

# Biostatistics

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit  
University of Zurich

# Overview

- 1 Introduction
- 2 Univariate descriptive statistics
- 3 Probability theory
- 4 Hypothesis testing and confidence intervals
- 5 Correlation and linear regression
- 6 Logistic regression
- 7 Survival analysis
- 8 Analysis of variance

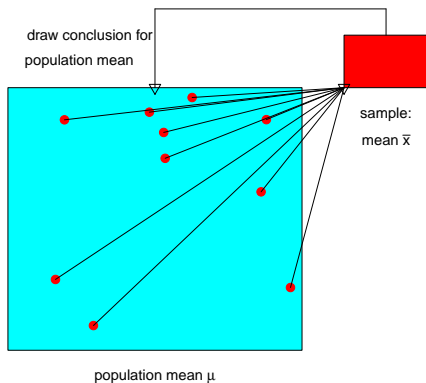
# Introduction

For which purpose does a medical biologist need statistics?

- in the own field of research
- study of literature
- consulting and support of the respective working group in quantitative methods

# Population and sample

- Data are based on one **sample**
- Data of different samples vary
- Conclusions are valid for a **population**



## Population and sample (II)

### Population

The population is the totality of all individuals for which conclusions should be made.

### Sample

A sample of a population is the set of individuals that are actually observed.

Example:

- Population = all human beings (all Swiss citizens)
- Sample = students of Medical Biology visiting this lecture

## Recommended literature

- Altman, D. G. (1991). *Practical statistics for medical research*. Chapman & Hall.  
- correct statistics, good examples, 600 pages.
- Armitage, P., Berry, G., and Matthews, J. N. S. (2002). *Statistical methods in medical research*. Blackwell, 4th edition.  
- comprehensive textbook, 817 pages.
- Bland, M. (1995). *An introduction to medical statistics*. Oxford Medical Publications.  
- very good introduction with many examples and exercises, 396 pages.
- Hüsler, J. and Zimmermann, H. (2006). *Statistische Prinzipien für medizinische Projekte*. Hans Huber, Bern.  
- clearly presented textbook, 355 pages.
- Johnson, R. A. and Bhattacharyya, G. K. (2001). *Statistics. Principles and methods*. Wiley, 4th edition.  
- light reading textbook, 236 pages.
- Kirkwood, B. R. and Sterne, J. A. C. (2006). *Essential Medical Statistics*. Blackwell, 4th edition.  
- extensive textbook, 502 pages.
- Matthews, D. E. and Farewell, V. T. (1988). *Using and understanding medical statistics*. Karger, 2nd edition.  
- offers logistic regression and survival analysis, 200 pages.
- Sachs, L. (2004). *Angewandte Statistik*. Springer, 11th edition.  
- in the German-speaking area the classical cook-book for all circumstances since 1968, no textbook, 890 pages.

# Biostatistics

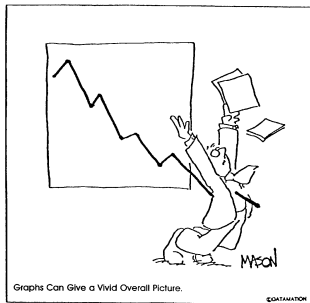
## Univariate descriptive statistics

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit  
University of Zurich

# Univariate descriptive statistics

- Approach “descriptive”, without “significance”
- Main types of data (scale types)
- Description of data
  - via tables
  - via graphics
  - via location- and dispersion statistics



## Data in a table

- In 2006, 245 students (16 groups) of the 2<sup>nd</sup> semester in medicine reported their body height and measured their hand length

sex	height	hand	group	tutor	gender
1	168.0	17.5	1	1	f
0	183.5	21.0	1	1	m
1	170.0	20.0	1	1	f
1	159.0	17.0	1	1	f
1	165.0	18.0	1	1	f
0	180.0	20.0	1	1	m
1	181.0	19.5	1	1	f
0	193.0	21.5	1	1	m
0	183.0	19.5	1	1	m
0	183.0	20.5	1	1	m
...	...	...	...	...	...

# Main types of data

## 1) **nominal**, categorical data

- Assignment to categories  
→ Counts and % meaningful  
Examples: Gender, blood type

sex	height	hand	group	tutor	gender
1	168.0	17.5	1	1	f
0	183.5	21.0	1	1	m
1	170.0	20.0	1	1	f
1	159.0	17.0	1	1	f
1	165.0	18.0	1	1	f
0	180.0	20.0	1	1	m
1	181.0	19.5	1	1	f
0	193.0	21.5	1	1	m
0	183.0	19.5	1	1	m
0	183.0	20.5	1	1	m
...	...	...	...	...	...

	Levels	Frequency	%	Cum. %
sex	m	106	43.3	43.3
	f	139	56.7	100.0
Total		245	100.0	

## 1-2) **ordinal** data (ordered categorical)

- have a ranking  
Example: Severity of a disease

# Describing data in tables and graphics

- Discrete data

$$\text{relative frequency} = \frac{\text{number of times an event occurred}}{\text{total number of events}}$$

**Example:** Proportion of blood types in a healthy population

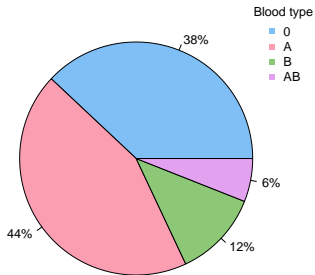
	Blood type	Frequency	Rel. frequency
Table	0	2313	38 %
	A	2678	44 %
	B	731	12 %
	AB	365	6 %
	Total	6087	100 %

Graphics are:

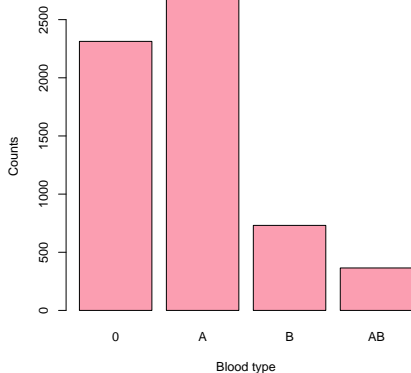
- easy to comprehend
- easy to create nowadays

# Graphics

## Pie chart

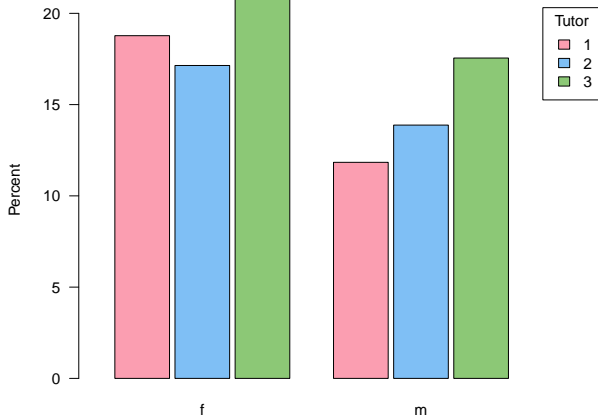


## Pareto or bar chart

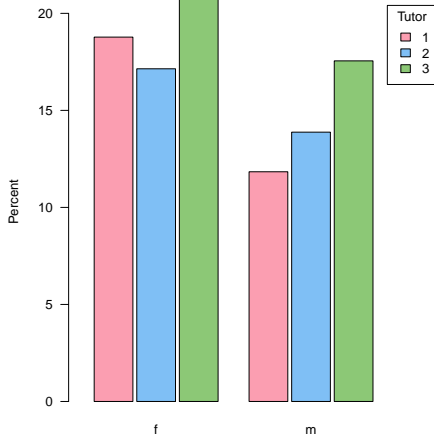


- Origin!

# Bar chart

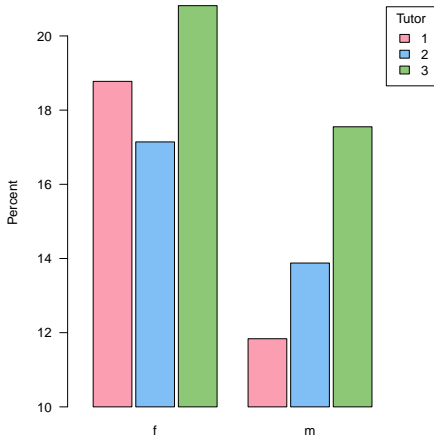


## Bar chart



- Don't trust a graphic which is higher than wide.

## Bar chart



- Don't trust a graphic which is higher than wide.
- Pay attention to the origin.

# Main types of data

## 2) continuous (numeric) data

- Differences and means meaningful

Example: Temperature in °C

- If a absolute zero point exists

→ Ratios meaningful

Examples: Temperature in K,  
body height, length of hand

sex	height	hand	group	tutor	gender
1	168.0	17.5	1	1	f
0	183.5	21.0	1	1	m
1	170.0	20.0	1	1	f
1	159.0	17.0	1	1	f
1	165.0	18.0	1	1	f
0	180.0	20.0	1	1	m
1	181.0	19.5	1	1	f
0	193.0	21.5	1	1	m
0	183.0	19.5	1	1	m
0	183.0	20.5	1	1	m
...	...	...	...	...	...

- Not meaningful: “There were times when the temperature was 60% higher than nowadays” *BBC 2006*

Now	Then
14 °C	22 °C
57 °F	91 °F = 33 °C
287 K	459 K = 186 °C

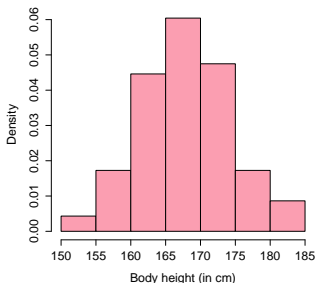
# Histogram

- Graphical visualisation of the data distribution, “data density”
- Continuous and ordinal data
- Group data into similar, non overlapping classes (intervals)

## Determine number of observations in interval

$$\text{Relative frequency in interval} = \frac{\text{number of observations in interval}}{\text{total number of observations}}$$

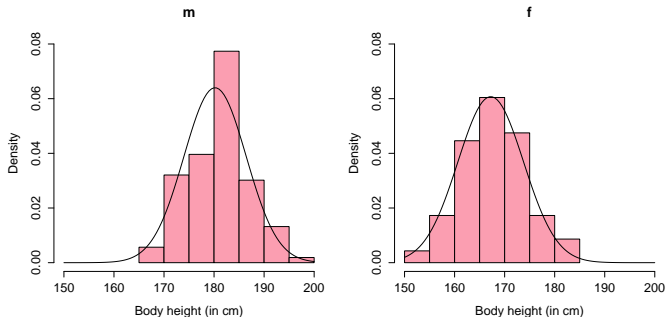
- Show relative (or absolute) frequencies of intervals in a bar chart



## Female body height ordered

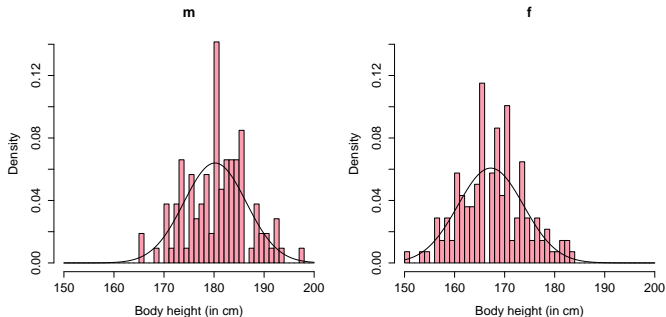
Interval	Height	n	# Observations	Relative frequency
150-154	150	1		
	153	1		
	154	1	3	2%
155-159	156	3		
	156.5	1		
	157	2		
	158	4		
	159	2	12	9%
160-164	160	8		
	161	6		
	162	5		
	163	5		
	164	7	31	22%
165-169	165	16		
	167	8		
	168	12		
	169	6	42	30%
170-174	170	14		
	171	2		
	172	4		
	173	9		
	174	4	33	24%
175-179	175	2		
	176	4		
	177	2		
	178	3		
	179	1	12	9%
180-184	180	1		
	181	2		
	182	2		
	183	1	6	4%
Total			139	100%

# Histogram



- Shows the distribution in the sample
- Meaningful interval length: 5 cm
- Fitted a “Gaussian normal distribution” to distribution in population

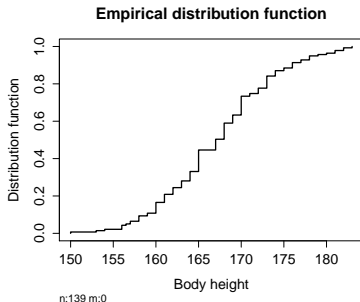
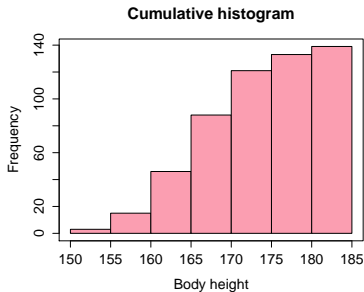
# Histogram



- Interval length: 1 cm (very variable)
- Statement depends mainly on **bin width** and slightly on **center**
- Histograms are simple and popular, but there are better density estimators

# Cumulative histogram

A cumulative histogram estimates the distribution function



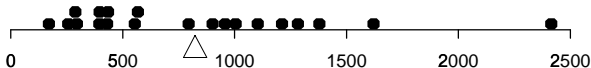
## Characterization of the centre of the data

- What is a typical, mean value?

**Mean**  $\bar{x}$ : measure of the “middle” (mean, average) value

$$\bar{x} = (x_1 + x_2 + \dots + x_n)/n$$

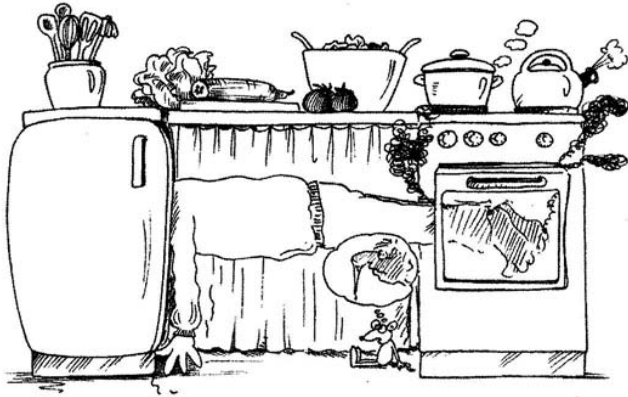
The mean is the value which balances the data on a set of scales.



With **normally distributed data** the mean in a sample is the best fit to the mean in the population.

- sensitive to outliers

## Dispersion or variation of a sample



*A statistician is a person who, if you've got your feet in the oven and your head in the refrigerator, will tell you that, on average, you're very comfortable.*

# Dispersion or variation of a sample

- How dispersed are the data around the mean position?

## Variance $s^2$ :

Compute deviations  $(x_1 - \bar{x}), \dots, (x_n - \bar{x})$

Mean? No - would result to be 0!

$$\Rightarrow s^2 = \{(x_1 - \bar{x})^2, \dots, (x_n - \bar{x})^2\} / (n - 1)$$

- Note:  $s^2$  in squared units (e. g.  $\text{cm}^2$ )

Standard deviation (SD):  $s = \sqrt{\text{variance}}$  (in cm)

For **normally distributed data** are 68% of the data in the interval mean  $\pm$  SD, 95% of the data in the interval mean  $\pm$  2 SD.

- sensitive to outliers
- no interpretation for non-normally distributed data



A statistician is a person who, if you've got your feet in the oven and your head in the refrigerator, will tell you that, on average, you're very comfortable.

## Descriptive statistics

- Data are often represented by the mean plus-minus the standard deviation (mean  $\pm$  SD).
- R-output summary():

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
f	150.0	163.0	167.0	167.2	171.5	183.0
m	165.0	176.0	180.0	180.2	184.0	197.0

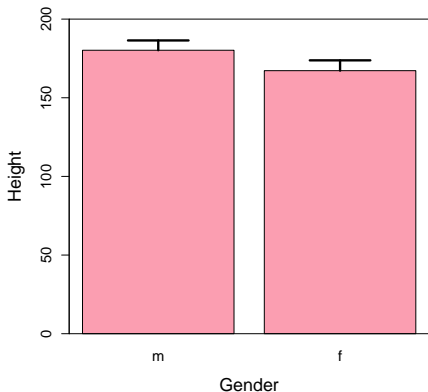
- R-output tableContinuos() (“reporttools”, v.1.0.4):

Gender	N	Min	Q1	Median	Mean	Q3	Max	SD	IQR	#NA
f	139	150	163	167	167.2	171.5	183	6.6	8.5	0
m	106	165	176	180	180.2	184.0	197	6.2	8.0	0

### Mean $\pm$ SD or Mean $\pm$ SEM ?

- The **standard error** of the mean (SEM) is the standard deviation of the mean:  $SEM = SD/\sqrt{n}$ .  
In descriptive statistics the SEM should not be used!

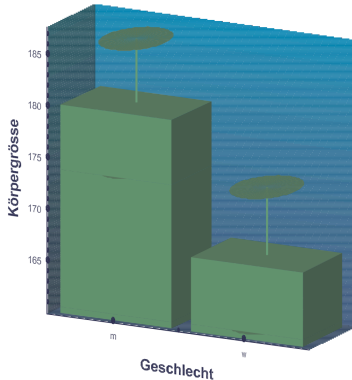
## Bar chart



Error bars show mean  $\pm$  1.0 SD  
Bars show means

- Bars stand on the floor, therefore pay attention to the origin
- Take care of 3-dimensional graphics

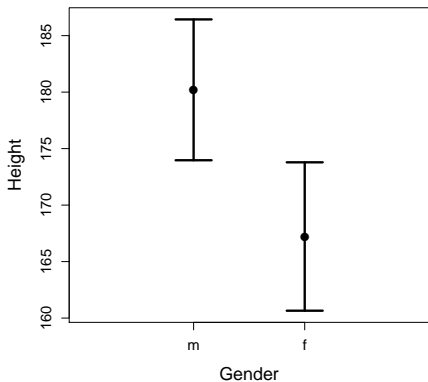
# Bar chart



Error Bars show Mean  $\pm$  1.0 SD

- Bars stand on the floor, therefore pay attention to the origin
- Take care of 3-dimensional graphics

## Dot charts



Error bars show mean  $\pm$  1.0 SD  
Dots show means

- The origin has no meaning here

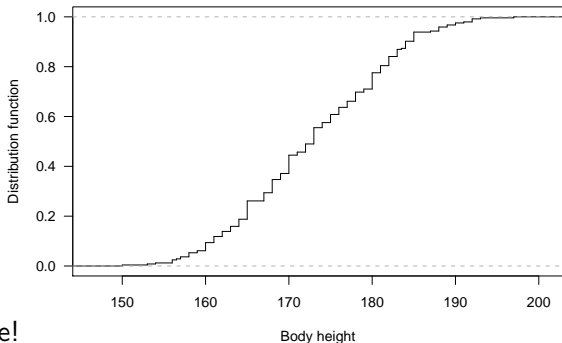
# Percentiles (quantiles)

$\alpha$ .- percentile ( $\alpha\%$  - quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$ . - percentile and  $(100 - \alpha)\%$  are larger or equal.

**Examples:** ● **Median** = 50. percentile

● **Quartile** = 25. and 75. percentiles



Not unique!

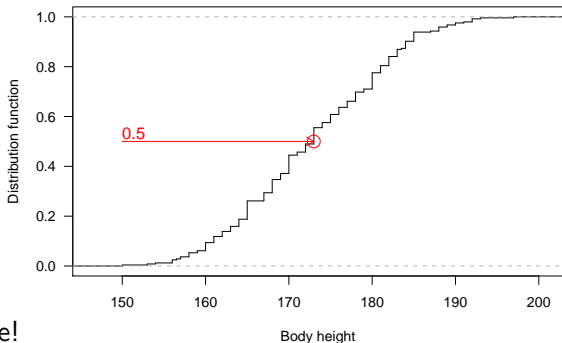
In R there are nine different quantile algorithms.

# Percentiles (quantiles)

$\alpha$ .- percentile ( $\alpha\%$  – quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$ . – percentile and  $(100 - \alpha)\%$  are larger or equal.

- Examples:**
- **Median** = 50. percentile
  - **Quartile** = 25. and 75. percentiles



Not unique!

In R there are nine different quantile algorithms.

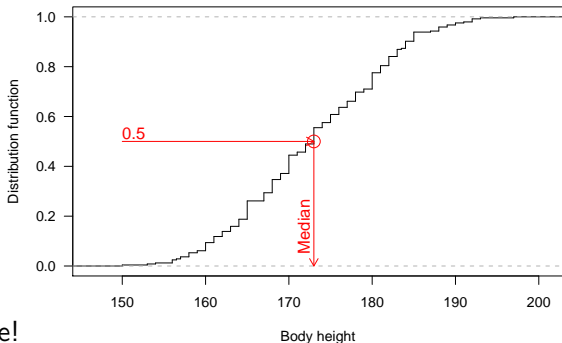
# Percentiles (quantiles)

$\alpha$ .- percentile ( $\alpha\%$  - quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$ . - percentile and  $(100 - \alpha)\%$  are larger or equal.

**Examples:** ● **Median** = 50. percentile

● **Quartile** = 25. and 75. percentiles



Not unique!

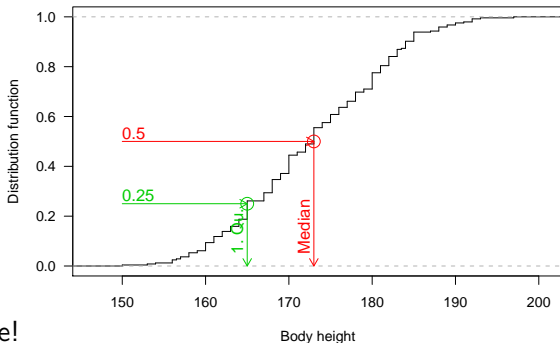
In R there are nine different quantile algorithms.

# Percentiles (quantiles)

$\alpha$ -percentile ( $\alpha\%$  – quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$  – percentile and  $(100 - \alpha)\%$  are larger or equal.

- Examples:**
- **Median** = 50. percentile
  - **Quartile** = 25. and 75. percentiles



Not unique!

In R there are nine different quantile algorithms.

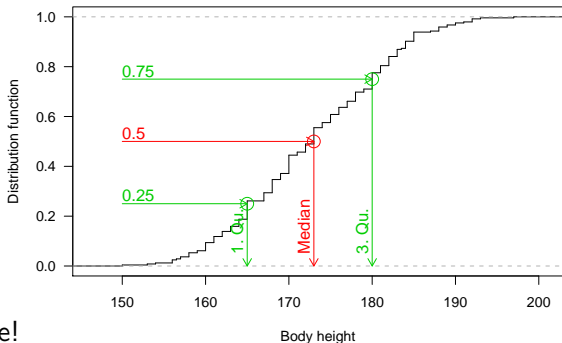
# Percentiles (quantiles)

$\alpha$ .- percentile ( $\alpha\%$  – quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$ . – percentile and  $(100 - \alpha)\%$  are larger or equal.

**Examples:** ● **Median** = 50. percentile

● **Quartile** = 25. and 75. percentiles



Not unique!

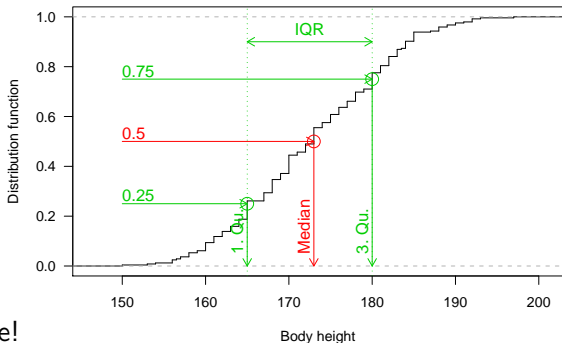
In R there are nine different quantile algorithms.

# Percentiles (quantiles)

$\alpha$ .- percentile ( $\alpha\%$  - quantile):

$\alpha\%$  of the data are smaller than or equal to the  $\alpha$ . - percentile and  $(100 - \alpha)\%$  are larger or equal.

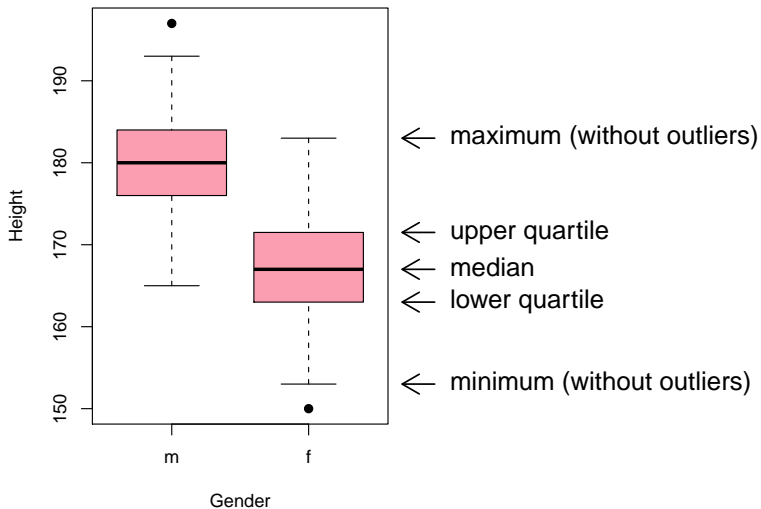
- Examples:**
- **Median** = 50. percentile
  - **Quartile** = 25. and 75. percentiles



Not unique!

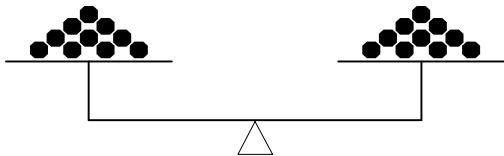
In R there are nine different quantile algorithms.

# Boxplot



## Characterization of the centre of the data

- **Median:** centre of the data, 50. percentile  
i.e. half of the sample is above the median, the other half below



The median is robust to outliers.

- **Mode:** (rarely used)
  - discrete data: most frequent value
  - continuous data: maximum of the density (population only)

# Dispersion of a sample

- **Range** = maximum – minimum
  - states the range of all values in the sample
  - strongly influenced by outliers
  - but: Minimum and maximum are easy to understand
- **Interquartile range (IQR)**
  - = 75. percentile – 25. percentile
  - = length of box in the boxplot, contains central 50% of data
  - as standard deviation a measure for the magnitude of the central range of the data

With normally distributed data half the IQR equals 0.67 SD.

- “Median(IQR)” tells nothing about skewness  
⇒ Data are often reported as  
“Median [lower quartile, upper quartile]”.