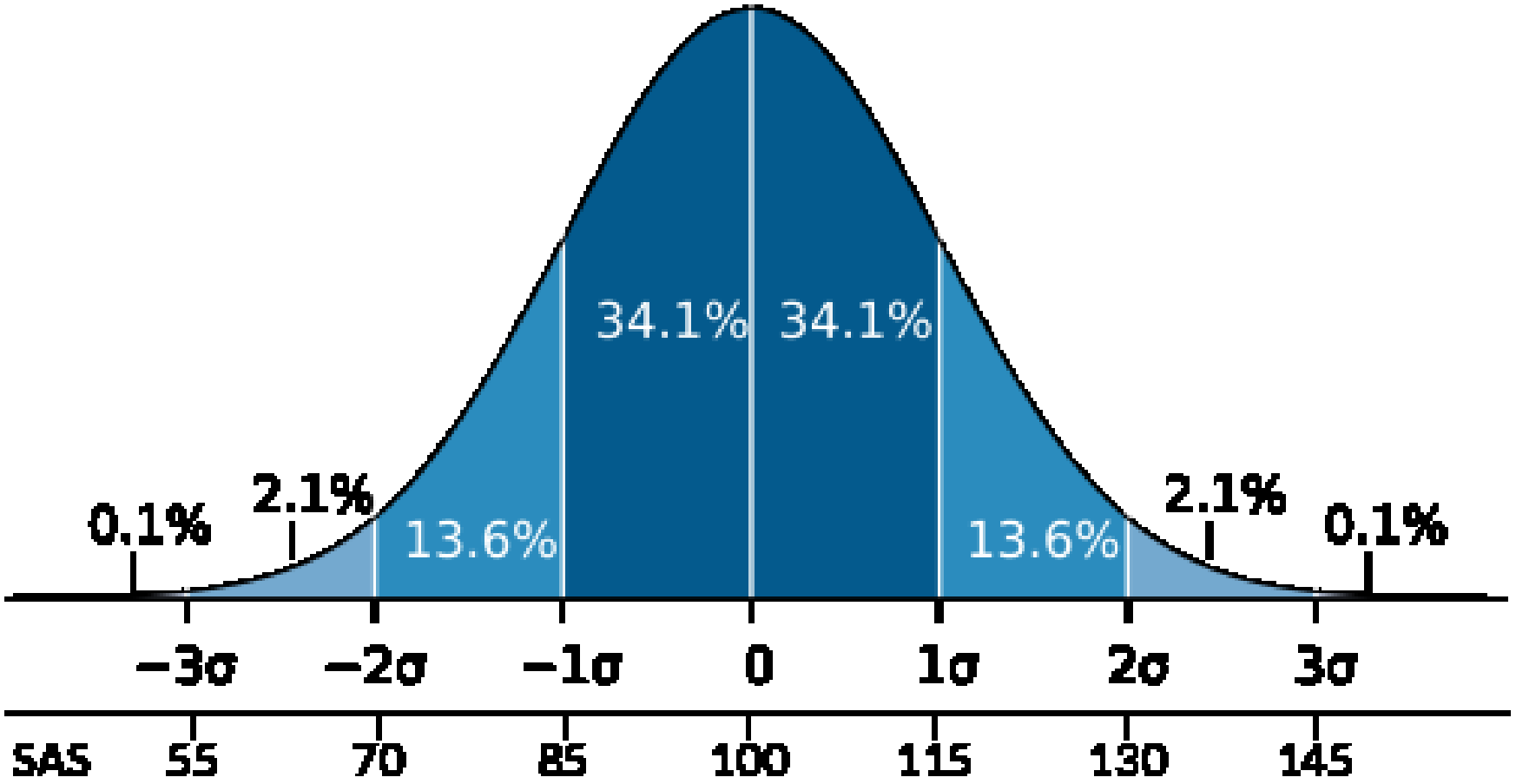
- Positively skewed (right-skewed), negatively skewed (left-skewed)

Bimodal distribution





Normal/Gaussian distribution

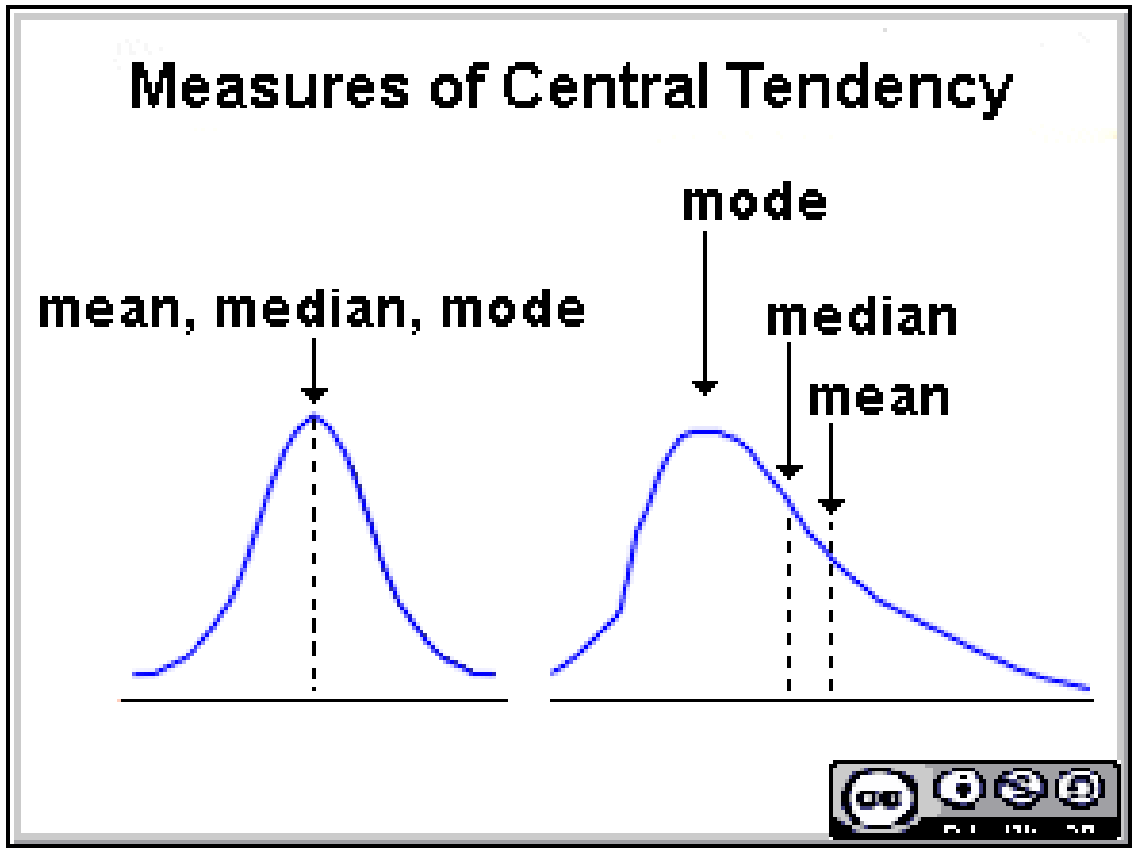


Galton board and the laws of nature

Measures of central tendency

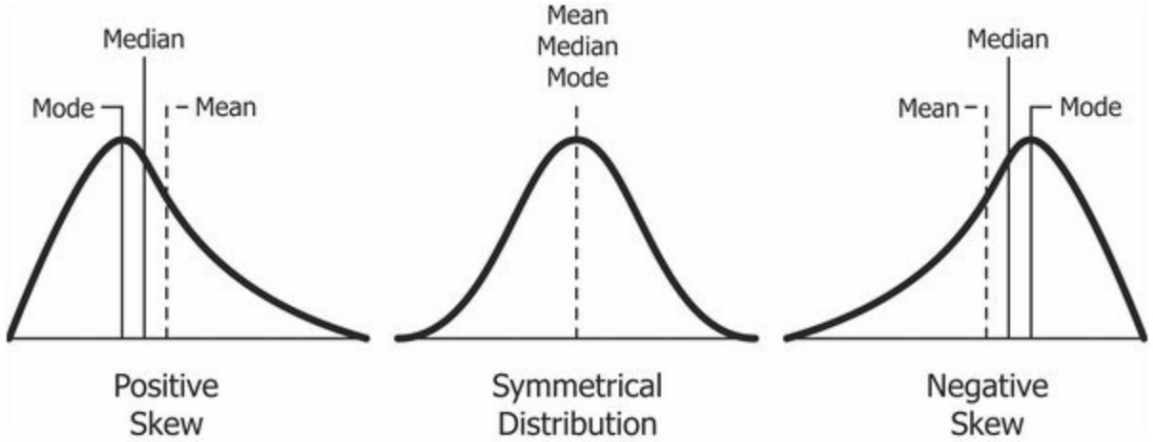
1. **Mean** – the average value (sum of values / number of values)
2. **Median** – the value in the middle of distribution (50th percentile)
3. **Mode** – the most frequent value

Measures of central tendency

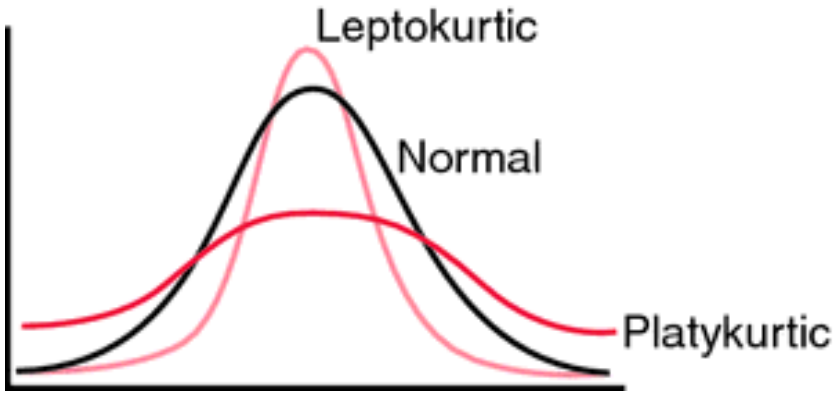


Non-normal distributions

Skewness

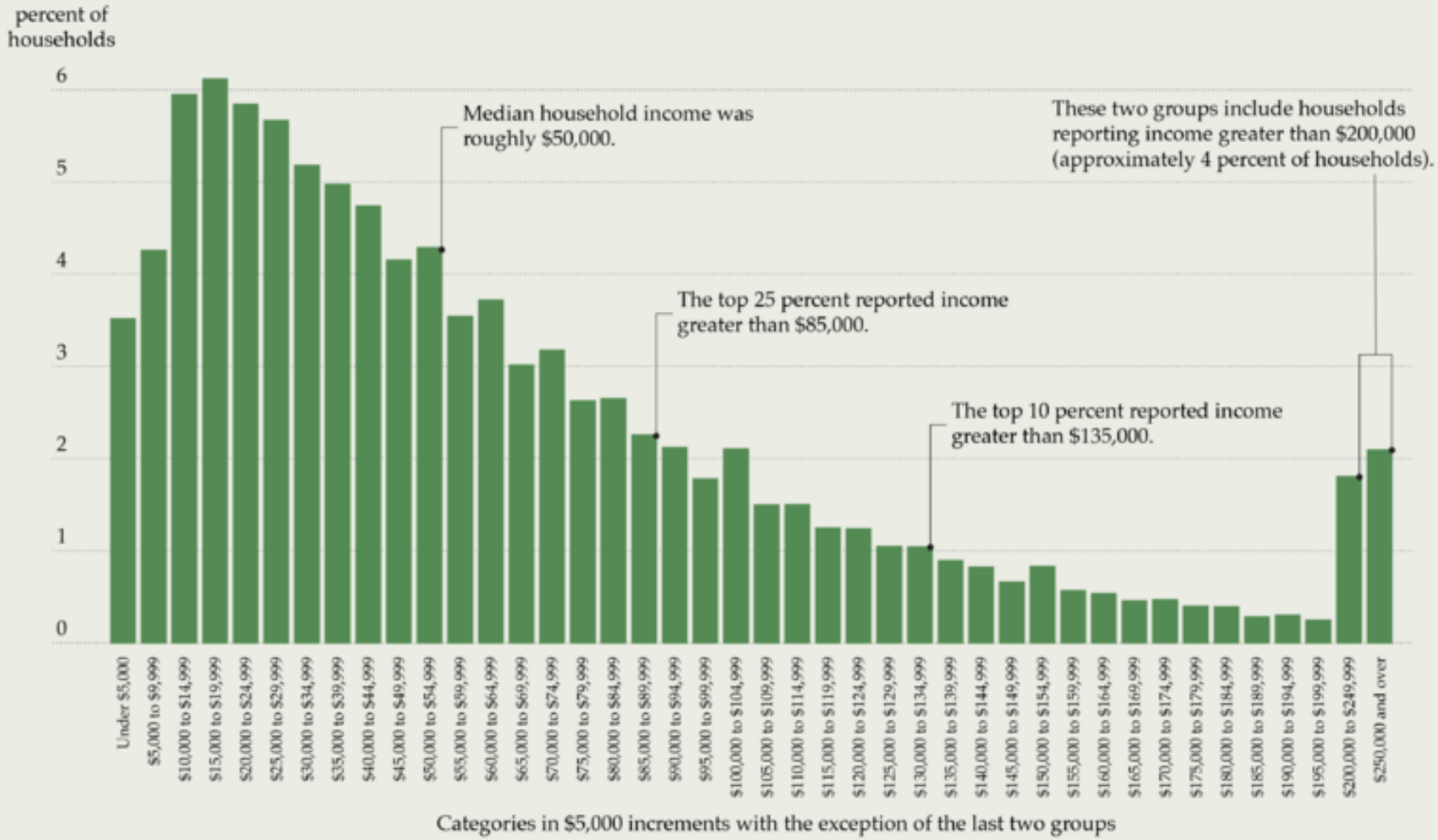


Kurtosis



Distribution of annual household income in the United States

2010 estimate



Categories in $5,000 increments with the exception of the last two groups

Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

KvIS 2

Measures of spread

How **spread** our data are around the central tendency

The lower the spread, the more representative the measures of central tendency are of the data

High spread = large variability

Standard deviation (SD)

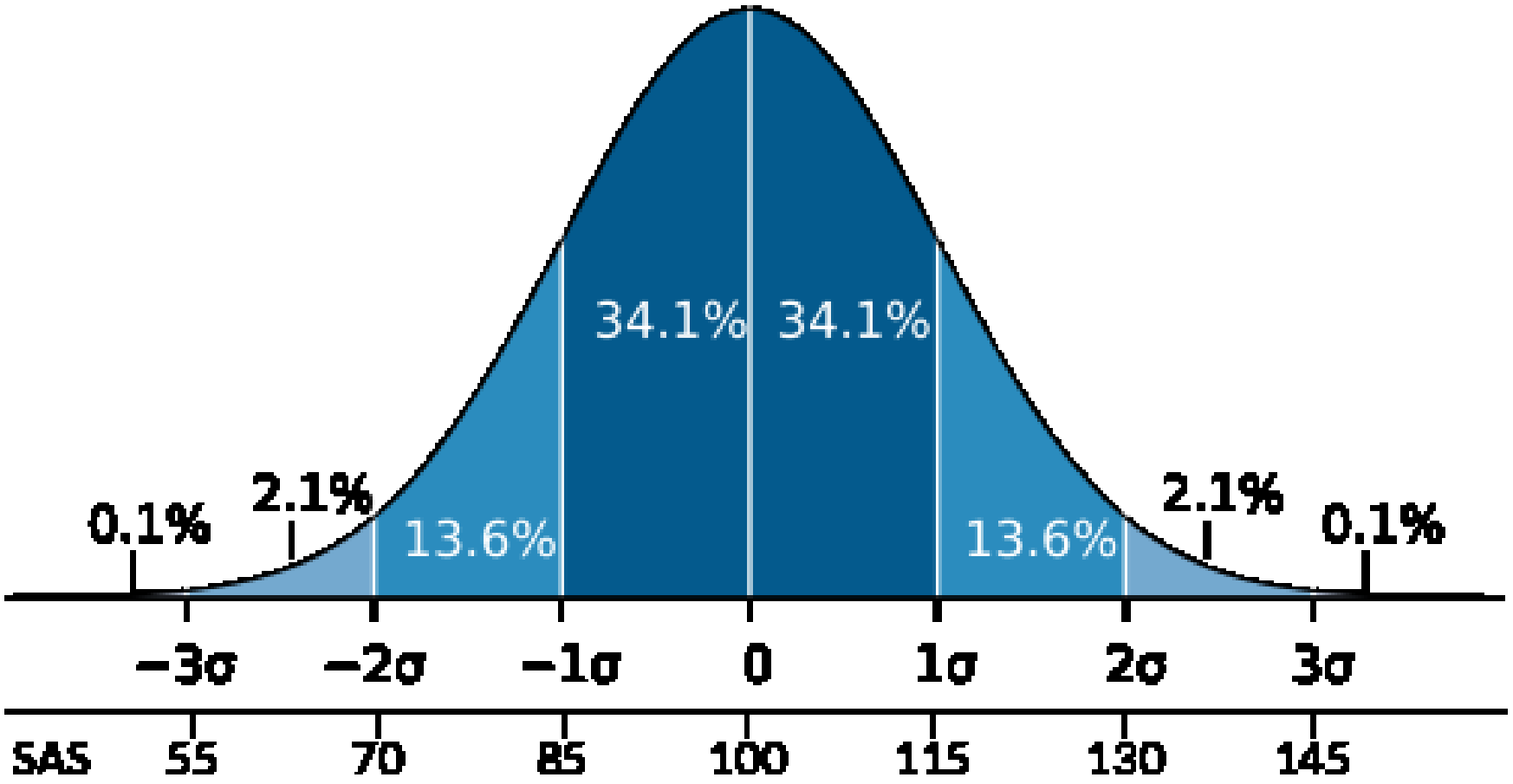
Amount of variation of the values from the mean

High SD = high variability

(population SD): square root of the mean of squared differences of individual values from the mean = **variance**

Sample SD = using n-1 instead of N

$$\sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$


34.1%
34.1%
0.1%
2.1%
13.6%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
SAS
55
70
85
100
115
130
145

Basic descriptive terms

Sum – adding values together

Mean (M) – sum of values divided by their count

Mode – most frequently occurring value

Median – value at the 50% (“in the middle”)

Standard deviation (SD) – distance of a value from a sample mean

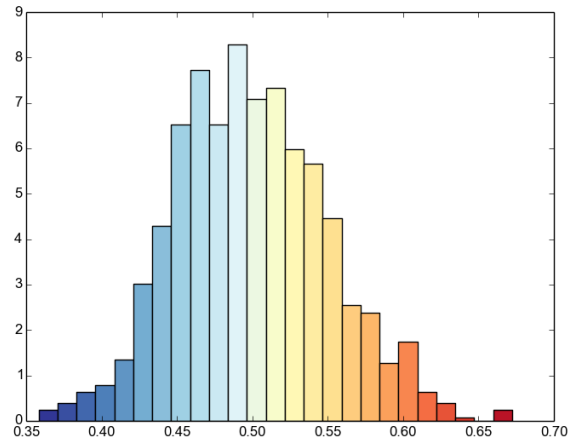
Variance – squared SD

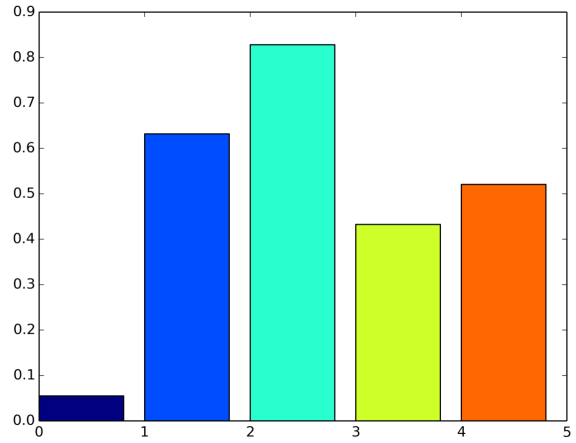
Quantile – cut point dividing the range of the distribution into intervals with equal probabilities

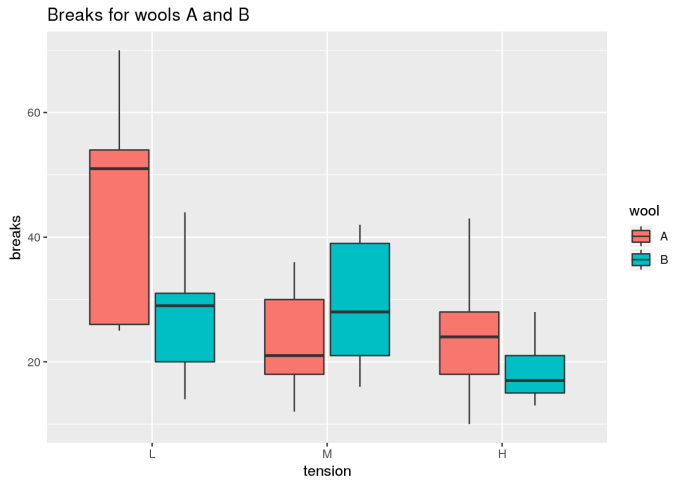
Minimum – the smallest value

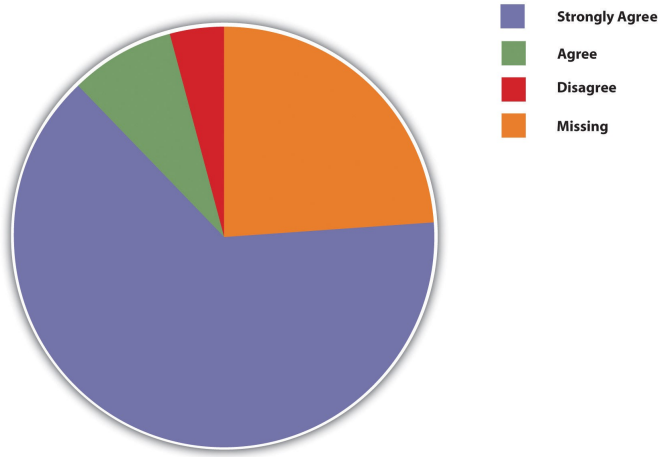
Maximum – the largest value

Visualizing the distribution







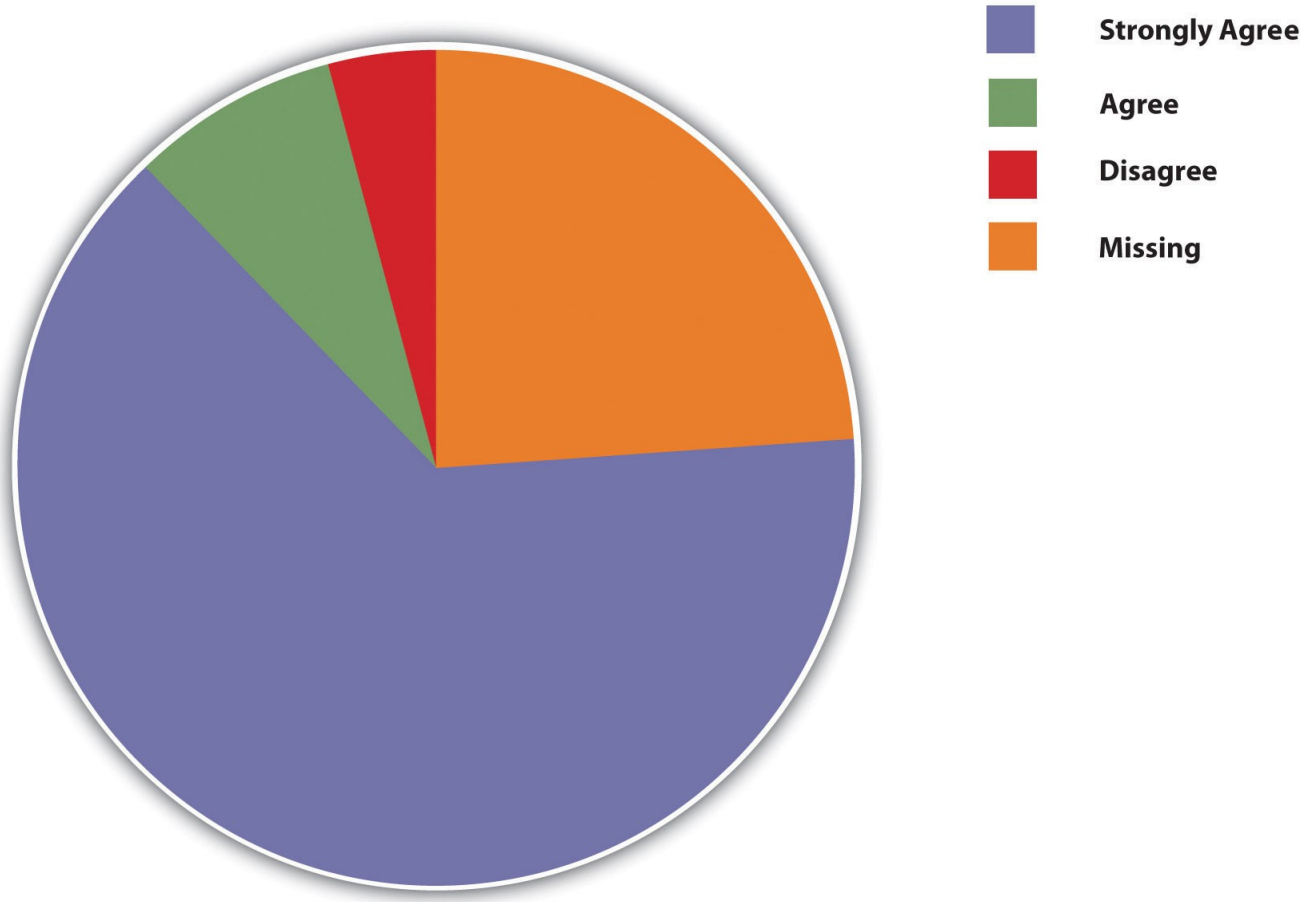


Pie chart

Categorical variables

Visualizing the proportions of values

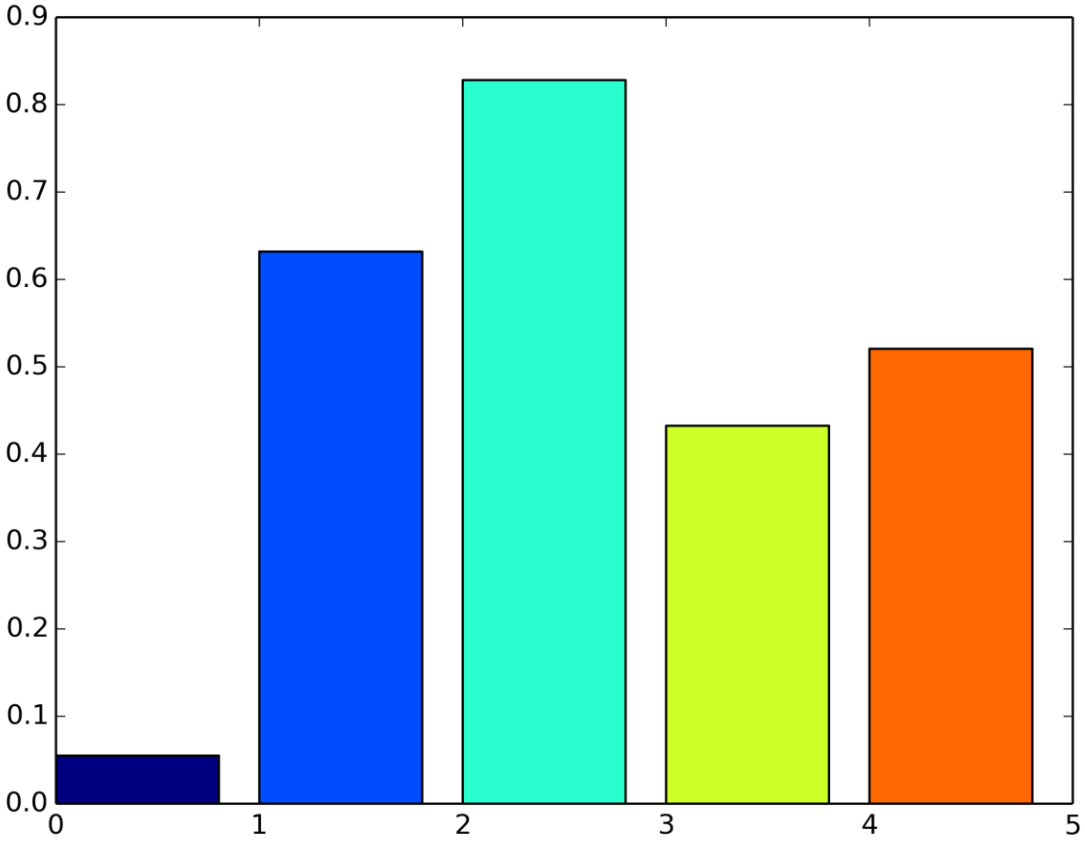
Typically as a percentage



Bar chart

Categorical variables

Frequency of values for each category



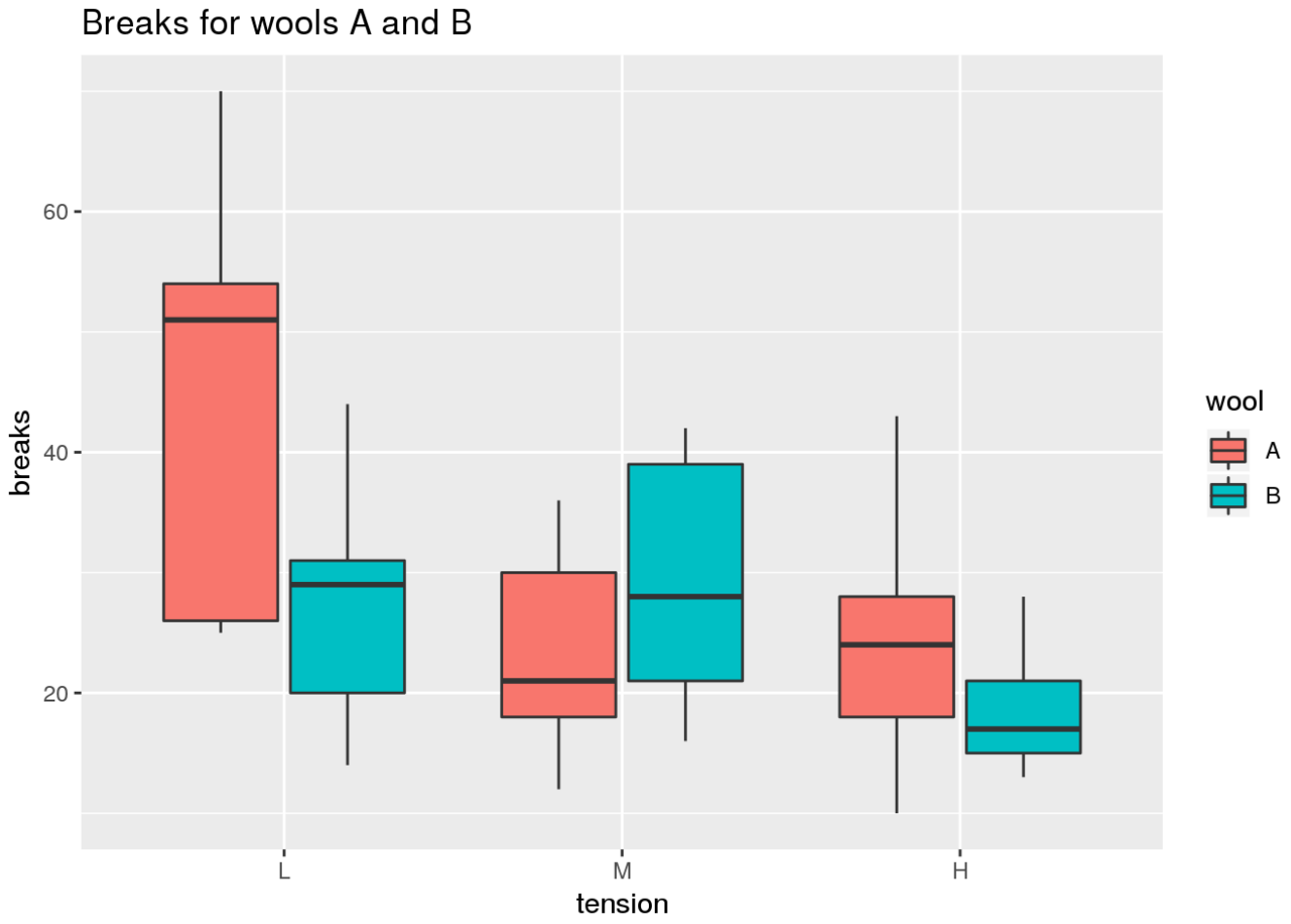
Boxplot

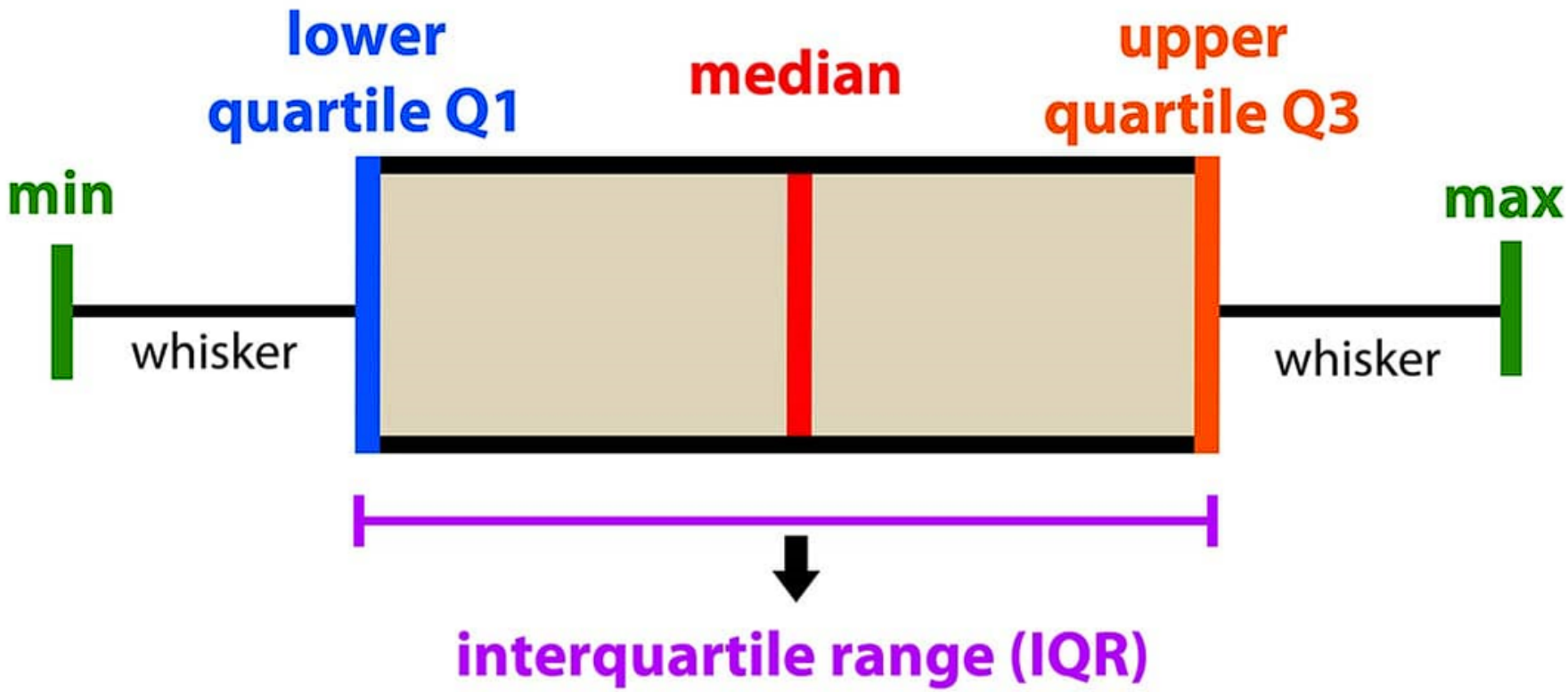
Continuous variables

Illustrates the spread of values

Median, percentiles, min/max

Outliers




lower
quartile Q1
median
upper
quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

Histogram

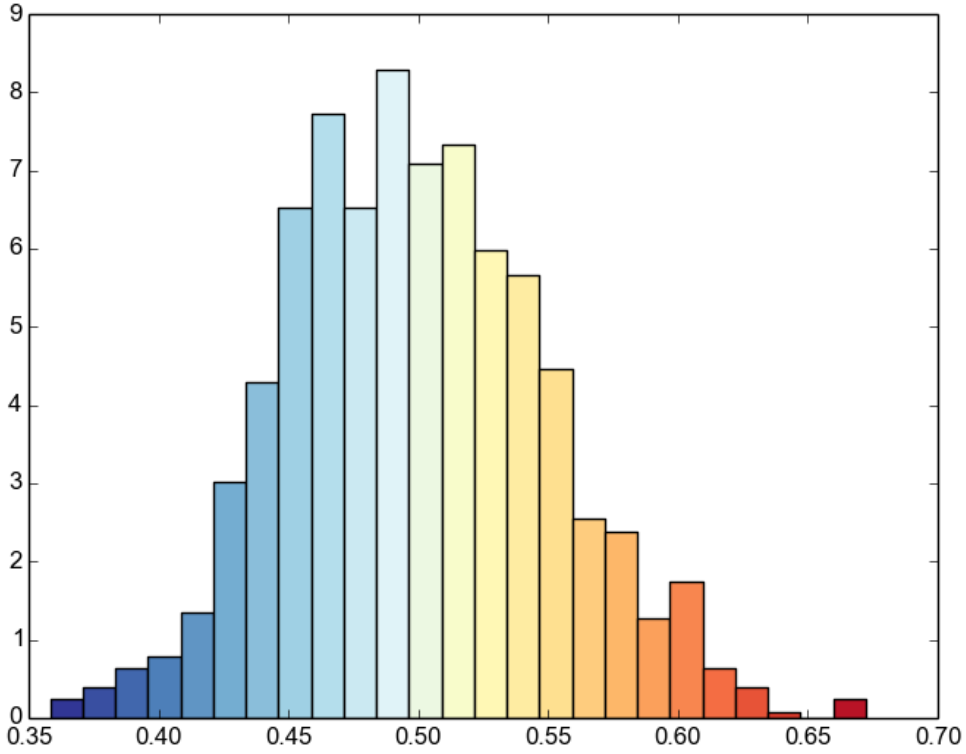
Continuous variables

Values are divided into bins = range of values

Visualizing the density

Bars in histogram vs bars in bar chart

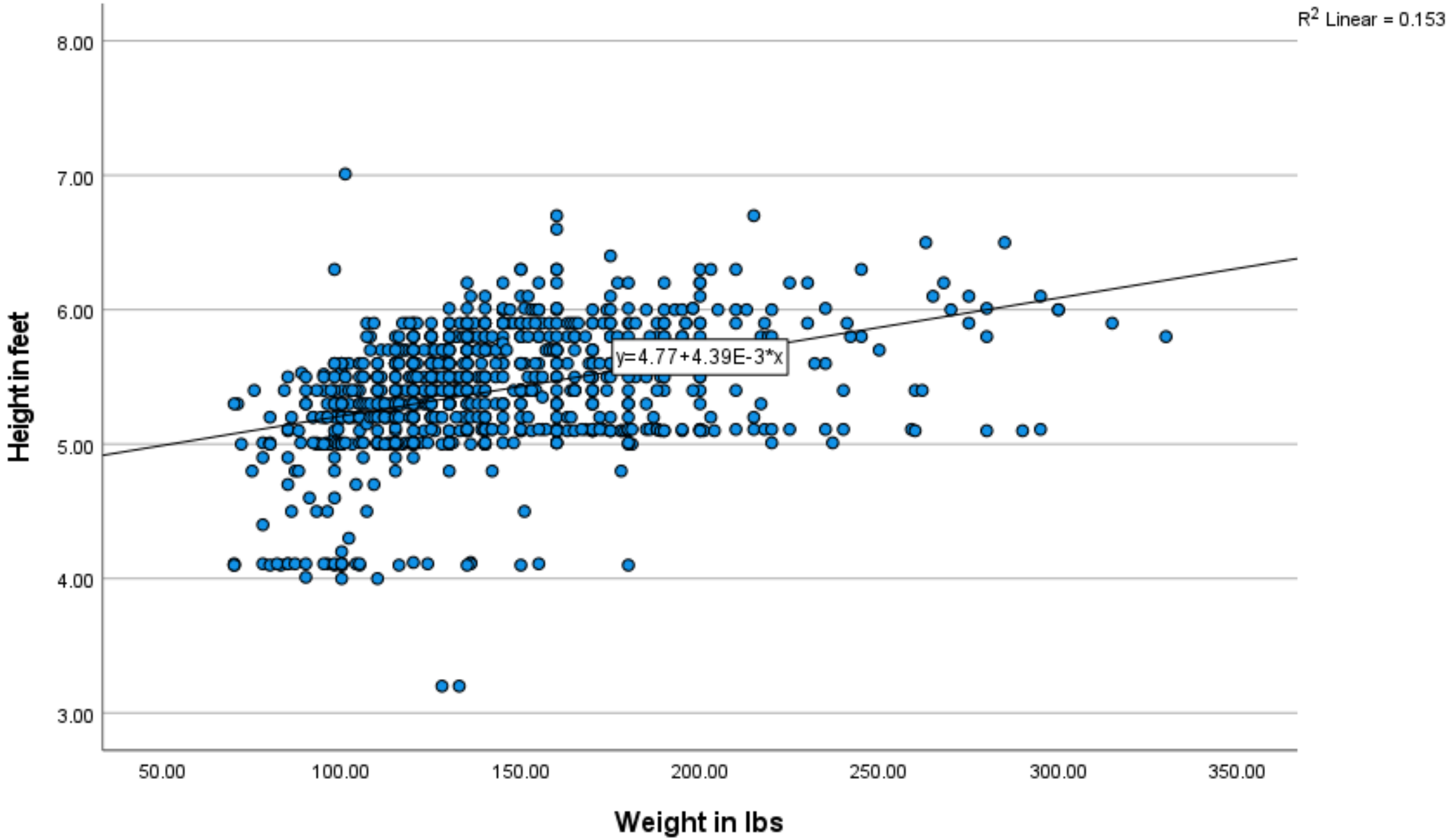
- Different range of values vs different categories



Scatterplot

Two continuous variables

Bivariate distribution


R2 Linear = 0.153
8.00
7.00
6.00
5.00
4.00
3.00
Height in feet
y=4.77+4.39E-3*x
50.00
100.00
150.00
200.00
250.00
300.00
350.00
Weight in lbs

KvIS 3

Data cleaning

Data often contain errors, missing values, outliers

This might be due to

- Contamination (biological samples)
- Error in data entry
- Just a really atypical case (with regards to outliers)

Outliers

atypical data point with regards to sample values

Example

- Erasmus students in class – 10 students

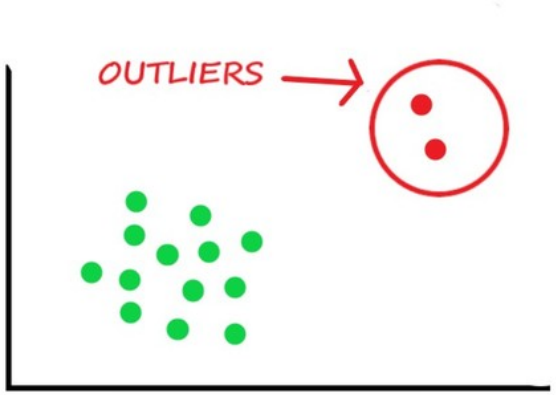
With outlier:

- M = 25.8
- SD = 15.9
- Median = 21

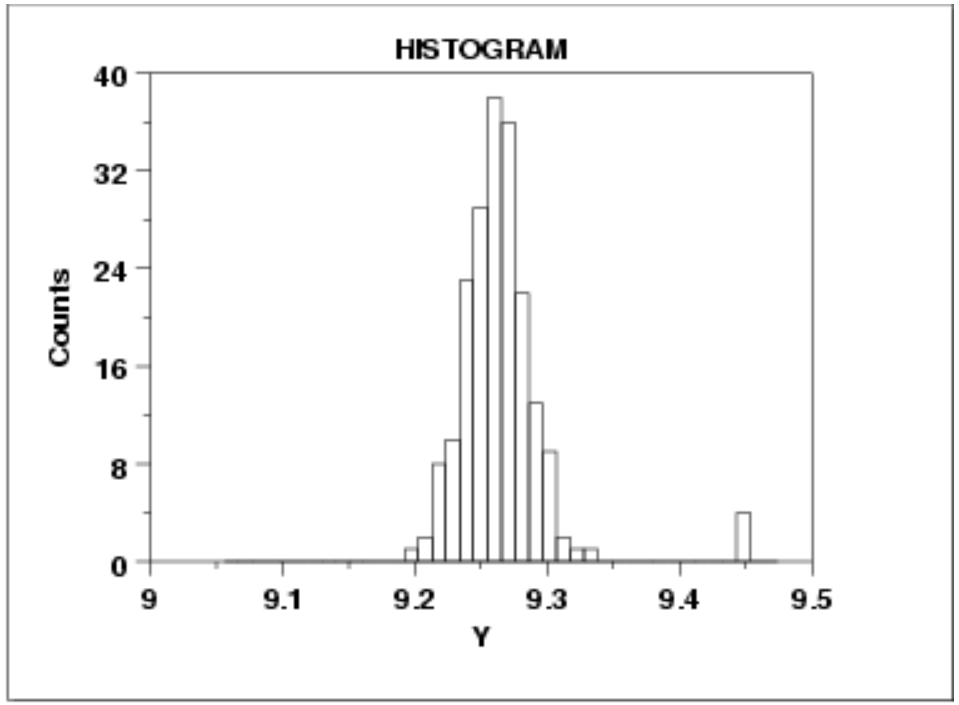
Without outlier:

- M = 20.8
- SD = 0.83
- Median = 21

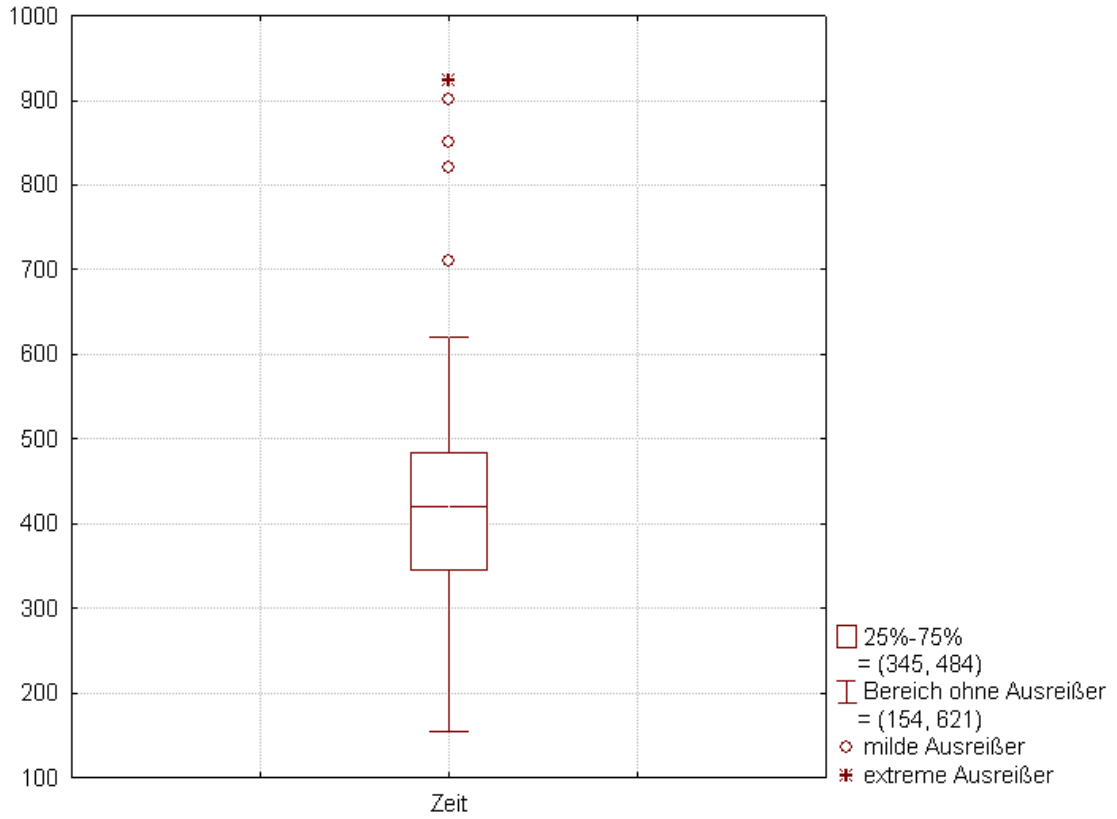
#	age
1	20
2	21
3	20
4	22
5	21
6	20
7	22
8	20
9	71
10	21



Identifying outliers – graphs



Box plot with 1.5 IQR = everything beyond that is outlier



Outliers – what should we do?

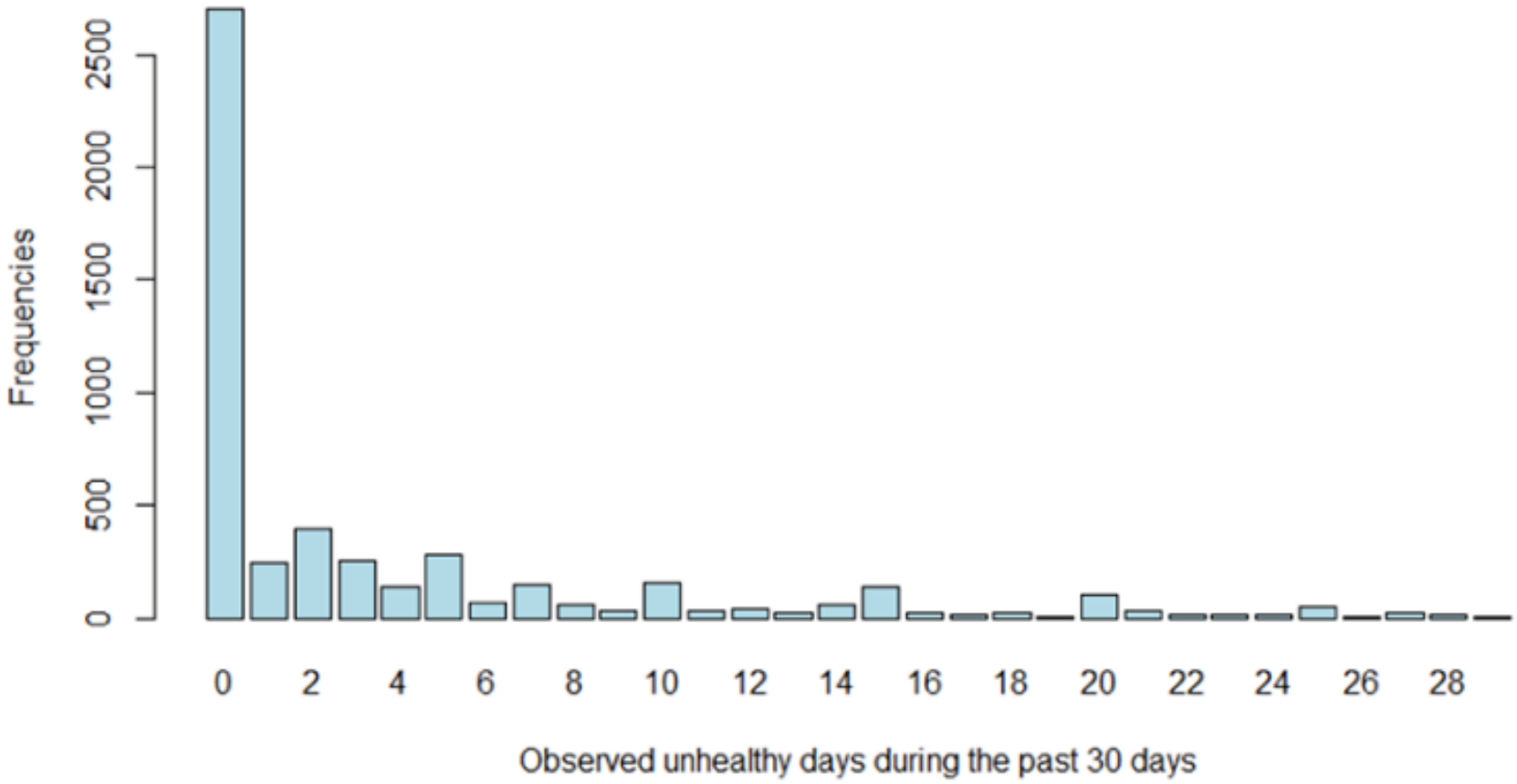
Errors in data entry – need to fix

Extreme values

- Remove?
- Keep in?
- Substitute?
- Transform?

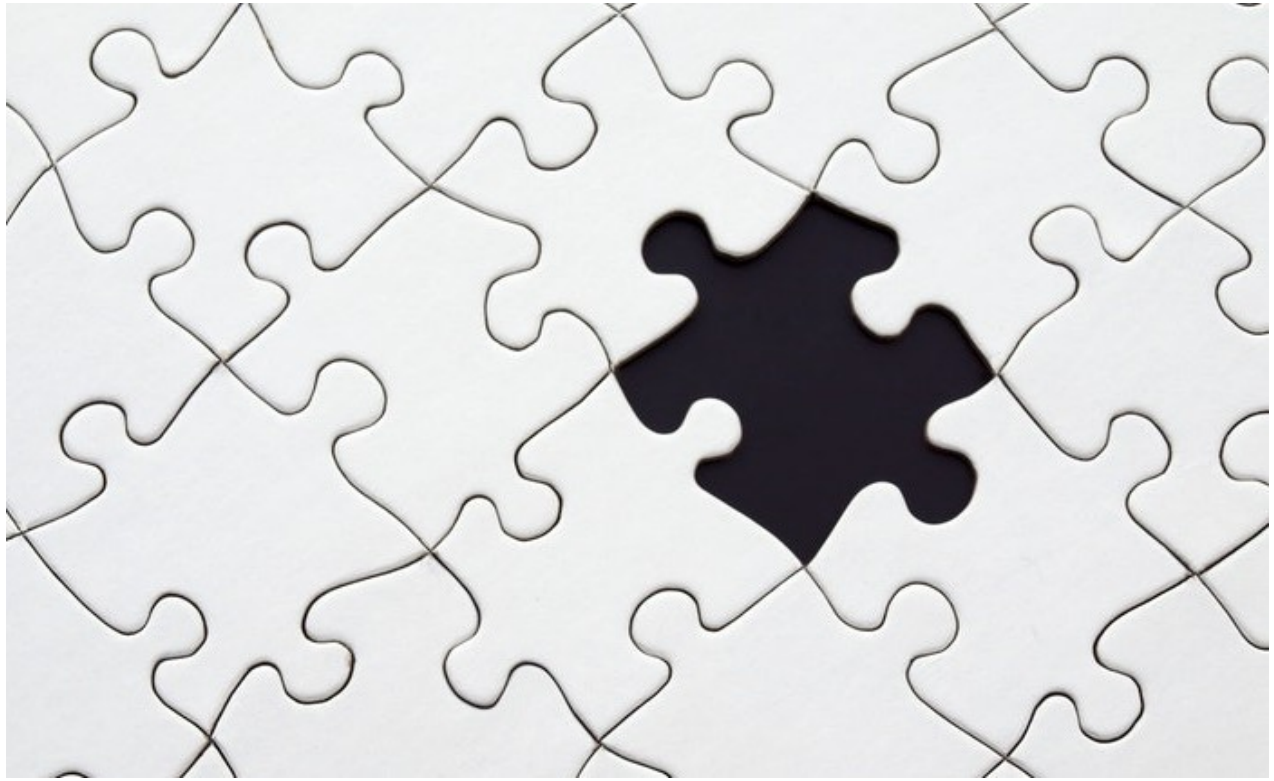
Depends on the type of data

Outliers?



Yang, S., Puggioni, G., Harlow, L. L., & Redding, C. A. (2017). A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys. *JMASM Editors*, *16*(1), 518-543.

Missing data



Missing data – no response for some or all of variables for an individual

- Missing data can lead to biased results

Handling missing data

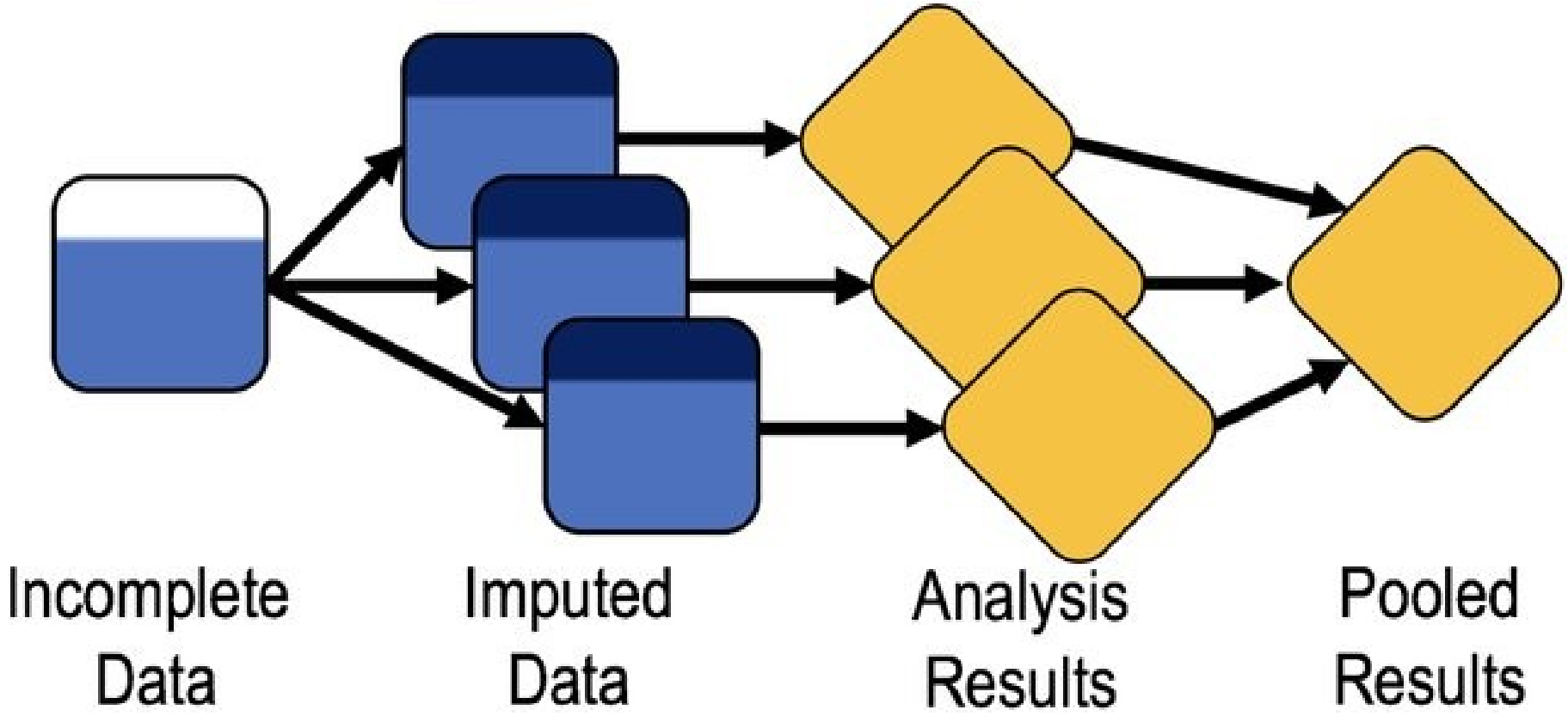
Poor handling:

- Listwise deletion
- Pairwise deletion
- Mean/median imputation

Good handling:

- Multiple imputation
- Full information maximum likelihood

Multiple imputation



Nissen, J., Donatello, R., & Van Dusen, B. (2019). Missing data and bias in physics education research: A case for using multiple imputation. Physical Review Physics Education Research, 15(2), 020106.