

# Elements of Statistics

## Chapter 1: Introduction to statistics

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# What is Statistics? (1)

**status** (lat.): State, Status

**statisticum** [Latin]: Concerning the state

**statista** [Italian]: Statesman

→ **Statistik** [German]: Science of state affairs

**Oxford English Dictionary:**

*Material* sense (statistic): A statistical fact, statement, or piece of data

*Instrumental* sense (statistics): The systematic collection and arrangement of numerical facts or data of any kind; (also) the branch of science or mathematics concerned with the analysis and interpretation of numerical data and appropriate ways of gathering such data

# What is Statistics? (2)

Statistics can be simple counting and measuring.

→ Presentation of the results in tables and graphics

Statistics is a scientific discipline which uses methods for description and analysis of aggregates through numbers.

→ Application of adequate methods in the applications

Statistics does not examine a specific field of human experience, but it is a methodical instrument for all sciences like economics, social sciences, medicine, natural sciences, etc.

# The German Census

- ▶ Census of a population (inhabitants of Germany)
- ▶ Last *Volkszählung* in West Germany: 25.05.1987 (East: 31.12.1981)
- ▶ A census should be carried out approximately every 10 years.
- ▶ Census program:
  - ▶ Size of the population (State, Federal States, Districts, Communities)
  - ▶ Distribution of gender and age
  - ▶ Professions
  - ▶ Household structure
  - ▶ Housing

→ Current population statistics

→ Communal fiscal adjustment

Register based census in Germany 2021:

Methodological research by Trier University (Münnich) and  
GESIS Leibniz-Institut für Sozialwissenschaften in Mannheim (Gabler);  
see: <https://www.uni-trier.de/index.php?id=57986>

# Further Examples for Statistics

- ▶ Inventory
  - ▶ Aim: Counting of items in, e.g., a warehouse
  - ▶ Methods: Complete inventory count vs. inventory sampling procedure
- ▶ National accounts
- ▶ Price indices and price changes
- ▶ Demographic development
  - ▶ Birth rate
  - ▶ Death rate
  - ▶ Migration
- ▶ Measurement of the progress in attaining EU targets  
(e.g. Lisbon Process) ⇒ need for *adequate* indicators

Stocks vs. flows

# The German Microcensus

- ▶ Classification of sample units
  - ▶ 214 regional strata (RS)
  - ▶ 5 house size classes (GGK)
  - ▶ Clusters of approx. 20 persons by RS x GGK:  
Sampling District (AWB)
- ▶ Sampling design
  - ▶ Pooling of 100 clusters each to one zone
  - ▶ Selection of one cluster per zone
  - ▶ ≈ 1% of persons/households
- ▶ Since 2005 during the year, before annually
- ▶ Basis for the Access-Panel
  - ▶ European Union Statistics on Income and Living Conditions (EU-SILC)
  - ▶ Survey about the use of information technologies (IKT)

# Unemployed in the Microcensus campus file

```
H <- with(subset(Y, Unemployed==1),  
table(Gender, AgeGroup)  
)  
H <- cbind(H, total=apply(H,1,sum))  
H <- rbind(H, total=apply(H,2,sum))
```

| [0;15) [15;25) [25;45) [45;65) at least 65 | total

---

male		0	124	288	253		1		666
female		0	66	235	222		1		524
total		0	190	523	475		2		1190

- ▶ Data taken from **Microcensus** 2002 campus file  
(Source: <http://www.forschungsdatenzentrum.de>)
- ▶ Subsequently: Projection of sample to federal level
- ▶ How high is the quality of the projections?

# Data quality: Eurostat definition (1)

## Relevance of statistical concepts

User needs, concepts and granularity

## Timeliness and punctuality

Time lag between data release and phenomenon and  
time lag between actual and targeted data release

## Accessability and clarity

Data (format) availability and documentation, visualisation etc.

## Completeness

## Coherence

- ▶ Preliminary and final statistics
- ▶ Yearly data and higher frequency data
- ▶ Definitions

# Data quality: Eurostat definition (2)

## Comparability

Geographical areas, domains, time

## Accuracy of estimates

- ▶ Sampling error:  
Standard error, confidence interval
- ▶ Non-sampling error:  
Non-response, coverage error, measurement error, model error

## Statistical production process

- ▶ Input — Transformation — Output (and meta data)
- ▶ Problem of statistical adequation
- ▶ Burden of respondents and costs incurred by data producer

# Population vs. sample

## Population

The set of all elements relevant to a specific research question is called population. Therefore, a population is a set of elements which share a certain system of defining characteristics relevant to the research question.

## Delimitation of a population

- ▶ Time  
(populace at point in time vs. births during period of time)
- ▶ Geography  
(GDP vs. GNP)
- ▶ Fact  
(resident population vs. expatriates)

## Sample

A subset of or selection from the population is called sample, where the sample can be **random** or **non-random**.

# Scaling of variables (1)

Aim: Specification of a research question

We have to distinguish

- ▶ statistical units from
- ▶ their attributes or statistical variables and those in turn from their
- ▶ values or realisations.

The definition of the set of possible values of a variable is called its scaling.

Statistical variables can be scaled differently and at different levels of intensity. Some examples:

- ▶ Marital status
- ▶ Social stratum
- ▶ Age
- ▶ Income

## Scaling of variables (2)

### Nominal variables

A variable is called nominal if its values only express equality or inequality but no order, no distance and no ratio.

Examples: Marital status, citizenship, soil colour

### Ordinal variables

A variable is called ordinal if its values express inequality as well as a certain order, but neither a distance nor a ratio. Assigned numbers are supposed to express the order of the values.

Examples: Ranks, risk preference, water quality

## Scaling of variables (3)

### Interval-scaled variables

A variable is called interval-scaled if its values express inequality, an order as well as a distance or difference, but no ratio.

Examples: Speeding, *historical* year, temperature ( $^{\circ}\text{C}$ )

### Ratio-scaled variables

A variable is called ratio-scaled if its values express an inequality, an order, a distance or difference as well as a ratio.

Examples: Income, duration, global radiation

Interval-scaled and ratio-scaled variables are both called **metric variables**.

# Intensity and measurement problems (metric variables)

**Discrete:** At most a countable number of values

- ▶ Finite
- ▶ Countably infinite

**Continuous:** An uncountable number of values

Problem of measurement accuracy: Age, weight, income, etc.

# Descriptive and inferential statistics

## Descriptive statistics

Collected data are typically processed (for better interpretation) or *condensed* (e.g. figures, tables, statement of *typical* values) in an appropriate manner. In statistics, this work is known as description; the methods used here are summarised under the term descriptive statistics.

## Inferential statistics

If data are sampled, descriptions only pertain to the sampled elements of the population at first. Nevertheless, typically conclusions about the population as a whole are desired. The transfer of sample results to the population is called statistical inference or inferential statistics. When dealing with sampled data, inferential statistics have to complement the description of the sampled population subset.

# Statistical institutions and types of statistics

## Non-official statistics

Businesses, research institutes, organisations

## Official statistics

National statistical institutes (NSIs), other public-law institutions,  
Eurostat, European Central Bank (ECB), United Nations (UN)

## Primary statistics

Statistics based on data **collected for the specific task** at hand

## Secondary statistics

Statistics based on data originally **collected for other purposes**  
(e.g. tax statistics)

# Elements of Statistics

## Chapter 2: Methods of data collection and visualisation

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# Eurostat database

Indicators for:

- ▶ Key indicators on EU policy
- ▶ General and regional statistics
- ▶ Economics and finance
- ▶ Population and social conditions
- ▶ Industry, trade and services
- ▶ Agriculture, forestry and fisheries
- ▶ International trade
- ▶ Transportation
- ▶ Environment and Energy
- ▶ Science, technology, digital technology

<https://ec.europa.eu/eurostat/data/database>

# European Social Survey (ESS)

- ▶ Introduced in 2001
- ▶ International **Survey** in Europe (more than 30 countries)
- ▶ *Academically driven*
- ▶ Surveyed every 2 years
- ▶ Measures the attitudes, beliefs and behaviour patterns
- ▶ Chart stability and change in social structure
- ▶ Introduce soundly-based indicators

Homepage:

<https://www.europeansocialsurvey.org/>

Documents and data files for German survey:

[https:](https://www.europeansocialsurvey.org/data/country.html?c=germany)

[//www.europeansocialsurvey.org/data/country.html?c=germany](https://www.europeansocialsurvey.org/data/country.html?c=germany)

# Eurosystem Household Finance and Consumption Survey (HFCS)

- ▶ Initiated in 2006
- ▶ Currently 19 countries of the Eurozone plus Croatia, Hungary and Poland
- ▶ Initiative of the European Central Bank (ECB)
- ▶ Approx. surveyed every 3 years
- ▶ Data on households:
  - ▶ Real assets and their financing
  - ▶ Liabilities
  - ▶ Income
  - ▶ Consumption
  - ▶ Socio economic information
  - ▶ socio demographic information
- ▶ High relevance for monetary and fiscal policy (see financial crisis)

[http://www.ecb.int/home/html/researcher\\_hfcn.en.html](http://www.ecb.int/home/html/researcher_hfcn.en.html)

# Terminology

POP	Population
$n$	Number of units to be analysed; enumerated by $i = 1, \dots, n$ (later on: $N$ for POP and $n$ for sample)
$m$	Number of categories or ranks
$x_i$	Value of variable $X$ for $i$ -th unit ( $i = 1, \dots, n$ )
$n_j$	Absolute frequency of $j$ -th category or $j$ -th rank ( $j = 1, \dots, m$ )
$p_j$	Relative frequency of $j$ -th category or $j$ -th rank ( $j = 1, \dots, m$ )

# Frequency distribution

The frequency distribution of a variable summarises its categories or ranks and the related frequencies. It can be determined in an absolute or relative sense and is presented in a frequency table.

We have

$$\sum_{j=1}^m n_j = n \quad \text{and} \quad \sum_{j=1}^m p_j = \sum_{j=1}^m \frac{n_j}{n} = 1.$$

## Example 2.1: Unemployment (1)

In an attempt to estimate the number of unemployed people in Germany, we first analyse the one-dimensional (univariate) frequency distributions.

- ▶ Distribution by gender:

male	female	$\sum$
666	524	1,190

$$m = 2$$

Realisation in R:

```
setwd("path") # Choose your working directory on your own.  
load("Example2-1.RData")  
table(Unemployment$Gender)
```

```
Male   Female  
666      524
```

```
length(Unemployment$Gender)
```

```
[1] 1190
```

## Example 2.2: Fields of study (1)

Students in the course *Mathematical Statistics* were asked about their field of study. The following original list emerged:

ECO, ECO, ECO, BA, SOC, MAT, BMAT, BMAT, ECO, ECO, ECO,  
SOC, SOC, ECO, ECO, SOC, MAT, CS, ECO, ECO.

Field of study	$j$	Tally	$n_j$	$p_j$ (in %)
Business administration	1		1	5
Economics	2		10	50
Sociology	3		4	20
Mathematics	4		2	10
Business mathematics	5		2	10
Computer sciences	6		1	5
			20	100

## Example 2.2: Fields of study (2)

Determination of frequency tables in R:

```
x2_2 <- factor(c("ECO", "ECO", "ECO", "BA", "SOC", "MAT",
                  "BMAT", "BMAT", "ECO", "ECO", "ECO", "SOC",
                  "SOC", "ECO", "ECO", "SOC", "MAT", "CS",
                  "ECO", "ECO"),
                  levels = c("BA", "ECO", "SOC", "MAT", "BMAT", "CS"))
```

```
)
```

```
n_j <- table(x2_2)
```

```
p_j <- prop.table(n_j)
```

```
n_j
```

```
x2_2
```

	BA	ECO	SOC	MAT	BMAT	CS
1	10	4	2	2	1	

```
p_j
```

```
x2_2
```

	BA	ECO	SOC	MAT	BMAT	CS
0.05	0.05	0.50	0.20	0.10	0.10	0.05

```
p_j * 100
```

```
x2_2
```

	BWL	VWL	SOZ	MAT	WIM	INF
5	5	50	20	10	10	5

# Grouping of metric variables

- ▶ Variables with few unique values can be treated as before (e.g. number of children in household)
- ▶ Variables with many unique values need grouping (e.g. income)

→ Splitting of range  $[x_0^o; x_m^o]$  into  $m$  classes:

$x_j^o$                       Upper boundary of  $j$ -th class

$x_0^o$                       Lower boundary of first class

$x'_j = \frac{1}{2}(x_j^o + x_{j-1}^o)$       Mid of  $j$ -th class

$n_j(p_j)$                       Number (share) of observations in  $j$ -th class

The  $i$ -th observation  $x_i$  falls into the  $j$ -th class if  $x_{j-1}^o \leq x_i < x_j^o$ .

## Example 2.3: compare Example 2.1 (1)

- Distribution of unemployed by age class:

[0; 15)	[15; 25)	[25; 45)	[45; 65)	$\geq 65$	$\sum$
0	190	523	475	2	1.190

Illustration of absolute frequency of classes  $m = 2$  (Ex. 2.1) resp.  $m = 5$  (here).

Application in R:

```
x_o      <- c(0, 15, 25, 45, 65, Inf)
age_class <- cut(x = Unemployment$Age, breaks = x_o,
                  right = FALSE)
table(age_class)
length(age_class)
```

```
age_class
[0,15) [15,25) [25,45) [45,65) [65,Inf)
      0       190     523     475       2
```

```
length(age_class)
```

```
[1] 1190
```

## Example 2.3: compare Example 2.1 (2)

Further analysis of the unemployment dataset in R:

```
attach(Unemployment)
summary(Income)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
247.0	564.9	644.4	642.9	723.1	946.3

- ▶ What would be the result for other surveys?
- ▶ Which influence do *cut* incomes have?
- ▶ How exact are the observations?

```
x_o_new <- c(200, 350, 500, 650, 800, Inf)
income_class <- cut(x = Income, breaks = x_o_new,
                     right = FALSE
)
summary(income_class)
[200,350) [350,500) [500,650) [650,800) [800,Inf)
      5        140       480       451       114
```

# Some visualisation tools

- ▶ Horizontal bar chart
  - ▶ Absolute
  - ▶ Relative
- ▶ Vertical bar chart
  - ▶ Absolute
  - ▶ Relative
- ▶ Pie chart
- ▶ Spider plot
- ▶ Histogram
- ▶ Sum function

You should always consider the scaling of the data about to be visualised!

## Example 2.4: German construction activity (1)

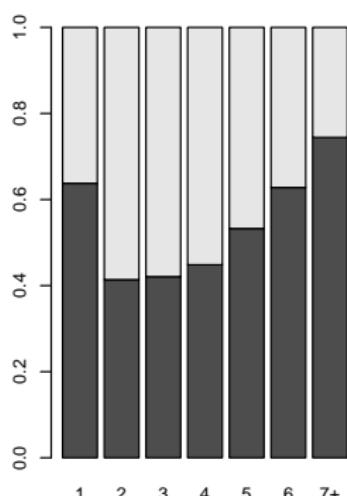
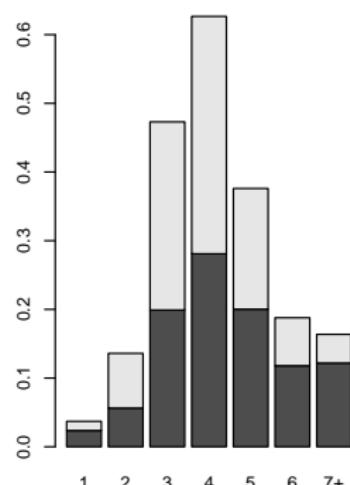
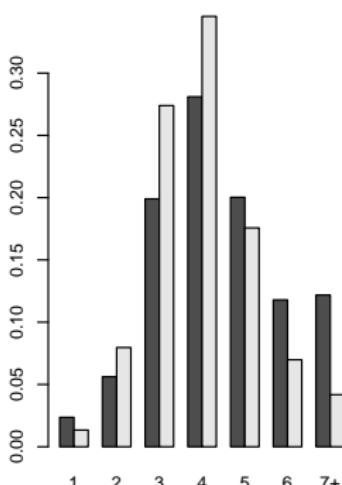
Object	Units	West	East	Total
West: former federal territory		East: new federal states and Berlin		
<b>Housing stock 2011 (available figures)</b>				
Flats	1 000	31 585.2	8 888.7	40 473.8
with ... rooms				
1	1 000	738.7	120.0	858.7
2	1 000	1 784.9	698.5	2 483.5
3	1 000	6 222.7	2 405.7	8 628.4
4	1 000	8 765.1	3 043.7	11 808.8
5	1 000	6 341.9	1 585.1	7 927.1
6	1 000	3 797.6	644.3	4 442.0
7 and more	1 000	3 934.2	391.3	4 325.5
Rooms in total	1 000	143 321.5	35 686.2	179 007.6
Difference to 2004	1 000	718.0	71.3	789.2
Living space in total	million sqm	2 862.1	654.1	3 516.2

Source (15/10/2012): See Bautätigkeit, Wohnungsbestand on

<https://www.destatis.de/DE/ZahlenFakten/Wirtschaftsbereiche/Bauen/Bautaetigkeit/Bautaetigkeit.html>

## Example 2.4: German construction activity (2)

Graphical presentation with bar charts (see table):



```
round(App_pj, digits = 4) # See next slide!
```

	1	2	3	4	5	6	7+
West	0.0234	0.0565	0.1970	0.2775	0.2008	0.1202	0.1246
East	0.0135	0.0786	0.2707	0.3424	0.1783	0.0725	0.0440

## Example 2.4: German construction activity (3)

Bar chart of the previous slide in R:

```
load("Example2-4.RData")

App_pj <- t(apply(Housing[, 2:3], MARGIN = 2,
                  FUN = prop.table))
colnames(App_pj) <- Housing$Number_of_rooms

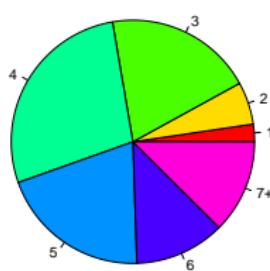
par(mfrow = c(1, 3))
barplot(App_pj, beside = TRUE)
barplot(App_pj)
barplot(apply(App_pj, 2, prop.table))
par(mfrow = c(1, 1))
```

You can create vertical bar charts by setting the argument `horiz` of the `barplot` function to `TRUE`.

## Example 2.4: German construction activity (4)

The intention here is to illustrate the number of flats subject to the number of rooms per flat and compare the relevant figures for Western and Eastern Germany (see table above).

Western Germany



Eastern Germany



## Example 2.4: German construction activity (5)

Pie charts of the previous slide in R:

```
sum_west <- sum(Housing$West)
sum_east <- sum(Housing$East)
radius_west <- min(1, sqrt(sum_west / sum_east))
radius_east <- min(1, sqrt(sum_east / sum_west))

par(mfrow = c(1, 2))
pie(Housing$West,
    col = rainbow(n = 7),
    radius = radius_west,
    labels = Housing$Number_of_rooms
)
title("Western Germany", line = -4)
pie(Housing$East,
    col = rainbow(n = 7),
    radius = radius_east,
    labels = Housing$Number_of_rooms
)
title("Eastern Germany", line = -4)
par(mfrow = c(1, 1))
```

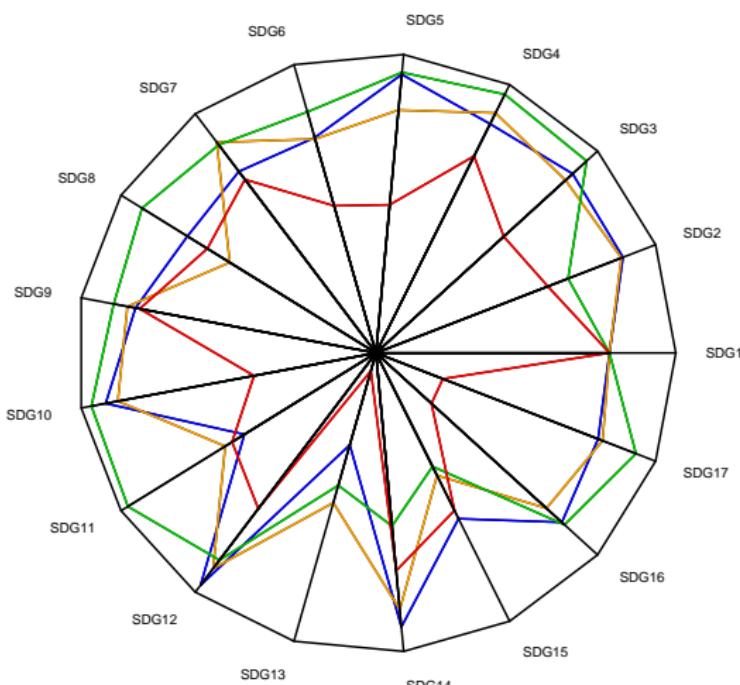
# Sustainable Development Goals (SDG)



Siehe Sustainable Development Goals Knowledge Platform:

<https://sustainabledevelopment.un.org/?menu=1300>

# Spider plot



Legend: GER, FRA, NOR, RUS

# Grouping of metric variables

- ▶ Number of classes
  - ▶ 5 – 20
  - ▶ Rule of thumb:  $\sqrt{n}$   
(problematic for official statistics; e.g. micro census or census)
- ▶ Location of classes
  - ▶ A higher denseness of observations should lead to narrower class intervals
  - ▶ Only few differing class widths
  - ▶ Class widths and class mids should rather be integers
  - ▶ No open marginal classes

Some special cases might have to be accounted for explicitly.

# The histogram

Starting from a given grouping of a metric variable without any open classes, the **areas of the rectangles** drawn above the class intervals  $[x_{j-1}^o; x_j^o)$  are matched to the relative frequencies  $p_j$ .

The rectangles' **heights  $h_j$**  are calculated by rearranging  $p_j = d_j \cdot h_j$ , with  $d_j = x_j^o - x_{j-1}^o$  as the class width.

width \* height

# The empirical distribution function

The **empirical distribution function** specifies the share of observations which exhibit a value less than or equal to  $x$ .

It holds that:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(x_i \leq x) \quad , \quad \text{with} \quad \mathcal{I}(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{else} \end{cases}$$

as the indicator function.

## Some remarks (1)

- ▶ The subscript  $n$ , giving the size of the population/sample, may be dropped due to redundancy.
- ▶ Instead of the empirical distribution function (also called the cumulative relative frequency distribution), the cumulative absolute frequency distribution is used occasionally.  
Then, the following holds:  $F_n^*(x) = n \cdot F_n(x)$ .
- ▶ The empirical distribution function requires at least an ordinal variable. An interpretation of distances on the abscissa is only possible for metric variables.

## Some remarks (2)

- ▶ When we start with grouped data, information on concrete values within classes is typically missing. Nevertheless, by using the class boundaries we may still calculate the cumulative relative frequencies. Given the assumption that the values are uniformly distributed within the classes, we can connect the cumulative relative frequencies at the class boundaries  $F(x_j^o)$  as a polygonal line, thereby connecting the following points:  $(x_0^o, 0); (x_1^o, F(x_1^o)); \dots; (x_{m-1}^o, F(x_{m-1}^o)); (x_m^o, 1)$ .

Then, within the  $j$ -th class ( $x_{j-1}^o \leq x < x_j^o$ ) we have

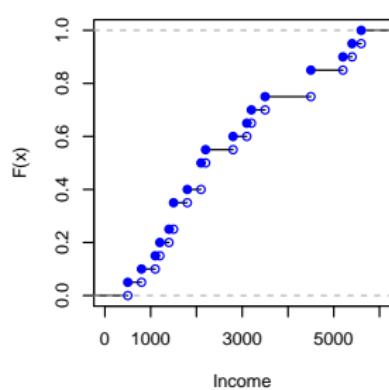
$$F(x) = F(x_{j-1}^o) + p_j \cdot \frac{x - x_{j-1}^o}{x_j^o - x_{j-1}^o} .$$

## Example 2.5: Income (1)

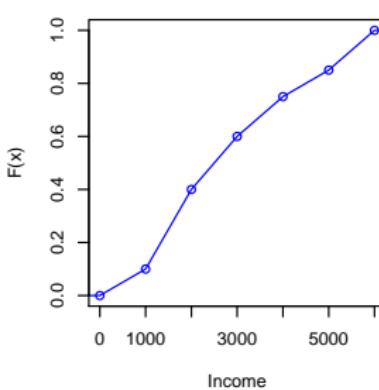
In an income study with a sample size of  $n = 20$ , the following values were recorded: 3500, 3200, 2100, 500, 1800, 2100, 5600, 4500, 1400, 1200, 1500, 2200, 3100, 1500, 2800, 1100, 5200, 4500, 5400, 800.

The resulting empirical distribution function (original and grouped data) as well as the histogram with class boundaries of 0, 1000, 2000, 3000, 4000, 5000 and 6000 are shown here:

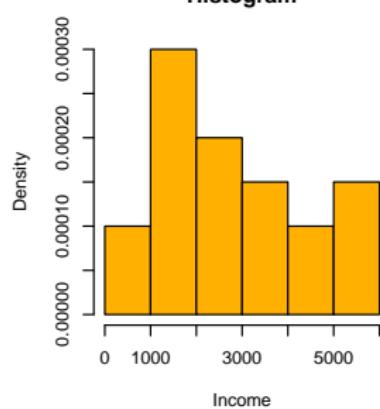
Empirical distribution function



... for 6 classes



Histogram



## Example 2.5: Income (2)

Graphics of the previous slide in R:

```
x2_5 <- c(3500,3200,2100,500,1800,2100,5600,4500,1400,1200,  
        1500,2200,3100,1500,2800,1100,5200,4500,5400,800)  
ant <- (1:length(x2_5)) / length(x2_5)  
  
par(mfrow=c(1,3))  
plot(ecdf(x2_5), xlab = "Income",  
      ylab = expression(F[n](x)),  
      main = "Empirical distribution function",  
      col = "blue")  
points(sort(x2_5), c(0, ant[-length(ant)]), col = "blue")  
  
x_o <- c(0,1000,2000,3000,4000,5000,6000) # Overwriting  
x2_5_kl <- cut(x2_5, x_o, right = FALSE)  
F_j <- cumsum(prop.table(table(x2_5_kl)))  
plot(x = c(0,0), main = "...for 6 classes",  
      xlab = "Income", ylab = expression(F[n](x)),  
      xlim = c(0,6000), ylim = c(0,1), type = "n")  
lines(x = x_o, y = c(0,F_j), col = "blue")  
points(x = x_o, y = c(0,F_j), col = "blue")
```

## Example 2.5: Income (3)

```
hist(x2_5, probability = TRUE, xlab = "Income",  
      ylab = "Density", main = "Histogram", col = "#FFB000")  
par(mfrow = c(1, 1))
```

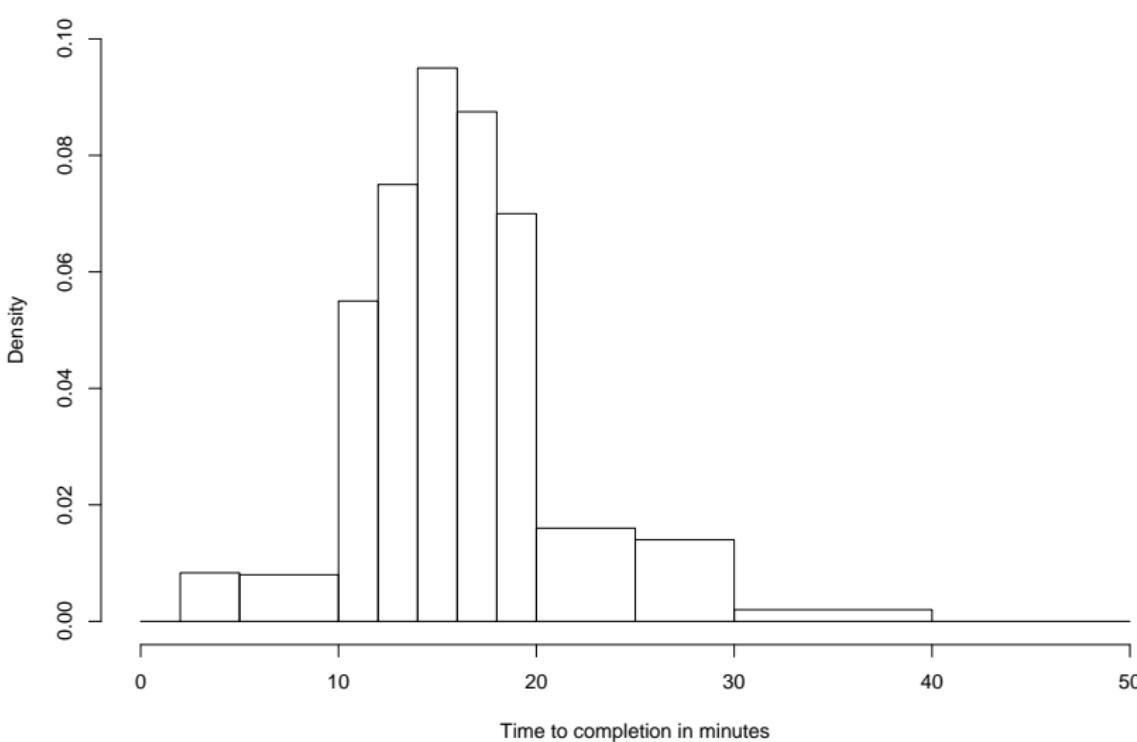
## Example 2.6: Time to completion (1)

For  $n = 200$  pupils, the time taken to solve a problem was recorded:

Class (min.)	$j$	$n_j$	$p_j$	$d_j$	$h_j$	$\sum_{\nu=1}^j p_\nu$
2 up to, but less than 5	1	5	0.025	3	0.0083	0.025
5 up to, but less than 10	2	8	0.040	5	0.0080	0.065
10 up to, but less than 12	3	22	0.110	2	0.0550	0.175
12 up to, but less than 14	4	30	0.150	2	0.0750	0.325
14 up to, but less than 16	5	38	0.190	2	0.0950	0.515
16 up to, but less than 18	6	35	0.175	2	0.0875	0.690
18 up to, but less than 20	7	28	0.140	2	0.0700	0.830
20 up to, but less than 25	8	16	0.080	5	0.0160	0.910
25 up to, but less than 30	9	14	0.070	5	0.0140	0.980
30 up to, but less than 40	10	4	0.020	10	0.0020	1.000
$\sum$		200	1.000			

See Schaich, E.: Schätz- und Testmethoden für Sozialwissenschaftler (1998), 3rd edition, Vahlen, p. 17 ff.

## Example 2.6: Time to completion (2)



## Example 2.6: Time to completion (3)

Histogram of the previous slide in R:

```
load("Example2-6.RData")

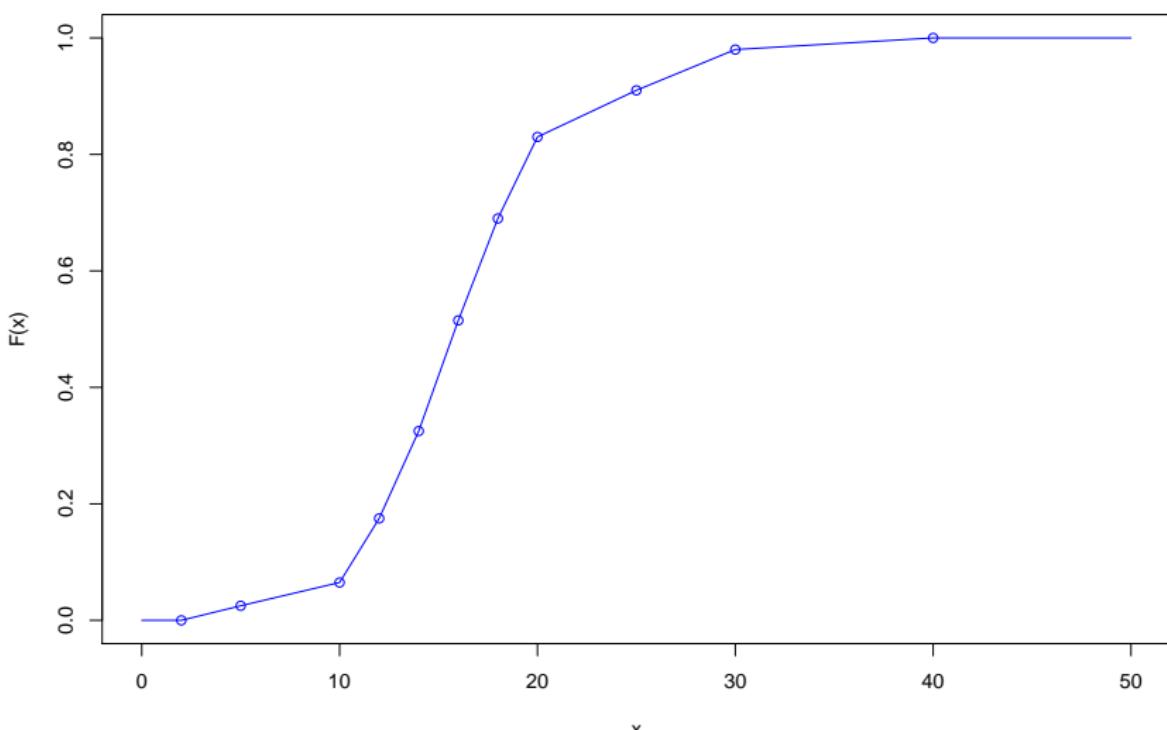
x_o <- c(2,5,10,12,14,16,18,20,25,30,40)
hist(Time, breaks = x_o,
      right = FALSE, main = "",
      xlim = c(0,50), ylim = c(0,0.1),
      xlab = "Time to completion in minutes", ylab = "Density")
```

Empirical distribution function of the next slide in R:

```
F_j <- cumsum(prop.table(table(cut(Time, x_o,
                                         right = FALSE))))
plot(x = c(0,0), type = "n",
      main = "Empirical distribution function",
      ylab = "F(x)", xlab = "x", xlim = c(0,50),
      ylim = c(0,1))
lines(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
points(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
```

## Example 2.6: Time to completion (4)

Empirical distribution function



# Kernel density estimation

Instead of histograms, which show discontinuities at class boundaries, *approximations* may be used.

## Kernel density estimator

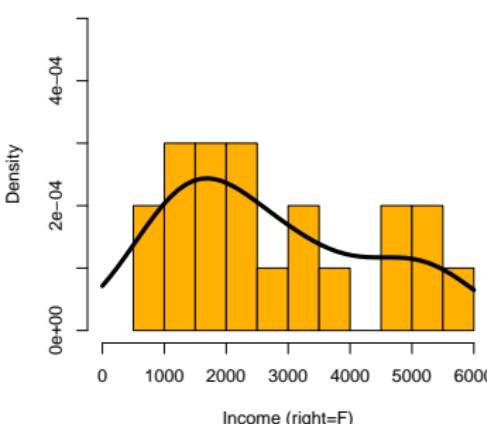
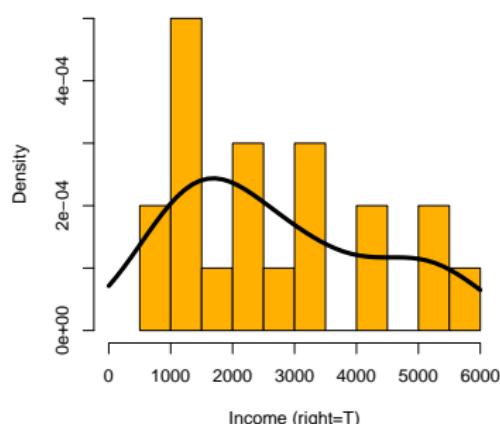
For a given kernel  $K(u)$  and the data  $x_1, \dots, x_n$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad , \quad x \in \mathbb{R}$$

is the kernel density estimator with kernel  $K$  and smoothing parameter  $h$ .

For example  $K(u) = \frac{3}{4}(1 - u^2)$  is the Epanechnikov kernel. The parameter  $h$  affects the width of the intervals within which observations are still considered in the kernel.

## Example 2.7: Histogram vs. kernel density estimation



```
x <- c(3500, 3200, 2100, 500, 1800, 2100, 5600, 4500, 1400,  
1200, 1500, 2200, 3100, 1500, 2800, 1100, 5200, 4500,  
5400, 800)  
hist(x, probability=TRUE, breaks=11, xlim=c(0,6000), right=F,  
col="#FFB000", main="", xlab="Income (right=F)",  
ylab="Density")  
lines(density(x, n=50, from=0, to=6000), lwd=4)
```

# A two-dimensional frequency table

Cat. of 1st variable \ Cat. of 2nd variable	1	$\dots$	$k$	$\dots$	$r$	Sum
1	$n_{11}$	$\dots$	$n_{1k}$	$\dots$	$n_{1r}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
$j$	$n_{j1}$	$\dots$	$n_{jk}$	$\dots$	$n_{jr}$	$n_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
$m$	$n_{m1}$	$\dots$	$n_{mk}$	$\dots$	$n_{mr}$	$n_{m\cdot}$
Sum	$n_{\cdot 1}$	$\dots$	$n_{\cdot k}$	$\dots$	$n_{\cdot r}$	$n$

# Terminology of two-dimensional variables (1)

X	First variable with $x_j$ as $j$ -th value ( $j = 1, \dots, m$ )
Y	Second variable with $y_k$ as $k$ -th value ( $k = 1, \dots, r$ )
$n_j$	Absolute frequency of $j$ -th value of variable X (marginal frequency)
$n_{\cdot k}$	Absolute frequency of $k$ -th value of variable Y (marginal frequency)
$n_{jk}$	Joint absolute frequency of $j$ -th value of variable X and $k$ -th value of variable Y
$p_j$	Relative frequency of $j$ -th value of variable X
$p_{\cdot k}$	Relative frequency of $k$ -th value of variable Y
$p_{jk}$	Joint relative frequency of $j$ -th value of variable X and $k$ -th value of variable Y

## Terminology of two-dimensional variables (2)

We have

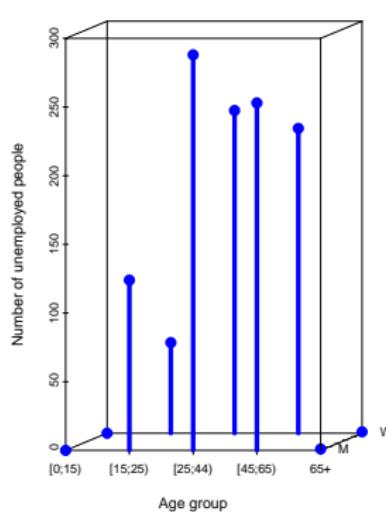
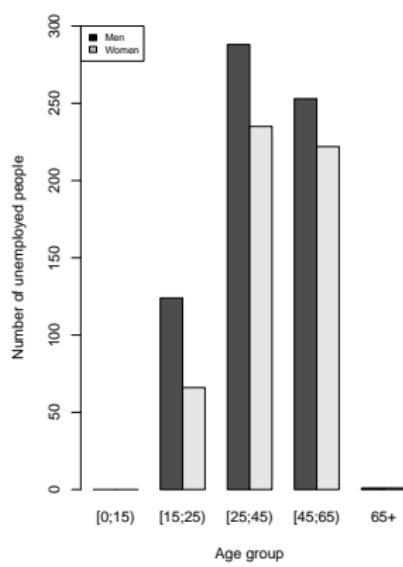
$$\sum_{j=1}^m n_{j \cdot} = \sum_{k=1}^r n_{\cdot k} = \sum_{j=1}^m \sum_{k=1}^r n_{jk} = n$$

and

$$\sum_{j=1}^m p_{j \cdot} = \sum_{k=1}^r p_{\cdot k} = \sum_{j=1}^m \sum_{k=1}^r p_{jk} = 1.$$

## Example 2.8: Unemployment (see Example 2.1)

Joint frequency distribution of unemployed people by age group and gender:



# Elements of Statistics

## Chapter 3: Measures of central tendency and measures of variation

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# Statistical measures

Statistical measures are calculated in order to characterise distributions by means of suitable parameters. In this context, the scaling of the variables of interest plays a major role, as it determines the suitability of different measures.

A distinction is made between

- ▶ measures of central tendency,
- ▶ measures of variation and
- ▶ further measures to describe a distribution.

## Properties of measures of central tendency

**Axiom of identity:** If all values are identical, the measure of central tendency should adopt that same value.

**Axiom of inclusion:** The value of the measure of central tendency should be in the interval  $[x_{\min}; x_{\max}]$ .

**Axiom of translation:** If all values are shifted by a common value, the value of the measure of central tendency should be shifted by this common value as well.

**Axiom of homogeneity:** If the frequencies of all  $m$  different values are (multiplicatively) changed by a common value in such a way that the relative frequencies stay constant, the value of the measure of central tendency should not change (homogeneity of degree zero).

In certain circumstances, restrictions regarding the scaling, like non-negativity, have to be respected as well.

(See Assenmacher, W. (2010): Descriptive Statistik, 4th edition, Springer.)

## The mode

Let a variable of an arbitrary scaling be given. The mode  $x_M$  is the value which occurs most frequently.

For  $i : x_i = x_M$  the following holds:

$$n_i \geq n_j \quad \forall j \neq i \quad .$$

Distributions which have exactly one mode are called unimodal.  
In this case  $n_i > n_j$  holds for all  $j \neq i$ .

In the context of continuous variables, we may call the mode the densest value.

# The quantile

The value  $x_p$  is called  $p$ -quantile if the following holds:

$$x_p = F^{-1}(p) := \inf\{x | F_n(x) \geq p\}$$

Remarks:

- ▶  $0 < p < 1$
- ▶ In this case, the quantile is defined by means of the inverse empirical distribution function.
- ▶ The definition may be broadened (symmetrical case; see Schaich and Münnich, 2001).
- ▶ The second *quartile*  $x_{0.50}$  is called median (see below).
- ▶  $x_{0.25}$  is the first quartile and  $x_{0.75}$  is the third quartile.

## Example 3.1: Quantiles (1)

Given a sample of unsorted original values ( $n = 10$ ):

5.7; 15.6; 12.6; 8.7; 11.9; 15.9; 1.6; 4.9; 19.8; 14.3

After reordering:

1.6; 4.9; 5.7; 8.7; 11.9; 12.6; 14.3; 15.6; 15.9; 19.8

Calculating the 0.2-quantile via the inverse distribution function:

$$x_{0.2} = x_{[2]} = 4.9 .$$

Calculating the first quartile (0.25-quantile) via the inverse distribution function:

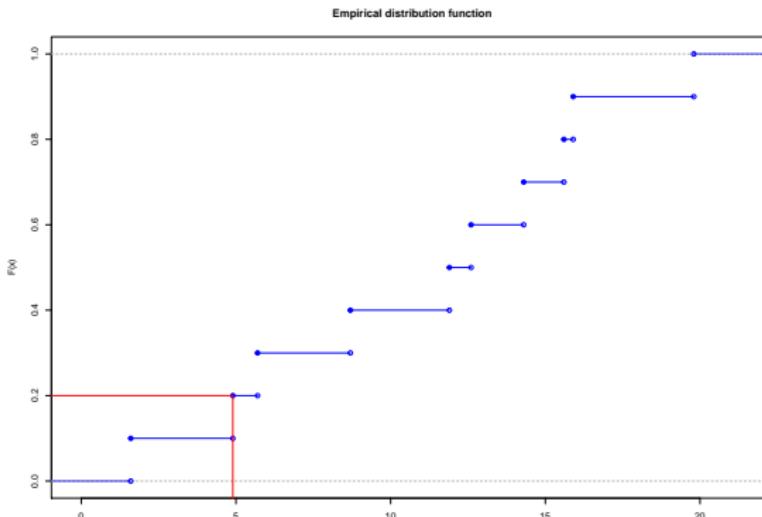
$$x_{0.25} = x_{[3]} = 5.7 .$$

Data input in R:

```
x3_1 <- c(5.7, 15.6, 12.6, 8.7, 11.9, 15.9, 1.6,  
        4.9, 19.8, 14.3)
```

## Example 3.1: Quantiles (2)

Graphical depiction of  $x_{0,2}$ :

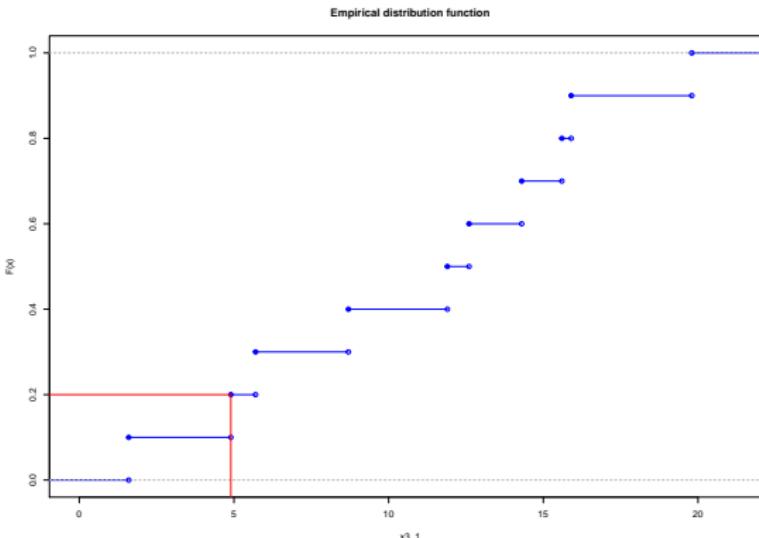


```

> plot(ecdf(sort(x3_1)), xlab = "x3_1", ylab = "F(x)",
+       col = "blue",
+       main = "Empirical distribution function")
> points(sort(x3_1), seq(from = 0,to = 0.9,by = 0.1),
+         col = "blue")
> lines(x = c(-1,4.9), y = c(0.2,0.2), col = "red")
> lines(x = c(4.9,4.9), y = c(0.2,-1), col = "red")
  
```

## Example 3.1: Quantiles (3)

Graphical depiction of  $x_{0,25}$ :



```
# New plot like on the last slide
lines(x = c(0,5.7), y = c(0.25,0.25), col = "red")
lines(x = c(5.7,5.7), y = c(0,0.3), col = "red")
```

## Example 3.1: Quantiles (4)

Determination of  $x_{0.2} = 4.9$  and  $x_{0.25} = 5.7$  in R:

```
sort(x3_1)
```

```
[1] 1.6 4.9 5.7 8.7 11.9 12.6 14.3 15.6 15.9 19.8
```

```
quantile(x = x3_1, probs = c(0.2, 0.25), type = 1)
```

```
20% 25%
```

```
4.9 5.7
```

The determination of a quantile **with the inverse distribution function** is done in R with `quantile(x, type = 1)`.

## The median

Let a variable of at least ordinal scaling be given. Then the median

$$z := x_{0.5}$$

is the 0.5-quantile.

The median divides the smaller 50% from the larger 50% of values of a distribution.

Computation of the median:

- ▶ For uneven  $n$ :  $x_{0.5} = x_{[(n+1)/2]}$
- ▶ For even  $n$ :  $x_{0.5} = \frac{1}{2} \cdot (x_{[n/2]} + x_{[n/2+1]})$   
(In a strict sense, metric scaling would be needed here.)

## Exemplary median computations (1)

### Example 3.2:

Original values ( $n = 9$ ):

13.1; 12.5; 8.3; 6.4; 9.1; 10.5; 10.8; 17.9; 22.3

Ordered values:

6.4; 8.3; 9.1; 10.5; 10.8; 12.5; 13.1; 17.9; 22.3

$$x_{0,5} = x_{[5]} = 10,8$$

Calculation in R:

```
x3_2 <- c(13.1, 12.5, 8.3, 6.4, 9.1, 10.5,  
        10.8, 17.9, 22.3)  
sort(x3_2)
```

```
[1] 6.4 8.3 9.1 10.5 10.8 12.5 13.1 17.9 22.3
```

```
median(x3_2)
```

```
[1] 10.8
```

## Exemplary median computations (2)

### Example 3.3:

(Ordered) original values ( $n = 10$ ):

6; 17; 22; 22; 23; 31; 34; 80; 90; 200

$$x_{0.5} = \frac{1}{2} \cdot (x_{[5]} + x_{[6]}) = 27$$

Calculation in R:

```
x3_3 <- c(6, 17, 22, 22, 23, 31, 34, 80, 90, 200)
sort(x3_3)
```

```
[1] 6 17 22 22 23 31 34 80 90 200
```

```
median(x3_3)
```

```
[1] 27
```

## Determination of quantiles using grouped frequencies

Given grouped frequency distributions, the  $p$ -quantile is approximatively determined by using the empirical distribution function:

- ▶  $j$  is the class for which  $F(x_{j-1}^o) \leq p < F(x_j^o)$ .
- ▶ Determination of quantile  $x_p$  by linear interpolation:

$$\frac{p - F(x_{j-1}^o)}{F(x_j^o) - F(x_{j-1}^o)} = \frac{x_p - x_{j-1}^o}{x_j^o - x_{j-1}^o}$$

↔

$$x_p = x_{j-1}^o + (x_j^o - x_{j-1}^o) \cdot \frac{p - F(x_{j-1}^o)}{F(x_j^o) - F(x_{j-1}^o)}$$

Remark:

Values within the classes are assumed to be *uniformly* distributed.

## Example 3.4: see Ex. 2.6 (1)

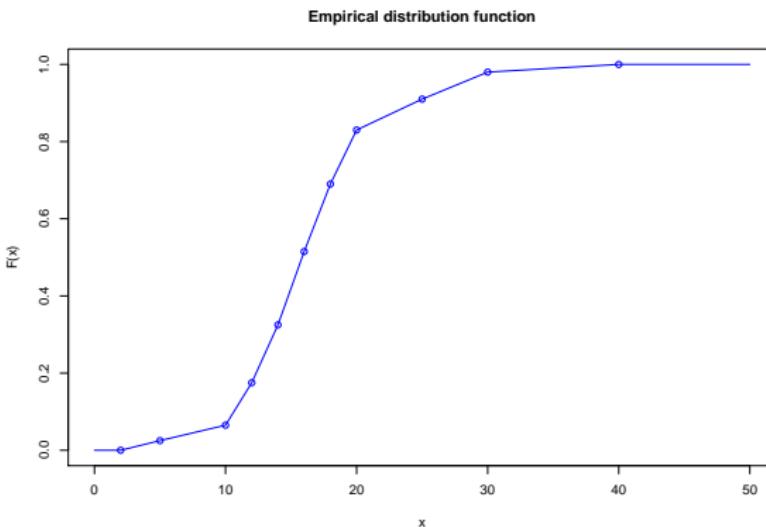
### Remember:

For  $n = 200$  pupils, the time taken to solve a problem was recorded:

	class (min.)	$j$	$n_j$	$p_j$
2	up to, but less than 5	1	5	0.025
5	up to, but less than 10	2	8	0.040
10	up to, but less than 12	3	22	0.110
12	up to, but less than 14	4	30	0.150
14	up to, but less than 16	5	38	0.190
16	up to, but less than 18	6	35	0.175
18	up to, but less than 20	7	28	0.140
20	up to, but less than 25	8	16	0.080
25	up to, but less than 30	9	14	0.070
30	up to, but less than 40	10	4	0.020
$\sum$			200	1.000

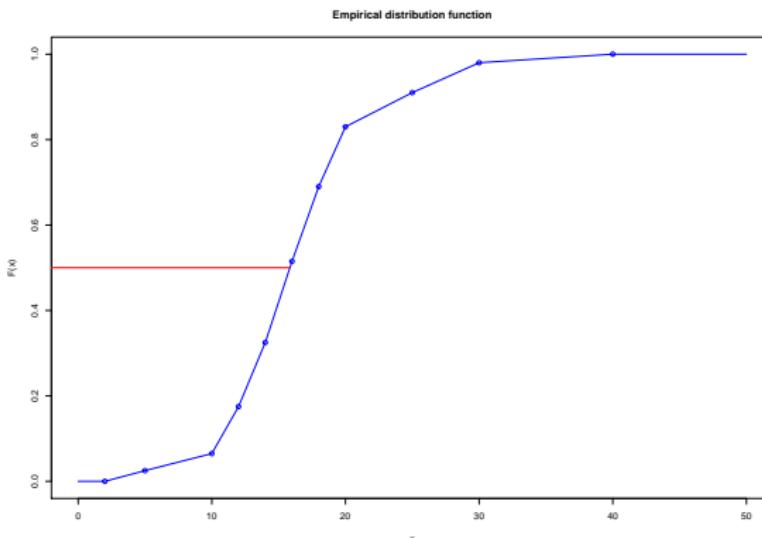
Which value will the median  $z$  have?

## Example 3.4: see Ex. 2.6 (2)



```
setwd("path"); load("Example3-4.RData") # your path
F_j <- cumsum(x = p_j); plot(x = c(0,0), type = "n",
  main = "Empirical distribution function",
  ylab = "F(x)", xlab = "x", xlim = c(0,50), ylim = c(0,1))
lines(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
points(x = x_o, y = c(0,F_j), col = "blue")
```

## Example 3.4: see Ex. 2.6 (3)

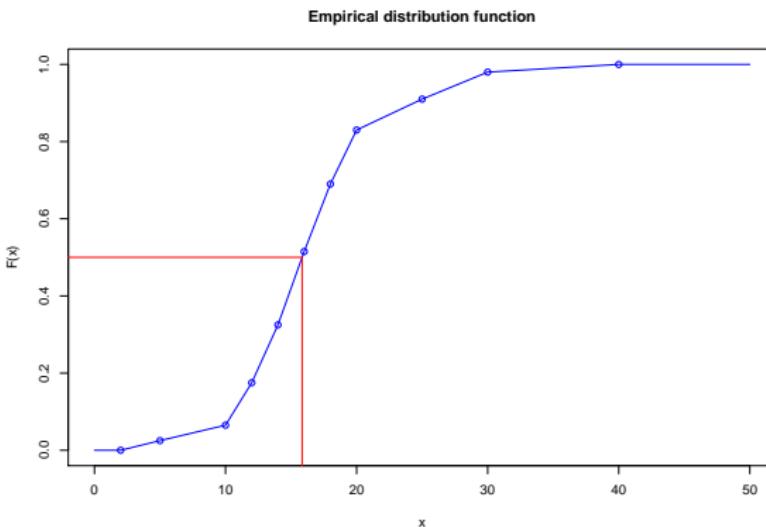


The median falls into class  $j = 5$ .

Graphical depiction in R:

```
lines(x = c(-5, 15.8421), y = c(0.5, 0.5), col = "red")
```

## Example 3.4: see Ex. 2.6 (4)



Finally, we reach

$$z = 14 + (16 - 14) \cdot \frac{0.5 - 0.325}{0.515 - 0.325} = 15.8421.$$

Graphical depiction in R:

```
lines(x = c(15.8421, 15.8421), y = c(0.5, -5), col = "red")
```

## Example 3.4: see Ex. 2.6 (5)

Calculation of  $z = 15.8421$  in R:

```
x_o
```

```
[1] 2 5 10 12 14 16 18 20 25 30 40
```

```
length(x_o)
```

```
[1] 11
```

```
F_j
```

```
[1] 0.025 0.065 0.175 0.325 0.515 0.690 0.830 0.910 0.980 1.000
```

```
length(F_j)
```

```
[1] 10
```

```
j <- 5
x_median <- x_o[j] + (x_o[j+1] - x_o[j]) *
    (0.5 - F_j[j-1])/(F_j[j] - F_j[j-1])
round(x_median, 4)
```

```
[1] 15.8421
```

## The arithmetic mean

Let a variable of metric scaling with  $n$  original values be given. Then

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

is the arithmetic mean.

Example 3.5: see Ex. 3.2

$$\bar{x} = \frac{1}{9} \cdot (13.1 + 12.5 + \dots + 22.3) = \frac{1}{9} \cdot 110.9 \approx 12.32$$

Calculation in R:

```
mean(x3_2)
```

```
[1] 12.32222
```

## The arithmetic mean for grouped data

The arithmetic mean can also be computed as a weighted mean of the means in the different classes:

$$\bar{x} = \frac{1}{n} \cdot \sum_{j=1}^m n_j \cdot \bar{x}_j = \sum_{j=1}^m p_j \cdot \bar{x}_j \quad .$$

It holds that

$$\bar{x} = \frac{1}{n} \cdot \sum_{j=1}^m n_j \cdot \underbrace{\left( \frac{1}{n_j} \sum_{\nu=1}^{n_j} x_{\nu j} \right)}_{\bar{x}_j} = \underbrace{\frac{1}{n} \cdot \sum_{j=1}^m \sum_{\nu=1}^{n_j} x_{\nu j}}_{\text{Overall sum of values}} \quad .$$

If no information on class means ( $\bar{x}_j$ ) is available, the arithmetic mean may still be determined by using the approximations  $\bar{x}_j \approx x'_j$  and

$$\bar{x}' = \sum_{j=1}^m p_j \cdot x'_j \quad .$$

## Example 3.6: see Ex. 2.5 (1)

For the following classification we get:

class	from	up to, but less than	$n_j$	$\bar{x}_j$	$x'_j$
1	0	1500	5	1000.00	750
2	1500	3000	7	2000.00	2250
3	3000	4500	3	3266.67	3750
4	4500	6000	5	5040.00	5250

The arithmetic mean of the original values is  $\bar{x} = 2700$ .

Calculation in R:

```
x2_5 <- c(3500, 3200, 2100, 500, 1800, 2100, 5600, 4500, 1400, 1200,
         1500, 2200, 3100, 1500, 2800, 1100, 5200, 4500, 5400, 800)
mean(x2_5)
[1] 2700
```

## Example 3.6: see Ex. 2.5 (2)

The same value results when using the grouped data

$$\bar{x} = \frac{5}{20} \cdot 1000 + \frac{7}{20} \cdot 2000 + \frac{3}{20} \cdot 3266.67 + \frac{5}{20} \cdot 5040 = 2700 .$$

Calculation in R:

```
x_o <- c(0,1500,3000,4500,6000)
x2_5_kl <- cut(x2_5, x_o, right = FALSE)
n_j <- table(x2_5_kl)

x_mean_j <- tapply(X = x2_5, INDEX = x2_5_kl, FUN = mean)

x_mean_kl <- sum(n_j/sum(n_j) * x_mean_j)

x_mean_kl

[1] 2700
```

## Example 3.6: see Ex. 2.5 (3)

If the arithmetic means of the classes were not available we would get

$$\bar{x}' = \frac{5}{20} \cdot 750 + \frac{7}{20} \cdot 2250 + \frac{3}{20} \cdot 3750 + \frac{5}{20} \cdot 5250 = 2850 \quad .$$

Calculation in R:

```
x_mean_approx_j <- x_o[-5] + (x_o[-1] - x_o[-5]) / 2
x_mean_aprrox_kl <- sum(n_j / sum(n_j) * x_mean_approx_j)
x_mean_aprrox_kl
```

[1] 2850

by using the approximations  $\bar{x}_j \approx x'_j$  and

$$\bar{x}' = \sum_{j=1}^m p_j \cdot x'_j \quad .$$

What happens if we approximate and have over .. up to classes?

## The geometric mean

Let a variable with strictly positive values and exhibiting a ratio scale be given ( $x_i > 0; i = 1, \dots, n$ ). Then

$$g = \sqrt[n]{\prod_{i=1}^n x_i} = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

is the geometric mean.

## Example 3.7: GDP growth

The growth of EU-25 GDP (in %) for the years from 1996 through 2011 is given in the New Cronos database as follows:

1.8; 2.8; 3.0; 3.0; 3.9; 2.1; 1.3; 1.4; 2.5; 2.1; 3.3; 3.2; 0.3; -4.3; 2.1; 1.5.

```
x3_7 <- c(1.8, 2.8, 3.0, 3.0, 3.9, 2.1, 1.3, 1.4, 2.5, 2.1, 3.3, 3.2,
        0.3, -4.3, 2.1, 1.5)
```

By transitioning to growth factors, the mean growth is calculated as

$$g = \sqrt[16]{1.018 \cdot 1.028 \cdots \cdot 1.015} = \sqrt[16]{1.342574} = 1.018582 \quad .$$

```
wa <- x3_7/100 + 1
x_geom <- prod(wa)^(1/length(wa))
x_geom
[1] 1.018582
```

The use of the arithmetic mean would have yielded  $\bar{x} = 1.01875$ .

After 16 years an overall growth of 34.2574% would have to be reported instead of 34.6114%.

## Example 3.8: German population (1)

Population figures for the FRG (in thousands) and the years from 1995 through 2005 are available in the New Cronos database as well:

81,538.6; 81,817.5; 82,012.2; 82,057.4; 82,037.0; 82,163.5;  
82,259.5; 82,440.3; 82,536.7; 82,531.7; 82,500.8.

Official Destatis figures for 2006 – 2011: 82,314.9; 82,217.8; 82,002.4;  
81,802.3; 81,751.6; 81,843.7.

Find the mean of the yearly population growth rate for the FRG in the time span from 1995 until 2005.

Data input in R:

```
x3_8 <- c(81538.6, 81817.5, 82012.2, 82057.4, 82037, 82163.5,  
        82259.5, 82440.3, 82536.7, 82531.7, 82500.8)
```

## Example 3.8: German population (2)

First, the 10 growth factors  $x_t/x_{t-1}$  have to be determined.

```
wa <- x3_8[-1] / x3_8[-length(x3_8)]
```

The corresponding geometric mean is

$$g = \sqrt[10]{\frac{81,817.5}{81,538.6} \cdot \frac{82,012.2}{81,817.5} \cdots \frac{82,500.8}{82,531.7}} = \sqrt[10]{1.011801} = 1.001174$$

```
x_geom <- prod(wa)^(1/length(wa))
```

```
x_geom
```

```
[1] 1.001174
```

It follows that  $81,538,600 \cdot g^{10} = 82,500,800$ .

This matches the actual population in 2005.

Using the arithmetic mean, we would get

$81,538,600 \cdot 1.001175^{10} = 82,501,380$  after 10 years, thus 580 more than actually existing.

## The harmonic mean

Let a variable with strictly positive values and exhibiting a ratio scale be given ( $x_i > 0; i = 1, \dots, n$ ). Then

$$h = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

is the harmonic mean.

## Example 3.9: New tires (1)

It takes a master mechanic 8 minutes to mount a new set of tires to a car. His apprentice completes this task in 12 minutes. How long does it take the two of them *together* to mount 100 sets of tires?

$$h = \frac{1}{\frac{1}{2} \cdot \left( \frac{1}{8} + \frac{1}{12} \right)} = \frac{48}{5} \left[ \frac{\text{min}}{\text{set}} \right]$$

Calculation in R:

```
x3_9 <- c(8,12)
n <- length(x3_9)

x_harm <- n / sum(1 / x3_9)
x_harm

[1] 9.6
```

## Example 3.9: New tires (2)

Hence, it takes *each* mechanic on average 48 minutes to mount 5 sets of tires. To mount 100 sets, both of them together need

$$100 \text{ [sets]} \cdot \frac{24}{5} \left[ \frac{\text{min}}{\text{set}} \right] = 480 \text{ [min]} \quad ,$$

filling 8 hours.

Using the arithmetic mean, the mounting of one set would have taken 10 minutes per mechanic, resulting in 8 hours and 20 minutes.

## Comparison of arithmetic and harmonic mean

Find the respective average speeds for the following two cases ( $v_1$  and  $v_2$ ):

- ▶ A car drives at 100 km/h for one hour and at 120 km/h for three hours.

We use the weighted arithmetic mean:

$$v_1 = \frac{1}{4} \cdot 100 + \frac{3}{4} \cdot 120 = 115 \quad [km/h]$$

- ▶ A car drives at 100 km/h for 115 km and at 120 km/h for 345 km.

We use the weighted harmonic mean:

$$v_2 = \frac{1}{\frac{115}{460} \cdot \frac{1}{100} + \frac{345}{460} \cdot \frac{1}{120}} = 114.2857 \quad [km/h]$$

The second car will take 90 seconds longer.

## Influence of changes in the unit of measurement

The observations  $x_i$  ( $i = 1, \dots, n$ ) are transformed linearly:

$$y_i = a \cdot x_i + b \quad , \quad a \neq 0, b \in \mathbb{R} .$$

It follows that:

$$y_M = a \cdot x_M + b$$

$$z_y = a \cdot z_x + b$$

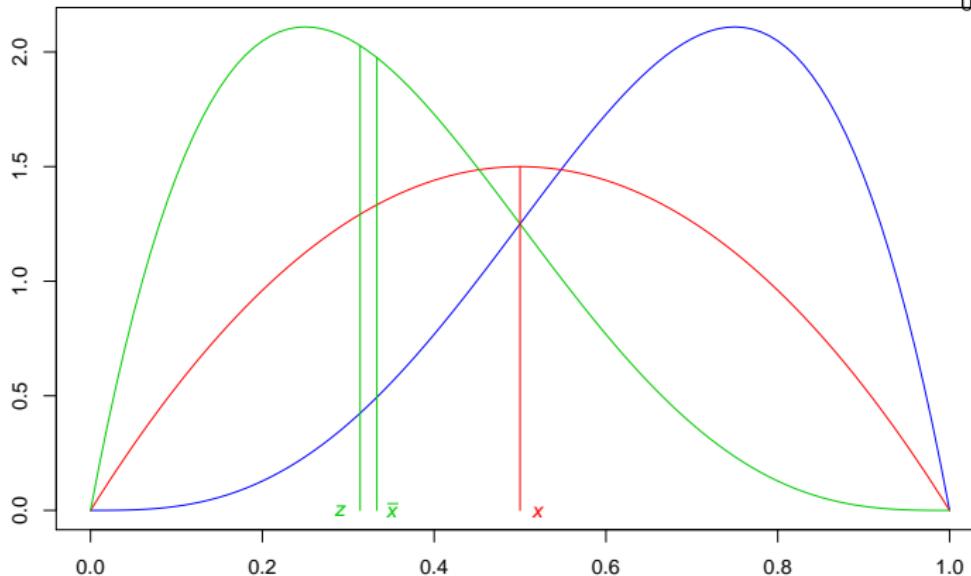
$$\bar{y} = a \cdot \bar{x} + b$$

$g$  and  $h$  only satisfy this transformation for  $b = 0$ . Therefore, the axiom of translation does not hold. For  $b = 0$  it follows that:

$$g_y = a \cdot g_x$$

$$h_y = a \cdot h_x$$

# Comparison of distributions (1)



Symmetric distribution

$$\bar{x} \approx z_x \approx x_M$$

Right-skewed (positively skewed) distribution

$$\bar{x} > z_x > x_M$$

Left-skewed (negatively skewed) distribution

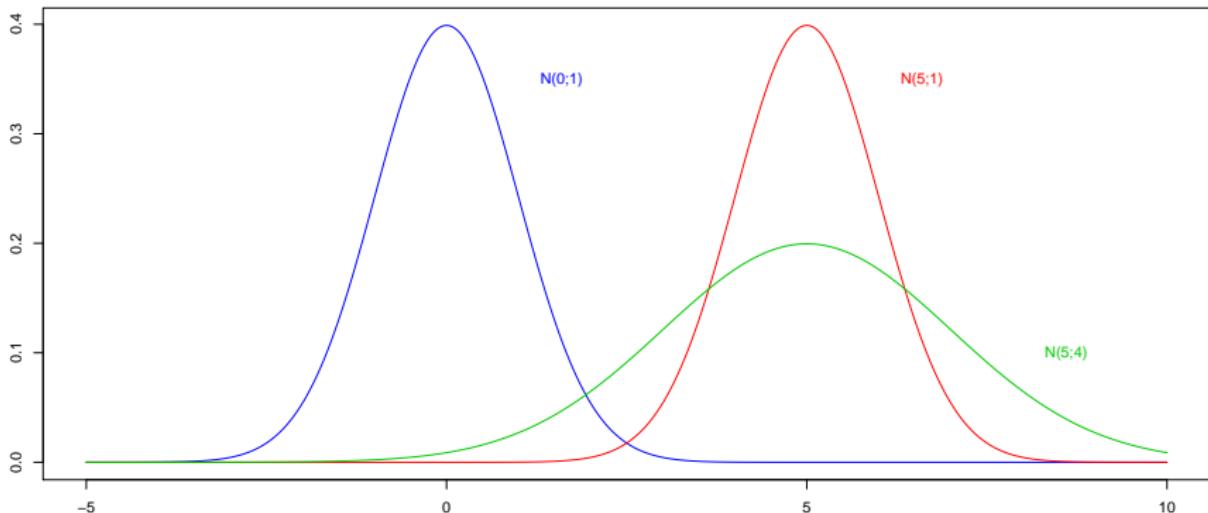
$$\bar{x} < z_x < x_M$$

Always:

$$\bar{x} \geq g \geq h$$

# Comparison of distributions (2)

Comparison of distributions



```
y <- seq(from = -5, to = 10, by = 0.01)
plot(y, dnorm(y, mean = 0, sd = 1), type = "l", xlab = "", 
      ylab = "", main = "Comparison of distributions", col=4)
lines(y, dnorm(y, mean = 5, sd = 1), col = 2)
lines(y, dnorm(y, mean = 5, sd = 2), col = 3)
text(x = 1.6, y = 0.35, label = "N(0;1)", col = 4)
text(6.6, 0.35, "N(5;1)", col=2); text(8.6, 0.1, "N(5;4)", col=3)
```

# Properties of measures of variation

## Degenerate distribution

If all observations are equal, there is no variation in the data and the measure of variation should take on the value zero.

## Positive value

When there are at least two unique observations, the measure of variation should take on a positive value.

## Translation invariance

If each observation is shifted by a common constant, the measure of variation should remain unchanged.

## Axiom of homogeneity

If the frequencies of all  $m$  different values are (multiplicatively) changed by a common value in such a way that the relative frequencies stay constant, the value of the measure of variation should not change (homogeneity of degree zero). The empirical distribution function will not be changed!

# Range and interquartile range

## Range

The range  $w$  is defined by

$$w = \max_{i=1,\dots,n} x_i - \min_{i=1,\dots,n} x_i = x_{[n]} - x_{[1]} .$$

$x_{[1]}$  ( $x_{[n]}$ ) is the first ( $n$ -th) element of the ordered list of values.

## Interquartile range

The interquartile range IQR is

$$\text{IQR} = x_{0.75} - x_{0.25}$$

and therefore equal to the difference of the third and first quartile.

## Mean linear deviation

The mean linear deviation  $l$  is given by

$$l = \frac{1}{n} \sum_{i=1}^n |x_i - z| .$$

The median minimises the mean linear deviation as a *measure of distance*:

$$\arg \min_{t \in \mathbb{R}} l(t) = \arg \min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |x_i - t| = z .$$

# Variance

The variance  $s^2$  is defined as

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 .$$

The arithmetic mean  $\bar{x}$  minimises the variance as a *measure of distance*:

$$\arg \min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (x_i - t)^2 = \bar{x}$$

*Proof using derivative...*

## Variance and *displacement law* - *Der Verschiebungssatz*

We have

$$s^{*2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \quad .$$

Furthermore, we use the *inferential variance*:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \cdot s^{*2} \\ &= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \cdot \bar{x}^2 \quad . \end{aligned}$$

The variance  $s^{*2}$  is often called the *empirical variance*.

# Standard deviation and coefficient of variation

The standard deviation is the square root of the variance:

$$s^* = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{or} \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} .$$

The standard deviation has the same unit of measurement as the arithmetic mean.

As a unit-independent measure of variation, we may use the coefficient of variation

$$v = \frac{s^*}{\bar{x}} \cdot 100\% .$$

It requires a ratio scale.

## Example 3.10: Screws (1)

Quality control yielded the following actual (and squared actual) lengths of two sets of 10 screws which were supposed to be 10 mm and 80 mm long:

$i$	$x_{1,i}$	$x_{1,i}^2$	$x_{2,i}$	$x_{2,i}^2$
1	10.40	108.16	81.00	6561.00
2	9.90	98.01	80.30	6448.09
3	10.30	106.09	80.00	6400.00
4	9.80	96.04	79.90	6384.01
5	9.90	98.01	79.80	6368.04
6	10.30	106.09	80.60	6496.36
7	10.40	108.16	80.30	6448.09
8	10.10	102.01	80.20	6432.04
9	10.20	104.04	80.30	6448.09
10	9.70	94.09	80.60	6496.36
$\Sigma$	101.00	1020.70	803.00	64482.08

## Example 3.10: Screws (2)

Data input in R:

```
x_1 <- c(10.4, 9.9, 10.3, 9.8, 9.9, 10.3, 10.4, 10.1, 10.2,  
        9.7)  
x_2 <- c(81, 80.3, 80, 79.9, 79.8, 80.6, 80.3, 80.2, 80.3,  
        80.6)
```

First, we calculate:

$$\bar{x}_1 = \frac{1}{10} \cdot (10.4 + \dots + 9.7) = 10.1$$
$$\bar{x}_2 = \frac{1}{10} \cdot (81.0 + \dots + 80.6) = 80.3$$

```
mean(x_1)  
[1] 10.1
```

```
mean(x_2)  
[1] 80.3
```

## Example 3.10: Screws (3)

We calculate further:

$$s_1^{*2} = \frac{1}{10} \cdot (10.4^2 + \dots + 9.7^2) - 10.1^2 = 0.06$$

$$s_2^{*2} = \frac{1}{10} \cdot (81.0^2 + \dots + 80.6^2) - 80.3^2 = 0.118$$

empirical variance

Calculation in R:

```
x_1_var <- (length(x_1) - 1) / length(x_1) * var(x_1)
x_2_var <- (length(x_2) - 1) / length(x_2) * var(x_2)
```

```
x_1_var
```

```
[1] 0.06
```

```
x_2_var
```

```
[1] 0.118
```

## Example 3.10: Screws (4)

From this, we get  $s_1^* = 0.244949$  and  $s_2^* = 0.3435113$ .

```
sqrt(x_1_var)
[1] 0.244949
sqrt(x_2_var)
[1] 0.3435113
```

The relative dispersion are  $v_1 = 2.425237\%$  and  $v_2 = 0.4277849\%$

Calculation in R:

```
x_1_var_koeff <- sqrt(x_1_var) / mean(x_1) * 100
x_2_var_koeff <- sqrt(x_2_var) / mean(x_2) * 100

x_1_var_koeff
[1] 2.425237
x_2_var_koeff
[1] 0.4277849
```

## Variance decomposition

If a population is divided into  $m$  subpopulations, the variance can be decomposed as follows:

$$s^*{}^2 = s_b^*{}^2 + s_w^*{}^2$$

( $b$ : between;  $w$ : within) with

$$s_b^*{}^2 = \sum_{j=1}^m p_j \cdot (\bar{x}_j - \bar{x})^2$$

$$s_w^*{}^2 = \sum_{j=1}^m p_j \cdot s_j^*{}^2$$

The two parts are the *external and internal variance*, where

$$s_j^*{}^2 = \frac{1}{n_j} \sum_{\nu=1}^{n_j} (x_{\nu j} - \bar{x}_j)^2 \quad , \quad \bar{x} = \sum_{j=1}^m p_j \cdot \bar{x}_j$$

# Variance and grouped data

- ▶ If class means and class variances are known:  
Variance decomposition
- ▶ If class means and class variances are unknown:

$$s^{*2} = \sum_{j=1}^m p_j \cdot (\underbrace{\bar{x}_j - \bar{x}}_{x'_j - \bar{x}'})^2 + \underbrace{\sum_{j=1}^m p_j \cdot s_j^{*2}}_{\approx 0} \quad \text{with} \quad \bar{x}' = \sum_{j=1}^m p_j \cdot x'_j.$$

Therefore, we have:

$$s'^{*2} = \sum_{j=1}^m p_j \cdot (x'_j - \bar{x}')^2 = \sum_{j=1}^m p_j \cdot x'^2_j - \bar{x}'^2.$$

## Example 3.11: Income classes (1)

Let the following  $n = 10$  income values be given:

3500; 3200; 2100; 500; 1800; 2100; 5600; 8500; 1400; 1200 Furthermore,

let the class boundaries be:

$$x_0^o = 0; x_1^o = 2500; x_2^o = 5000; x_3^o = 7500; x_4^o = 10000.$$

Data input in R:

```
x3_11 <- c(3500,3200,2100,500,1800,2100,5600,8500,1400,1200)
x_o <- c(0,2500,5000,7500,10000)
x3_11_kl <- cut(x3_11, x_o, right = FALSE)
```

$j$	$p_j$	$\bar{x}_j$	$s_j^{*2}$	$p_j(\bar{x}_j - \bar{x})^2$	$p_j s_j^{*2}$	$x'_j$	$p_j x'_j$	$p_j(x'_j - \bar{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350.00	22500	25920	4500	3750	750	112500
3	0.1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

## Example 3.11: Income classes (2)

$j$	$p_j$	$\bar{x}_j$	$s_j^{*2}$	$p_j(\bar{x}_j - \bar{x})^2$	$p_js_j^{*2}$	$x'_j$	$p_jx'_j$	$p_j(x'_j - \bar{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350.00	22500	25920	4500	3750	750	112500
3	0.1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

Calculation of  $p_j$  and  $n_j$  in R:

```
p_j <- prop.table(x = table(x3_11_kl))
n_j <- p_j * length(x3_11)
```

Calculation of  $\bar{x}_j$  in R:

```
x_mean_j <- tapply(X = x3_11, INDEX = x3_11_kl, FUN = mean)
```

Calculation of  $s_j^{*2}$  in R:

```
x_var_j <- (n_j - 1)/n_j *
tapply(X = x3_11, INDEX = x3_11_kl, FUN = var)
x_var_j
```

[0,2.5e+03) [2.5e+03,5e+03) [5e+03,7.5e+03) [7.5e+03,1e+04)  
 318055.6 22500.0

## Example 3.11: Income classes (3)

$j$	$p_j$	$\bar{x}_j$	$s_j^{*2}$	$p_j(\bar{x}_j - \bar{x})^2$	$p_js_j^{*2}$	$x'_j$	$p_jx'_j$	$p_j(x'_j - \bar{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350.00	22500	25920	4500	3750	750	112500
3	0.1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

$$s^{*2} = s_b^{*2} + s_w^{*2} = 5045567 + 195333 = 5240900$$

Calculation of  $s^{*2}$  in R:

```
x_var_j[is.na(x_var_j)] <- 0

x_var_between <- sum(p_j * (x_mean_j - mean(x3_11))^2)
x_var_within <- sum(p_j * x_var_j)

x_var_kl <- x_var_between + x_var_within
```

x\_var\_between

[1] 5045567

x\_var\_within

[1] 195333.3

x\_var\_kl

[1] 5240900

## Example 3.11: Income classes (4)

$j$	$p_j$	$\bar{x}_j$	$s_j^{*2}$	$p_j(\bar{x}_j - \bar{x})^2$	$p_js_j^{*2}$	$x'_j$	$p_jx'_j$	$p_j(x'_j - \bar{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350.00	22500	25920	4500	3750	750	112500
3	0.1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

$$s'^{*2} = 6312500$$

Calculation of  $x'_j$  and  $\bar{x}'$  in R:

```
x_mean_approx_j <- x_o[-5] + (x_o[-1] - x_o[-5]) / 2
x_mean_approx_kl <- sum(p_j * x_mean_approx_j)
```

Calculation of  $s'^{*2}$  in R:

```
x_var_approx_kl <- sum(p_j * (x_mean_approx_j -
x_mean_approx_kl)^2)
x_var_approx_kl
```

[1] 6312500

## Example 3.12: see Ex. 3.11

An alternative grouping according to

- ▶  $x_0^o = 0; x_1^o = 2000; x_2^o = 4000; x_3^o = 10000$  would lead to

$$s'^{*2} = 4,800,000.$$

- ▶  $x_0^o = 0; x_1^o = 2000; x_2^o = 5000; x_3^o = 10000$  would lead to

$$s'^{*2} = 5,660,000.$$

In both cases we have

$$s_b^{*2} = 4,570,900 \quad \text{and} \quad s_w^{*2} = 670,000.$$

Calculation in R with use of the R code from Ex. 3.11.

For this, the vector **x\_o** from Ex. 3.11 must be changed.

E.g. for the first case above of alternative grouping of the data:

```
x_o <- c(0,2000,4000,10000)
```

## Influence of changes in the unit of measurement

If the  $n$  original values  $x_i$  are linearly transformed ( $y_i = a_0 + a_1 \cdot x_i$ ), we have

$$\bar{y} = a_0 + a_1 \cdot \bar{x} \quad , \quad s_y^{*2} = a_1^2 \cdot s_x^{*2} \quad , \quad s_y^* = |a_1| \cdot s_x^*$$

and

$$v_y = \frac{|a_1| \cdot s_x^*}{a_0 + a_1 \cdot \bar{x}} \quad .$$

For the **standard transformation**

$$y_i = \frac{x_i - \bar{x}}{s_x^*}$$

we specifically have  $\bar{y} = 0$  and  $s_y^{*2} = 1$ .

# Entropy

The *spread* of values of variables on a nominal or an ordinal scale can be measured by entropy, which is:

$$E = - \sum_{j=1}^m p_j \cdot \ln p_j = \ln n - \frac{1}{n} \sum_{j=1}^m n_j \cdot \ln n_j .$$

Entropy reaches its maximum when frequencies are identical.

As relative entropy

$$E_r = \frac{E}{\ln m}$$

is used. Then we have  $0 \leq E_r \leq 1$ .

Notice that measures of variation are usually based on metric scales!

## Skewness and kurtosis

- ▶ The skewness of a distribution is measured by

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \quad \text{or} \quad \frac{m_3}{s^{*3}} .$$

- ▶ The kurtosis of a distribution is measured by

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 \quad \text{or} \quad \frac{m_4}{s^{*4}} .$$

The normal distribution exhibits a kurtosis of 3.

(see *Comparison of distributions (1) and (2)*) (slide 33)

# Stem-and-leaf plot

As a stem, the first digit is used ( $0, \dots, 5$ ). Then, the leafs (next digit) will be attached on the stem in increasing order. The number of stems can be variated corresponding to the size of the data.

**Example 3.13:** see Ex. 2.5 resp. 3.6

```
sort(x2_5)
```

```
[1] 500 800 1100 1200 1400 1500 1500 1800 2100 2100 2200 2800 3100 3200  
[15] 3500 4500 4500 5200 5400 5600
```

```
stem(x = x2_5, scale = 2)
```

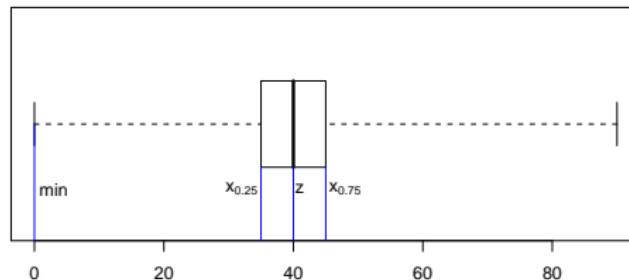
The decimal point is 3 digit(s) to the right of the |

```
0 | 58  
1 | 124558  
2 | 1128  
3 | 125  
4 | 55  
5 | 246
```

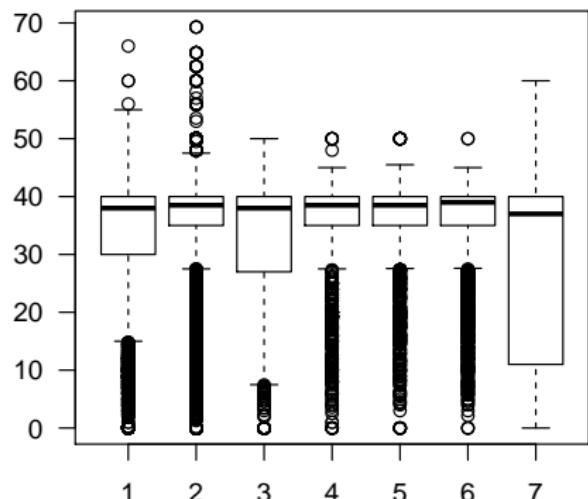
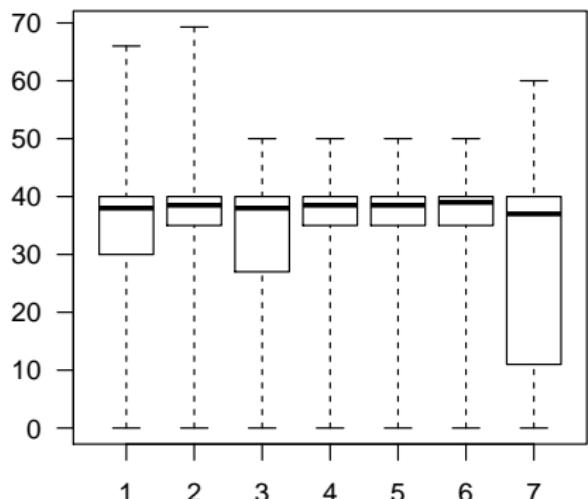
## The boxplot

A boxplot (or box-whisker-plot) describes a distribution by means of five chosen parameters: minimum,  $x_{0.25}$ ,  $z$ ,  $x_{0.75}$  and maximum. The box consists of  $x_{0.25}$ ,  $z$  and  $x_{0.75}$ , the length of the box being equal to the interquartile range  $IQR := x_{0.75} - x_{0.25}$ . The *whiskers* reach from the ends of the box to the minimum and maximum, respectively.

On a modified boxplot the whiskers may be bounded by the values  $x_{0.25} - 1.5 \cdot IQR$  and  $x_{0.75} + 1.5 \cdot IQR$ , respectively. Each observation outside of those boundaries is plotted as an individual value.



# Working hours by qualification



Box-Plots in R:

```
boxplot(Hours ~ Qualification, data = AZ, range = 0)
boxplot(Hours ~ Qualification, data = AZ, range = 1.5)
```

# Elements of Statistics

## Chapter 4:

### Measures of association and regression analysis

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# Measures of association for multidimensional distributions

Here: 2 variables with pairs of variates ( $x_i; y_i$ )

## 1. Univariate analysis for each variable

$$\bar{x}, \bar{y}, s_x^{*2}, s_y^{*2}, v_x, v_y$$

## 2. Analysis of the variables' relationship

Problems:

- ▶ Are we able to infer the value of one variable from the value of the other variable?
  - Measurement of (strength of) variables' relationship
- ▶ Is there an *algorithm* (function), governing the variables' relationship?
  - Regression analysis

# Two-dimensional distributions (1)

		cat. of 2nd variable	1	...	$k$	...	$r$	sum
		1	$n_{11}$	...	$n_{1k}$	...	$n_{1r}$	$n_{1\cdot}$
cat. of 1st variable	1	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	$j$	$n_{j1}$	...	$n_{jk}$	...	$n_{jr}$	$n_{j\cdot}$	$\vdots$
		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$m$	$n_{m1}$	...	$n_{mk}$	...	$n_{mr}$	$n_{m\cdot}$
		sum	$n_{\cdot 1}$	...	$n_{\cdot k}$	...	$n_{\cdot r}$	$n$

 $n_{jk}$ 

Joint absolute frequency of  $j$ -th value of the first variable and  $k$ -th value of the second variable

 $n_{j\cdot}$ 

Absolute frequency of  $j$ -th value of the first variable

 $n_{\cdot k}$ 

Absolute frequency of  $k$ -th value of the second variable

## Two-dimensional distributions (2)

The relative frequencies are given by

$$p_{jk} = \frac{n_{jk}}{n},$$

while

$$p_{k|j} = \frac{n_{jk}}{n_j}.$$

are conditional relative frequencies.

Condition: The first variable's realisation is  $j$ .

## Example 4.1: Unemployment (see Example 2.1):

Unemployed	[0; 15)	[15; 25)	[25; 45)	[45; 65)	at least 65	$\sum$
Men	0	124	288	253	1	666
Women	0	66	235	222	1	524
$\sum$	0	190	523	475	2	1190

The absolute marginal distributions were already given in Example 2.1.  
 The distribution of unemployed females across age groups is given by:

Unemployed	[0; 15)	[15; 25)	[25; 45)	[45; 65)	at least 65	$\sum$
Women	0	66	235	222	1	524

The relative conditional distribution shows the respective share of unemployed women in the different age groups and is reached by dividing the absolute frequencies by the marginal sum (here: 524).

```
load("Example2-8.RData")
round(FQttable["Female",]/FQttable["Female","Sum"],4)
[0;15) [15;25) [25;45) [45;65) at least 65 Sum
Female    0    0.126   0.4485   0.4237    0.0019    1
```

# Empirical distribution function (bivariate case)

For bivariate data, the empirical distribution function is given by

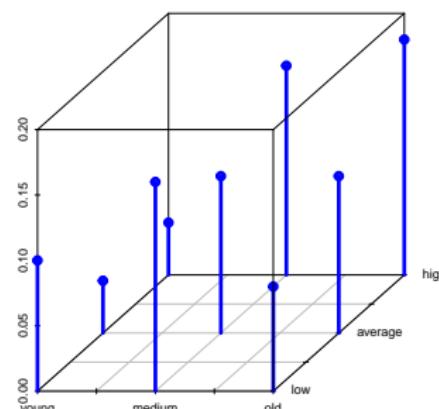
$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(x_i \leq x \wedge y_i \leq y).$$

- ▶ for simplicity reasons, the ECDF is displayed in a tabular format
- ▶ ordinal scale is required at least
- ▶ graphical representation is analogous to the univariate case

## Example 4.2: Kindergarten (1)

The relationship between age group and the interest in doing handicrafts is investigated in a kindergarten accomodating  $n = 50$  children. The results are as follows:

$p_{jk}$	young	medium	old
low	5/50	8/50	4/50
average	2/50	6/50	6/50
high	2/50	8/50	9/50



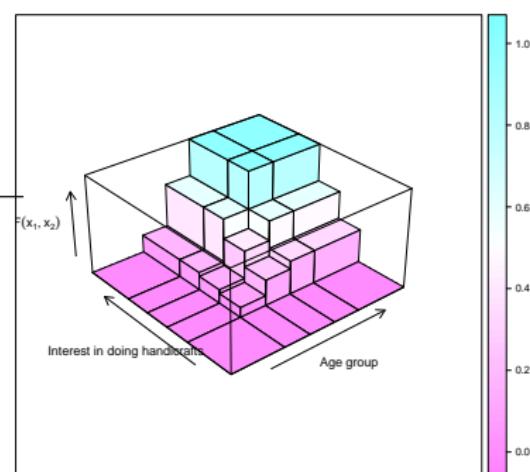
Load data in R:

```
load("Example4-2.RData")
```

## Example 4.2: Kindergarten (2)

Aggregation of values in both directions up to the relevant cell yields the tabulated empirical distribution function  $F_{50}(x, y)$ :

$F_{50}(x, y)$	young	medium	old
low	5/50	13/50	17/50
average	7/50	21/50	31/50
high	9/50	31/50	50/50



## Example 4.2: Kindergarten (3)

Calculation in R:

```
F_j_k <- t(apply(apply(p_j_k, 2, cumsum), 1, cumsum))  
F_j_k
```

	young	medium	old
low	0.10	0.26	0.34
average	0.14	0.42	0.62
high	0.18	0.62	1.00

ECDF

# Coefficient of contingency (1)

- ▶ Nominal scale
- ▶ Contingency table (two-dimensional frequencies)

We need a measure which accounts for the relationship between the two variables.

**Value → 1:** We can infer the value of one variable from the value of the other variable.

**Value → 0:** We cannot even infer a *tendency* for the value of one variable from the value of the other variable.

## Example 4.3: Rompers

Relationship between rompers' colours and babies' gender:

	blue	pink
m	10	0
f	0	10

	blue	pink
m	5	5
f	5	5

	blue	pink
m	8	2
f	2	8

# Independence of variables

Aim:

Comparison of actual distribution of variables and reference distribution which does not allow any *inference*.

## Definition

Two variables are called independent if and only if

$$n_{jk} = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$$

holds for all  $j = 1, \dots, m$  and  $k = 1, \dots, r$ .

Problem: The resulting values  $\frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$  may not be integers.

## Coefficient of contingency (2)

1. Calculation of  $n_{jk}^* = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$   $j = 1, \dots, m$   $k = 1, \dots, r$ .
2. Determination of deviation between actual and theoretical values (independence):

$$\chi^2 = \sum_{j=1}^m \sum_{k=1}^r \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}$$

$$3. K = \sqrt{\frac{\chi^2}{n + \chi^2}} \quad ; \quad 0 \leq K < 1$$

4. Standardisation:

$$K_* = \frac{K}{K_{\max}}; \quad K_{\max} = \sqrt{\frac{M-1}{M}}; \quad M = \min(m, r)$$

## Coefficient of contingency (3)

$K_*$  is called standardised coefficient of contingency. We have:  $0 \leq K_* \leq 1$ .

Independence  $\Rightarrow K_* = 0$

Perfect relation  $\Rightarrow K_* = 1$

## Example 4.4: Field of study and gender (1)

We are interested in the relation between field of study and gender.

$x$  : Field      **B**usiness administration, **E**conomics, **G**eography  
 $y$  : Gender    m, f

$n_{jk}$	m	f	
B	2	1	3
E	2	2	4
G	1	2	3
	5	5	10

→

$n_{jk}^*$	m	f	
B	1.5	1.5	3
E	2	2	4
G	1.5	1.5	3
	5	5	10

## Example 4.4: Field of Study and gender (2)

Calculations in R:

```
load("Example4-4.RData")
addmargins(n_j_k)
```

	m	f	Sum
B	2	1	3
E	2	2	4
G	1	2	3
Sum	5	5	10

```
n_j_k_star <- margin.table(n_j_k,1) %*%
                  t(margin.table(n_j_k,2))/margin.table(n_j_k)
```

```
n_j_k_star
```

	m	f
B	1.5	1.5
E	2.0	2.0
G	1.5	1.5

## Example 4.4: Field of Study and gender (3)

$$\begin{aligned}\chi^2 &= \frac{(2 - 1.5)^2}{1.5} + \frac{(1 - 1.5)^2}{1.5} + \frac{(2 - 2)^2}{2} + \frac{(2 - 2)^2}{2} + \\ &\quad \frac{(1 - 1.5)^2}{1.5} + \frac{(2 - 1.5)^2}{1.5} \\ &= \frac{1/2^2}{3/2} \cdot 4 = \frac{2}{3}\end{aligned}$$

Calculation of  $\chi^2$  in R:

```
chisq <- summary(n_j_k)$statistic  
chisq
```

```
[1] 0.6666667
```

Note that in R, the object `n_j_k` above has to have the structure table (Checking possible with `str()`).

## Example 4.4: Field of Study and gender (4)

$$\Rightarrow K = \sqrt{\frac{(2/3)}{10 + (2/3)}} = \sqrt{\frac{1}{16}} = \frac{1}{4}$$

With  $M = 2$  and  $K_{max} = \sqrt{\frac{1}{2}}$  we have  $K_* = \frac{1}{4} \cdot \sqrt{2} = 0.3536$ .

Calculations in R:

```
KK <- sqrt(chisq/(sum(n_j_k) + chisq))
M <- min(dim(n_j_k))
K_max <- sqrt((M-1)/M)
K_star <- KK/K_max
```

KK	M	K_max	K_star
[1] 0.25	[1] 2	[1] 0.7071068	[1] 0.3535534

## Pearson's $\Phi$ and Cramer's coefficient of contingency

Pearson's  $\Phi$ -coefficient is defined as:

$$\sqrt{\frac{\chi^2}{n}}.$$

We have  $0 \leq \Phi \leq \sqrt{M - 1}$ , where in case of  $\min(m, r) = 2$  we have  $M - 1 = 1$ .

Cramér's coefficient of contingency (Cramér's  $V$ ) is defined as:

$$V = \sqrt{\frac{\chi^2}{n \cdot (M - 1)}}.$$

We have:  $0 \leq V \leq 1$ . In case of  $\min(m, r) = 2$ , we have  $\Phi = V$ .

Both coefficients may only be used for nominal variables.

# Rank correlation coefficient of Spearman (1)

- ▶ **Ordinal scale**
- ▶ Each rank is unique

Additional information compared to coefficient of contingency:

**Positive correlation** The higher the value of one variable, the higher is the value of the other variable

**Negative correlation** The higher the value of one variable, the lower is the value of the other variable

## Rank correlation coefficient of Spearman (2)

Let  $x$  and  $y$  have at least ordinal scaling and no duplicated values in  $x_i$  and  $y_i$ , respectively. The rank correlation coefficient of Spearman is then given by

$$r_{sp} = 1 - \frac{6 \sum_{i=1}^n (\text{Rg}(x_i) - \text{Rg}(y_i))^2}{n(n^2 - 1)}$$

$r_{sp} = +1$  : All ranks are identical

$r_{sp} = -1$  : All ranks are contrary to each other

In order to determine the rank of an attribute in R we can use the function `rank()`.

## Example 4.5: Alpine skiing

Let the results for a combined alpine skiing event be given, where  $x$  is the time measured for downhill and  $y$  is the time measured for slalom.

$Rg(x_i)$	3	1	5	2	4
$Rg(y_i)$	1	3	5	4	2
$(Rg(x_i) - Rg(y_i))^2$	4	4	0	4	4

$$r_{sp} = 1 - \frac{6 \cdot 16}{5(25 - 1)} = 1 - \frac{4}{5} = 0.2$$

Calculation of  $r_{sp}$  in R:

```
load("Example4-5.RData")
cor_SP <- cor(Rg_x5_5, Rg_y5_5, method = "spearman")
cor_SP
```

```
[1] 0.2
```

Does it matter that the rankings in both disciplines are opposed to the time ranks used here?

## Tied ranks

If there are ties, we may replace ranks for identical values by a mean rank of the observations affected. Then we can use:

$$r_{sp} = \frac{\sum_{i=1}^n Rk(x_i) \cdot Rk(y_i) - \frac{1}{n} \sum_{i=1}^n Rk(x_i) \sum_{i=1}^n Rk(y_i)}{\sqrt{\sum_{i=1}^n Rk(x_i)^2 - \frac{1}{n} \left( \sum_{i=1}^n Rk(x_i) \right)^2} \cdot \sqrt{\sum_{i=1}^n Rk(y_i)^2 - \frac{1}{n} \left( \sum_{i=1}^n Rk(y_i) \right)^2}}.$$

This matches the correlation coefficient of Bravais-Pearson for the ranks of the observations (instead of their values).

For contingency tables we use:

$$r_{sp} = \frac{\sum_{j=1}^m \sum_{k=1}^r Rk(x_j)Rk(y_k) n_{jk} - \frac{1}{n} \sum_{j=1}^m Rk(x_j) n_j \cdot \sum_{k=1}^r Rk(y_k) n_{\cdot k}}{\sqrt{\left( \sum_{j=1}^m Rk(x_j)^2 n_j \cdot - \frac{1}{n} \left( \sum_{j=1}^m Rk(x_j) n_j \cdot \right)^2 \right) \cdot \left( \sum_{k=1}^r Rk(y_k)^2 n_{\cdot k} - \frac{1}{n} \left( \sum_{k=1}^r Rk(y_k) n_{\cdot k} \right)^2 \right)}}.$$

## Example 4.6: Scholarships (1)

Two reviewers had to give their assessment of  $n = 50$  students applying for a scholarship. In the final round, the following ratings could be awarded: *excellent* (A), *very good* (B) and *good* (C). The following table contains the results:

		Reviewer II		
		A	B	C
Reviewer I	A	3	2	0
	B	1	12	2
	C	0	4	26

Due to the large number of ties, a specification of the mean ranks is required at first.

For reviewer I, we have:

Rating A: Rank 1 – 5, 5 times a mean rank of 3

Rating B: Rank 6 – 20, 15 times a mean rank of 13

Rating C: Rank 21 – 50, 30 times a mean rank of 35.5

## Example 5.6: Scholarships (2)

Analogously, we get the mean ranks 2.5, 13.5 and 36.5 for Reviewer II.

Calculation in R:

```
load("Example4-6.RData")
Rg_x5_6_mean <- c(3.0, 13.0, 35.5)
Rg_y5_6_mean <- c(2.5, 13.5, 36.5)
```

Finally, using the formula for contingency tables, we get:

$$r_{sp} = \frac{38797.5 - \frac{1}{50} \cdot 1275 \cdot 1275}{\sqrt{7875 \cdot 8096}} = \frac{6285}{7984.735} = 0.7871.$$

## Example 4.6: Scholarships (3)

Calculation of  $r_{sp}$  in R:

```
n <- sum(n_j_k)

Rx_Ry_sum <- sum(n_j_k[1, ] * Rg_x5_6_mean[1] * Rg_y5_6_mean) +
  sum(n_j_k[2, ] * Rg_x5_6_mean[2] * Rg_y5_6_mean) +
  sum(n_j_k[3, ] * Rg_x5_6_mean[3] * Rg_y5_6_mean)

Rx_sum <- sum(n_j_k * Rg_x5_6_mean)
Ry_sum <- sum(t(n_j_k) * Rg_y5_6_mean)
Rx_2 <- sum(Rg_x5_6_mean^2 * margin.table(n_j_k, 1))
Ry_2 <- sum(Rg_y5_6_mean^2 * margin.table(n_j_k, 2))

cor_SP <- (Rx_Ry_sum - 1/n * Rx_sum * Ry_sum) /
  (sqrt(Rx_2 - 1/n * Rx_sum^2) *
  sqrt(Ry_2 - 1/n * Ry_sum^2))

round(cor_SP, 4)

[1] 0.7871
```

## Example 4.6: Scholarships (4)

Alternatively, we could use individual data on ranks and the first formula to reach the same result ( $Rk(x_i); Rk(y_i)$ ):

$\underbrace{(3; 2.5)}_{3x}$	$\underbrace{(3; 13.5)}_{2x}$	$\underbrace{(3; 36.5)}_{0x}$
$\underbrace{(13; 2.5)}_{1x}$	$\underbrace{(13; 13.5)}_{12x}$	$\underbrace{(13; 36.5)}_{2x}$
$\underbrace{(35.5; 2.5)}_{0x}$	$\underbrace{(35.5; 13.5)}_{4x}$	$\underbrace{(35.5; 36.5)}_{26x}$

## Preliminary remarks on $\tau$ and $\gamma$ measures

We define the following two relationship patterns for two pairs of values  $(x_i; y_i)$  and  $(x_j; y_j)$ :

**Concordant pair:** We have  $x_i < x_j$  and  $y_i < y_j$  or  $x_i > x_j$  and  $y_i > y_j$ , so that the comparison of the components of the pairs is unidirectional.

**Discordant pair:** We have  $x_i < x_j$  and  $y_i > y_j$  or  $x_i > x_j$  and  $y_i < y_j$ , so that the comparison of the components of the pairs is counterdirectional.

The number of concordant and discordant pairs is labelled  $n_c$  and  $n_d$ , respectively.

Additionally, we may have to take ties into account.  $T_x$  is the number of ties of the first variable and  $T_y$  is the number of ties of the second variable.

# Kendall's $\tau$ and Goodman and Kruskal's $\gamma$

We have:

$$\tau_a = \frac{n_c - n_d}{\frac{1}{2} \cdot n \cdot (n - 1)}$$

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + T_x) \cdot (n_c + n_d + T_y)}}$$

$$\tau_c = \frac{n_c - n_d}{\frac{1}{2} \cdot n^2 \cdot \frac{M-1}{M}} = \frac{2M \cdot (n_c - n_d)}{n^2 \cdot (M - 1)}$$

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

$\tau_a$  requires contingency tables without ties.  $\tau_b$  is commonly used, but only takes on the value 1 for quadratic contingency tables.  $\tau_c$  accounts for differing numbers of rows and columns.

## Example 4.7: see Ex. 4.6 (1)

In order to calculate the  $\tau$  and *gamma* measures, we have to find the concordant and discordant pairs as well as the ties in  $X$  and  $Y$ . We get:

$$n_c = 3(12 + 2 + 4 + 26) + 2(2 + 26) + 1(4 + 26) + 12 \cdot 26 = 530$$

$$n_d = 2(1 + 0) + 0 + 12 \cdot 0 + 2(0 + 4) = 10$$

Calculation of  $n_c$  and  $n_d$  in R:

```
load("Example4-6.RData")
```

```
nc <- sum(n_j_k[1,1]*n_j_k[2:3,2:3], n_j_k[1,2]*n_j_k[2:3,3],  
          n_j_k[2,1]*n_j_k[3,2:3], n_j_k[2,2]*n_j_k[3,3])
```

```
nd <- sum(n_j_k[3,1]*n_j_k[1:2,2:3], n_j_k[2,1]*n_j_k[1,2:3],  
          n_j_k[3,2]*n_j_k[1:2,3], n_j_k[3,1]*n_j_k[2,2])
```

nc

[1] 530

nd

[1] 10

## Example 4.7: see Ex. 4.6 (2)

For the ties, we get:

$$T_x = 3(2 + 0) + 1(12 + 2) + 0 + 0 + 12 \cdot 2 + 4 \cdot 26 = 148$$

$$T_y = 3(1 + 0) + 2(12 + 4) + 0 + 0 + 12 \cdot 4 + 2 \cdot 26 = 135$$

Calculation of  $T_x$  and  $T_y$  in R:

```
Tx <- sum(n_j_k[,1]*n_j_k[,2:3], n_j_k[,2]*n_j_k[,3])
Ty <- sum(t(n_j_k)[,1] * t(n_j_k)[,2:3],
          t(n_j_k)[,2] * t(n_j_k)[,3])
```

Tx

```
[1] 148
```

Ty

```
[1] 135
```

## Example 4.7: see Ex. 4.6 (3)

We finally reach:

$$\tau_a = \frac{530-10}{\frac{1}{2} \cdot 50 \cdot (50-1)} = 0.4245$$

$$\tau_b = \frac{530-10}{\sqrt{(530+10+148) \cdot (530+10+135)}} = 0.7631$$

$$\tau_c = \frac{530-10}{\frac{1}{2} \cdot 50^2 \cdot \frac{3-1}{3}} = 0.624$$

$$\gamma = \frac{530-10}{530+10} = 0.9630$$

In Example 4.6, the result was  $r_{sp} = 0.7871$ .

## Example 4.7: see Ex. 4.6 (4)

Calculation of  $\tau_a$ ,  $\tau_b$ ,  $\tau_c$  and  $\gamma$  in R:

```
n <- sum(n_j_k)
M <- min(dim(n_j_k))

tau_a <- (nc - nd) / (1/2 * n * (n - 1))
tau_b <- (nc - nd) / sqrt((nc + nd + Tx) * (nc + nd + Ty))
tau_c <- 2 * M * (nc - nd) / (n^2 * (M - 1))
gamma <- (nc - nd) / (nc + nd)
```

```
round(tau_a, 4)
```

```
[1] 0.4245
```

```
round(tau_b, 4)
```

```
[1] 0.7631
```

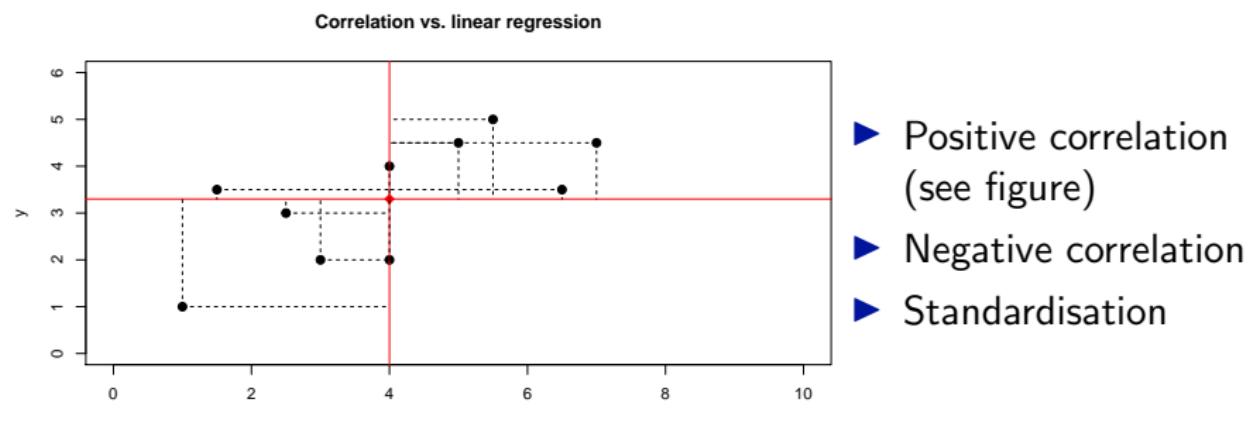
```
round(tau_c, 4)
```

```
[1] 0.624
```

```
round(gamma, 4)
```

```
[1] 0.963
```

# Correlation and covariance



The covariance of two metric variables  $x$  and  $y$  is given by

$$s_{xy}^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Some kind of standardisation is needed!

# Correlation coefficient of Bravais-Pearson (1)

For metrically scaled variables  $x$  and  $y$  with positive variances of  $x$  and  $y$ , the correlation coefficient of Bravais-Pearson is defined as

$$r_{xy} = \frac{s_{xy}^*}{s_x^* \cdot s_y^*} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

# Correlation coefficient of Bravais-Pearson (2)

Properties:

1.  $-1 \leq r_{xy} \leq 1$
2. Special cases:  $r_{xy} = -1; 0; +1$
3. Transformation

$$u_i = a_0 + a_1 \cdot x_i; v_i = b_0 + b_1 \cdot y_i$$

$$r_{uv} = \text{sgn}(a_1 \cdot b_1) \cdot r_{xy}$$

Frequently, only the algebraic sign ( $r_{xy} \gtrless 0$ ) is of interest in economics.

# Problems of the correlation coefficient (1)

## Non-linear relationships

Let  $x_i = -2; -1; 1; 2$  and  $y_i = x_i^2$ . The resulting correlation coefficient is

$$r_{xy} = 0 \quad .$$

There is a quadratic relationship in the data, which is not comprehended by the correlation coefficient!

## Problems of the correlation coefficient (2)

### Correlation and causality

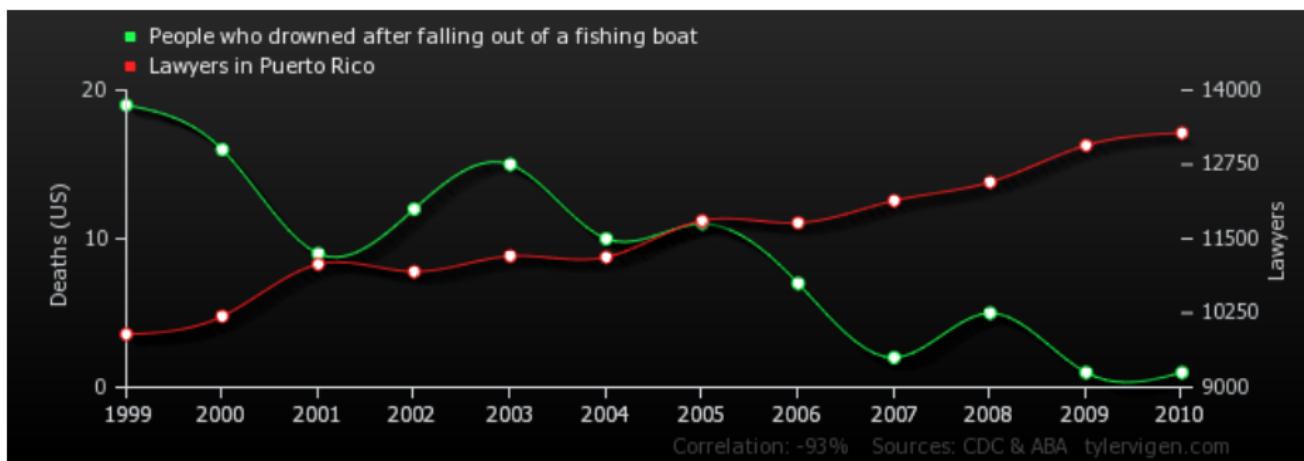
Interpretation of  $r_{xy} = 0.98$ :

A statistical interrelation does not necessarily indicate a theoretical interrelation.

- ▶ Presidential elections / Superbowl in the USA  
(<http://www.theguardian.com/sport/blog/2012/feb/01/super-bowl-ology-science-impotence-2012>)
- ▶ Beer consumption / Count of unemployed people per month

→ Spurious correlation

# Spurious correlation



Source:

Tyler Vigen – spurious correlations (2019).

[http://tylervigen.com/view\\_correlation?id=30074](http://tylervigen.com/view_correlation?id=30074)

## Example 4.8: Correlation coefficient (1)

The following table contains information on  $n = 10$  units and two variables:

$i$	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i \cdot y_i$
1	1	1	1	1	1
2	1.5	3.5	2.25	12.25	5.25
3	2.5	3	6.25	9	7.5
4	3	2	9	4	6
5	4	2	16	4	8
6	4	4	16	16	16
7	5	4.5	25	20.25	22.5
8	5.5	5	30.25	25	27.5
9	6.5	3.5	42.25	12.25	22.75
10	7	4.5	49	20.25	31.5
$\sum$	40	33	197	124	148

## Example 4.8: Correlation coefficient (2)

Data input in R:

```
x4_8 <- c(1, 1.5, 2.5, 3, 4, 4, 5, 5.5, 6.5, 7)  
y4_8 <- c(1, 3.5, 3.2, 2, 4, 4.5, 5, 3.5, 4.5)
```

At first, the univariate measures needed are computed . . .

$$\bar{x} = 40/10 = 4$$

$$\bar{y} = 33/10 = 3.3$$

$$s_x^{*2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \quad s_x^{*2} = \frac{1}{10} \cdot 197 - 4^2 = 3.7$$

$$s_y^{*2} = \frac{1}{10} \cdot 124 - 3.3^2 = 1.51$$

## Example 4.8: Correlation coefficient (3)

... thereupon, these are combined in order to determine a result for the covariance and correlation coefficient, respectively:

$$s_{xy}^* = \frac{1}{10} \cdot 148 - 4 \cdot 3.3 = 1.6$$

$$r_{xy} = \frac{1.6}{\sqrt{3.7 \cdot 1.51}} = 0.6769$$

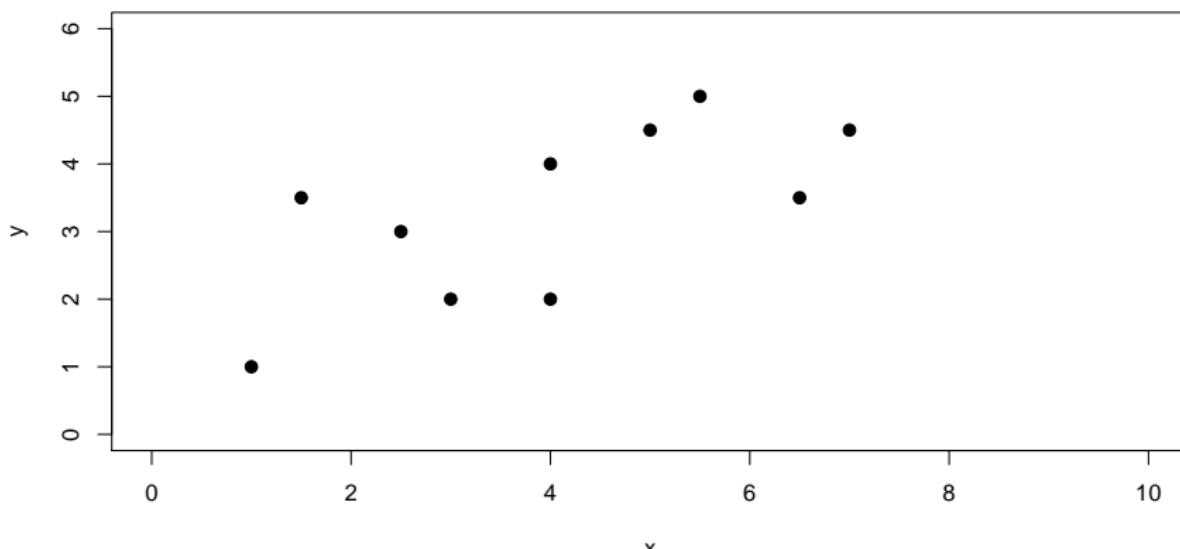
Calculation of  $s_{xy}^*$  and  $r_{xy}$  in R:

```
n <- length(x4_8)
x_y_cov <- (n-1)/n * cov(x4_8, y4_8)
cor_BP <- cor(x4_8, y4_8, method = "pearson")
x_y_cov
[1] 1.6
round(cor_BP, 4)
[1] 0.6769
```

```
x_1_var <- (length(x_1) - 1) / length(x_1) * var(x_1)
x_2_var <- (length(x_2) - 1) / length(x_2) * var(x_2)
```

## Example 4.8: Correlation coefficient (4)

Correlation vs. linear regression

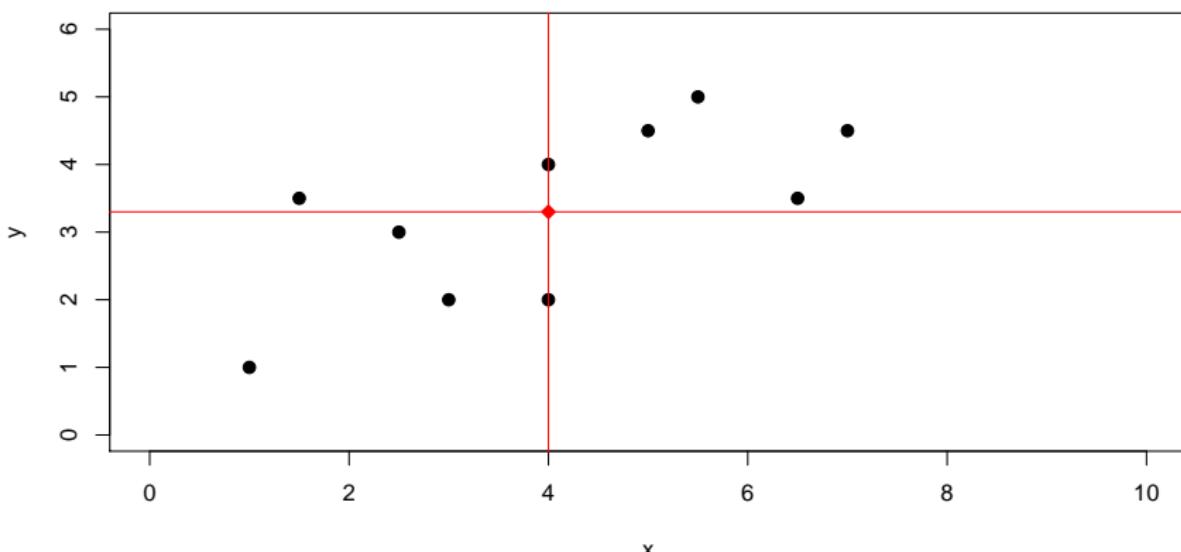


In R:

```
plot(x4_8,y4_8,xlim=c(0,10),ylim=c(0,6),xlab="x",ylab="y",  
type="p",pch=16)
```

## Example 4.8: Correlation coefficient (5)

Correlation vs. linear regression

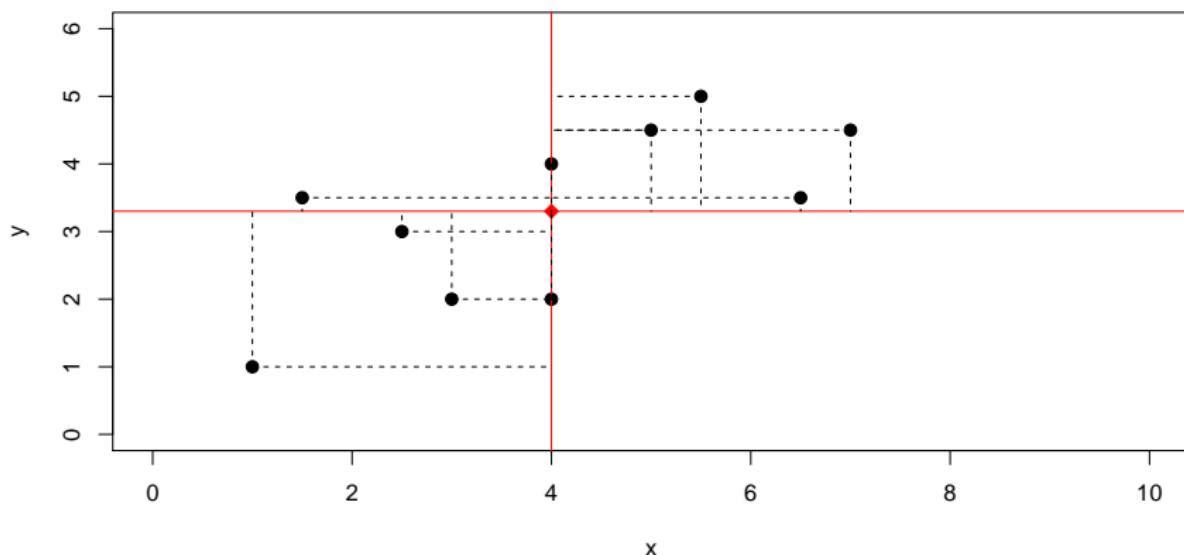


In R:

```
abline(v = mean(x4_8), col = "red")
abline(h = mean(y4_8), col = "red")
points(x = mean(x4_8), y = mean(y4_8), col = "red", pch=19)
```

## Example 4.8: Correlation coefficient (6)

Correlation vs. linear regression



# Linear regression

$x$  and  $y$  are continuous metric variables.

$x$  is the so-called independent variable.

$y$  is the so-called dependent variable.

We are looking for a relationship:

$$y = f(x).$$

→ Dependency analysis

Linear relationships are of primary interest:  $y = a + b \cdot x$ .

Problem: The observations typically do not lie on a line. Why is that the case?

## Simple linear regression

We assume a linear model:

$$Y = \alpha + \beta \cdot X \quad .$$

There might be more than one value of  $y$  that is corresponding to a certain value of  $x$  (random error). We use capital letters when talking about models.

We would like to determine the parameters  $a$  and  $b$  of

$$\hat{y}_i = a + b \cdot x_i \quad ,$$

where  $\hat{y}_i$  is the vertical projection of  $y_i$  to the regression line.  $e_i = y_i - \hat{y}_i$  is the residual corresponding to observation  $x_i$ . The method of ordinary least squares (OLS) determines estimates for the parameters  $a$  and  $b$ :

$$Z(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 \rightarrow \min$$

# Solution to minimisation problem

Using the normal equations

$$n \cdot a + b \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (1\text{st normal equation})$$

$$a \cdot \sum_{i=1}^n x_i + b \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i \quad (2\text{nd normal equation})$$

we get

$$b = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}^*}{s_x^{*2}}$$

and  $a = \bar{y} - b \cdot \bar{x}$ . Finally, for the sample regression line we have:

$$\hat{y} = \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot (x - \bar{x}) \quad .$$

# Coefficient of determination and properties of OLS

The measure  $r_{xy}^2 = \frac{s_{\hat{y}}^2}{s_y^2}$  is called coefficient of determination. For simple linear regression, it equals the squared correlation coefficient of Bravais-Pearson (for  $x$  and  $y$ ). The special cases of  $r_{xy}^2 = 0$  and  $r_{xy}^2 = 1$  are particularly interesting.

Properties:

1. Centre of gravity:  $(\bar{x}, \bar{y})$  is a point on the sample regression line.
2. The residuals cancel each other out:  $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$ .
3.  $b = r_{xy} \cdot \frac{s_y^*}{s_x^*}$   $\left( = \frac{s_{xy}^*}{s_x^* \cdot s_y^*} \cdot \frac{s_y^*}{s_x^*} = \frac{s_{xy}^*}{s_x^{*2}} \right)$

## Example 4.9: see Ex. 4.8 (1)

We get:

$$b = \frac{1.6}{3.7} = 0.6769 \cdot \sqrt{\frac{1.51}{3.7}} = 0.4324$$

$$a = 3.3 - 0.4324 \cdot 4 = 1.5703 \quad \hat{y}(8) = 1.5703 + 0.4324 \cdot 8 = 5.0297$$

$$r^2 = 0.6769^2 = 0.4582 \quad \hat{y}(9) = 1.5703 + 0.4324 \cdot 9 = 5.4622$$

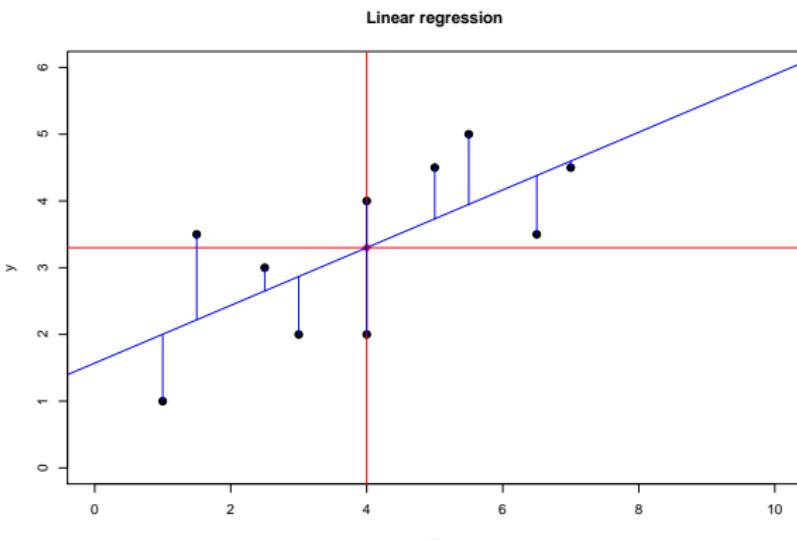
Calculations in R:

```
reg_mod <- lm(y4_8 ~ x4_8)

a <- summary(reg_mod)$coeff[1,1]
b <- summary(reg_mod)$coeff[2,1]
r_2 <- summary(reg_mod)$r.squared
y_hat_8 <- a + b * 8
y_hat_9 <- a + b * 9

round(a, 4) round(b, 4) round(r_2, 4) y_hat_8      y_hat_9
[1] 1.5703      [1] 0.4324      [1] 0.4582      [1] 5.02973     [1] 5.462162
```

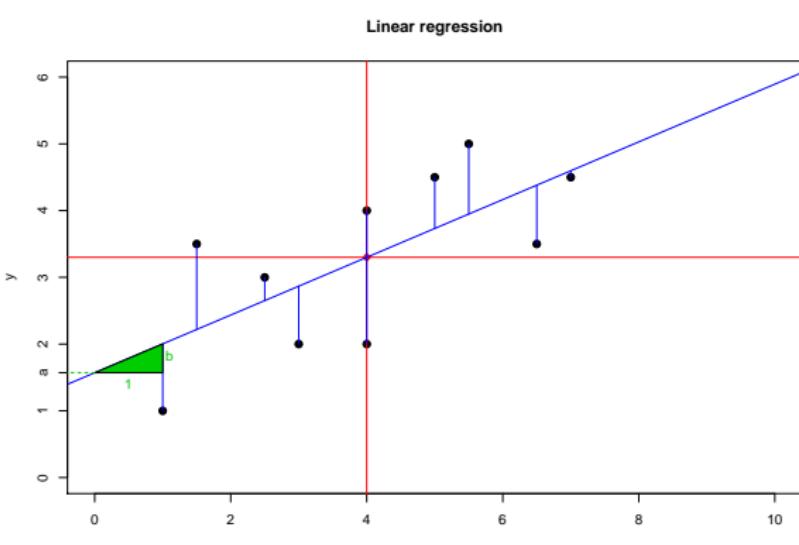
## Example 4.9: see Ex. 4.8 (2)



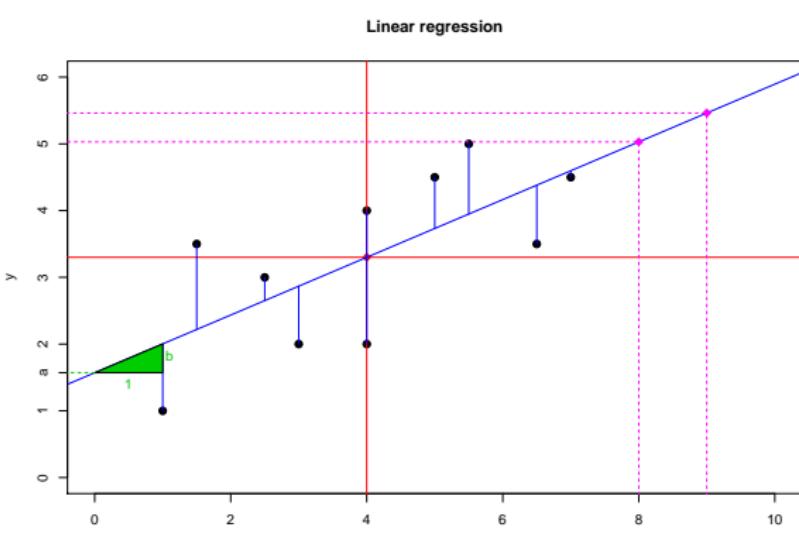
In R:

```
plot(x4_8, y4_8, xlim=c(0,10), ylim=c(0,6), type="p", pch=16)
abline(v = mean(x4_8), col = "red")
abline(h = mean(y4_8), col = "red")
points(x = mean(x4_8), y = mean(y4_8), col = "red", pch=19)
abline(reg_mod, col="blue")
```

## Example 4.9: see Ex. 4.8 (3)



# Example 4.9: see Ex. 4.8 (4)



# Elements of Statistics

## Chapter 5: Probability theory

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

## Example 5.1: Some experiments

- a) Rolling a dice
  - b) Playing the lottery
  - c) Duration between two computer malfunctions
  - d) Kilometer reading of a car
  - e) Burning life of a bulb
  - f) Turnover of a pharmacy on a Friday
  - g) Life span of a male live birth
  - h) Time to solve a problem
- ▶ What are the outcomes of these experiments?
  - ▶ Are these experiments reproducible?

# Experiment and sample space

## Definition 5.1: Experiment

*Procedures*, which can actually or at least theoretically be repeated under a constant set of conditions and the outcomes of which cannot be forecast precisely, are called experiments.

## Definition 5.2: Sample space

The set of all possible, mutually exclusive outcomes of an experiment is called sample space  $\Omega$ .

Sample spaces may be:

- ▶ Finite
- ▶ Infinite
  - ▶ (Un)countable

## Example 5.2:

One roll of a dice:

- a)  $\Omega = \{i \mid i \in \mathbb{N}; 1 \leq i \leq 6\}$   
 $i$  is number of pips
- b)  $\Omega = \{\omega_g, \omega_u\}$   
Even / uneven number of pips
- c)  $\Omega = \{\{1, 2, 3\}, \{4, 5\}, \{6\}\}$

## Example 5.3:

Burning life of a bulb:

- a)  $\Omega = \{x \mid x \in \mathbb{R}; 0 \leq x \leq 10,000\}$   
*Infinite accuracy of measurement*  
 $\rightarrow$  continuous variable
- b)  $\Omega = \{x \mid x \in \mathbb{N}_0, x \leq 10,000\}$   
*Measurement in hours*  
 $\rightarrow$  discrete variable

## Definition 5.3:

Each subset of a sample space is called event.

We write:

$A, B, C \dots$

We say:

An experiment with sample space  $\Omega$  yielded event  $A$  ( $B, C, \dots$ ).

## Example 5.4:

One roll of a dice  $\Omega = \{1, 2, 3, 4, 5, 6\}$ :

$$A = \{1, 3, 5\} \quad B = \{2, 4, 6\} \quad C = \{3\}$$

Application in R:

```
Omega <- 1:6  
  
A <- Omega[c(1,3,5)] ; B <- Omega[c(2,4,6)] ; C <- Omega[3]
```

**Bear in mind:** Set theory and its calculation rules

## Example 5.5: see Ex. 5.4 (1)

$$B = \overline{A}$$

```
A_Bar <- Omega[-A]
setequal(x = B, y = A_Bar)
[1] TRUE
```

$$\overline{C} = \{1, 2, 4, 5, 6\}$$

```
C_Bar <- Omega[-C]
C_Bar
[1] 1 2 4 5 6
```

## Example 5.5: see Ex. 5.4 (2)

$$A \cap B = \emptyset$$

```
intersect(x = A, y = B)  
integer(0)
```

$$B \cup C = \{2, 3, 4, 6\}$$

```
sort(union(x = B, y = C))  
[1] 2 3 4 6
```

Events may be characterised using *admissible questions*.

→ System of events (*set of admissible questions*)

# Event space

## Definition 5.4:

Let  $\Omega$  be a sample space and let  $\mathcal{K}$  be a non-empty subset of the power set of the sample space  $\mathcal{P}(\Omega)$ .  $\mathcal{K}$  is called an event space if the following properties hold:

I:  $\emptyset \in \mathcal{K}$

II: If  $A \in \mathcal{K}$ , then  $\overline{A} \in \mathcal{K}$

III: If  $A \in \mathcal{K}$  and  $B \in \mathcal{K}$ , then  $A \cup B \in \mathcal{K}$

We may call  $\mathcal{K}$  an algebra over  $\Omega$  as well.

## Example 6.6:

$$\mathcal{K}_1 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$$

is an event space.

$$\mathcal{K}_2 = \{\emptyset, \{1, 2\}, \{3, 4\}, \{5, 6\}, \Omega\}$$

is not an event space.

$\mathcal{P}(\Omega)$  is always an event space (containing  $2^n$  elements).

# Sigma algebra

## Definition 5.5:

Let  $\Omega$  be a sample space. A non-empty subset  $\mathcal{S}$  of  $\mathcal{P}(\Omega)$  is called sigma algebra if the following holds:

I:  $A \in \mathcal{S} \Rightarrow \bar{A} \in \mathcal{S}$  (closed under complementation)

II:  $A_i \in \mathcal{S}$  for all  $i = 1, 2, \dots$

$\Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{S}$  (closed under countable union)

- ▶ Every countable union of events itself is part of  $\mathcal{S}$ .
- ▶ All set operations can be derived from these two set operations

## Definition 5.6:

Let  $\mathcal{K}$  be an event space over  $\Omega$ . The (set) function  $P : \mathcal{K} \rightarrow \mathbb{R}$  is called a *probability content*, if the following holds:

1. If  $A \in \mathcal{K}$ , then  $P(A) \geq 0$
2. If  $A, B \in \mathcal{K}$  and  $A \cap B = \emptyset$ , then

$$P(A \cup B) = P(A) + P(B)$$

3.  $P(\Omega) = 1$

# Implications

1.  $P(\bar{A}) = 1 - P(A)$
2.  $P(A \cap B)$  follows from Definition 6.6.
3. Finite unions follow from Definition 6.6.
4. Countable unions do not follow from Definition 6.6  
(e. g. even numbers within the natural numbers).

## Definition 5.7:

Let  $\mathcal{S}$  be a sigma algebra over a sample space  $\Omega$ . A (set) function  $P : \mathcal{S} \rightarrow \mathbb{R}$  is called *probability measure*, if the following holds:

1. If  $A \in \mathcal{S}$ , then  $P(A) \geq 0$  **(Non-negativity)**
2. If  $A_i \in \mathcal{S}$ ,  $i = 1, 2, \dots$  and all  $A_i$  are pairwise disjoint, we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

**( $\sigma$ -additivity)**

3.  $P(\Omega) = 1$  **(Standardisation)**

Bear in mind:

*Pairwise disjoint* and *overall disjoint* have to be distinguished!

# Probability measure and implications

## Definition 5.8:

A triple  $(\Omega; \mathcal{S}; P)$  consisting of a sample space  $\Omega$ , a sigma algebra  $\mathcal{S}$  over  $\Omega$  and a probability measure  $P$  over  $\mathcal{S}$  is called *probability space*.

## Further implications

1.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2.  $P(A) \leq 1$
3.  $P(\emptyset) = 0$
4.  $A \subset B \quad \Rightarrow \quad P(A) \leq P(B)$
5. 
$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

## Classical or Laplace's concept - Principle of symmetry

- ▶ Finite sample space  $\Omega$
- ▶ Elementary events (outcomes) have equal probability of occurrence

$\nu(\Omega)$  is the number of elementary events and  $\nu(A)$  is the number of *favourable* cases.

Then

$$P(A) = \frac{\nu(A)}{\nu(\Omega)}$$

is the probability that event  $A$  occurs (number of favourable cases divided by number of possible cases).

## Example 5.7: Two rolls of a dice (1)

Two rolls of a dice:  $\Omega = \{(i,j) | i, j \in \mathbb{N} \text{ mit } 1 \leq i, j \leq 6\}$

There are 36 possible events.

Application in R:

```
Omega <- expand.grid(Dice1 = 1:6, Dice2 = 1:6)
head(Omega, n = 9)
```

	Dice1	Dice2
1	1	1
2	2	1
3	3	1
4	4	1
5	5	1
6	6	1
7	1	2
8	2	2
9	3	2

```
length(Omega[,1])
[1] 36
```

## Example 5.7: Two rolls of a dice (2)

a) Sum of pips is **at least** 10:

$$A = \{(4; 6); (5; 5); (5; 6); (6; 4); (6; 5); (6; 6)\}$$

$$W(A) = \frac{\nu(A)}{36} = \frac{6}{36} = \frac{1}{6}$$

Calculations in R:

```
Sum_of_pips <- apply(Omega, 1, sum)
Omega <- cbind(Omega, Sum_of_pips)
A <- Omega[Omega$Sum_of_pips >= 10, ]
A
```

	Dice1	Dice2	Sum_of_pips
24	6	4	10
29	5	5	10
30	6	5	11
34	4	6	10
35	5	6	11
36	6	6	12

```
length(A[,1])/length(Omega[,1])
[1] 0.1666667
```

## Example 5.7: Two rolls of a dice (3)

b) Sum of pips is **exactly** 4:

$$B = \{(1; 3); (2; 2); (3; 1)\}$$

$$W(B) = \frac{3}{36} = \frac{1}{12}$$

Calculations in R:

```
B <- Omega[Omega$Sum_of_pips == 4, ]
```

```
B
```

	Dice1	Dice2	Sum_of_pips
3	3	1	4
8	2	2	4
13	1	3	4

```
length(B[,1])/length(Omega[,1])
```

```
[1] 0.08333333
```

# Statistical concept of probability - Principle of frequency

- ▶ Experiments can be arbitrarily repeated
- ▶  $A$  as an event in the experiment
- ▶  $n$  repetitions (independent trials)

$p_n(A)$  is the relative frequency of occurrences of event  $A$  in  $n$  repeated experiments.

Properties: Non-negativity, additivity and standardisation!

Then  $P(A) = \lim_{n \rightarrow \infty} p_n(A)$ .

Law of large numbers (see later).

# Subjectivistic concept of probability

- ▶ Probability of rain
- ▶ Investment in shares
- ▶ Risk of an accident for a certain new car

Subjectivistic determination:

- ▶ Expert knowledge
- ▶ Experience
- ▶ Intuition

Verifiability is a problem here.

## Example 5.8: Probability of survival

We determine the *probability* for a 50 year old man to survive the next year using the life table 2008/2010.

$$p_{50} = 0.995868.$$

Contrary to this:  $\frac{l_{51}}{l_0} = 0.95169$ .

Segmentation possible regarding further information:

- ▶ Person smokes
- ▶ Person has not been ill since 20 years
- ▶ Person has had a heart attack
- ▶ ...

## Example 5.9

We want to determine the probability of obtaining a 6 in a roll of a dice.

We are informed that the experiment yielded an even number of pips.  
Does that change the probability?

$$\text{We have: } P(\{6\}|\{2, 4, 6\}) = \frac{P(\{6\})}{P(\{2, 4, 6\})} = \frac{\frac{1}{6}}{\frac{3}{2}} = \frac{1}{3} .$$

Calculations in R:

```
Prob <- length(6)/length(c(2,4,6))
Prob
[1] 0.3333333
```

## Definition 5.9:

Let  $A$  and  $B$  be two events and let  $P(B) \neq 0$ . Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

is called the probability of  $A$  conditional on the occurrence of event  $B$ .

Analogously:  $P(B|A)$  with  $P(A) \neq 0$ .

The multiplication theorem follows:

$$P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A)$$

## Example 5.10: Evaluation of a course (1)

The evaluation of a course yielded the following frequency table:

	C male	D female	$\Sigma$
A good	0.45	0.35	0.8
B bad	0.15	0.05	0.2
$\Sigma$	0.6	0.4	1

Calculations in R:

```
load("Example5-10.RData")
p_j_k <- addmargins(p_j_k)
p_j_k
```

	C	D	Sum
A	0.45	0.35	0.8
B	0.15	0.05	0.2
Sum	0.60	0.40	1.0

## Example 5.10: Evaluation of a course (2)

Then:

$$P(A) = 0.8$$

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{0.45}{0.6} = 0.75$$

$$P(A|D) = \frac{0.35}{0.4} = \frac{7}{8} = 0.875$$

Calculations in R:

```
Prob_A     <- p_j_k[1,3]
Prob_A_C   <- p_j_k[1,1]/p_j_k[3,1]
Prob_A_D   <- p_j_k[1,2]/p_j_k[3,2]
```

Prob\_A

[1] 0.8

Prob\_A\_C

[1] 0.75

Prob\_A\_D

[1] 0.875

## Definition 5.10:

Let  $A$  and  $B$  be two random events.  $A$  and  $B$  are called stochastically independent if and only if:

$$P(A \cap B) = P(A) \cdot P(B)$$

Then we have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{st. ind.} \quad = \quad \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

## Example 5.11:

A coin is tossed twice (heads or tails). Consider the following two events:

- A:** 1st toss yields heads
- B:** 2nd toss yields tails

Then:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

## Example 5.12: see Ex. 5.10

We want to check if rating and gender are stochastically independent.

Inter alia we should have:

$$P(A \cap C) = P(A) \cdot P(C).$$

But we actually have:

$$P(A) \cdot P(C) = 0.8 \cdot 0.6 = 0.48 \neq 0.45 = P(A \cap C).$$

It follows that A and C are stochastically dependent..

Calculations in R:

```
Prob_C <- p_j_k[3,1]  
Prob_A * Prob_C == 0.45
```

```
[1] FALSE
```

Addendum: What would the joint probabilities be, if the variables were stochastically independent and the marginal distributions were unchanged?

## Example 5.13: Doorknobs (1)

A supplier to the automobile industry produces 10,000 door handles per day on 4 different machines. Production is distributed as follows:

- $M_1$  1000 pieces with 8% scrap,
- $M_2$  2000 pieces with 5% scrap,
- $M_3$  3000 pieces with 3% scrap,
- $M_4$  4000 pieces with 2% scrap.

One door handle is randomly chosen from the daily production. What is the probability that the item is defective?

Data input in R:

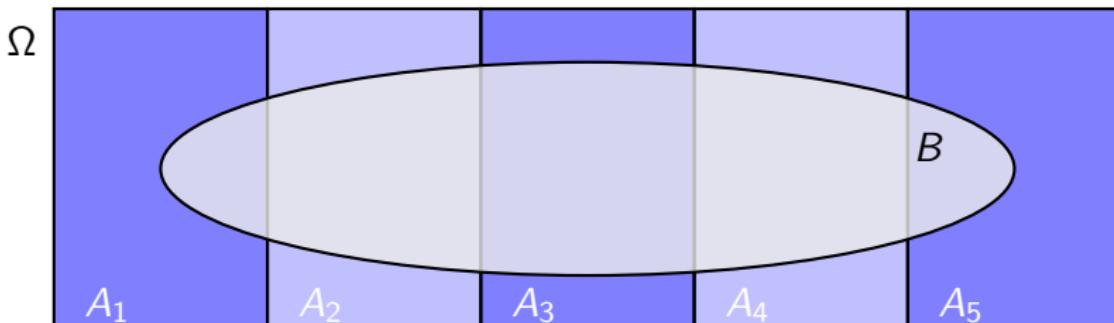
```
Doorknobs_per_machine <- c(1000,2000,3000,4000)
Number_doorknobs
Rejects_per_machine <- c(0.08,0.05,0.03,0.02)
```

## Theorem 5.1:

### Law of total probability:

Let  $A_1, \dots, A_m$  be a disjoint decomposition of  $\Omega$ . Then, for  $B \subset \Omega$  we have:

$$\begin{aligned} P(B) &= \sum_{i=1}^m P(B \cap A_i) \\ &= \sum_{i=1}^m P(B|A_i) \cdot P(A_i) \end{aligned}$$



## Example 5.13: Doorknobs (2)

Let  $A_i$  for  $i = 1, \dots, 4$  be the event that the door handle has been produced on machine  $M_i$ . Let  $F$  be the event that the door handle is faulty. Furthermore, we have:

$$P(A_1) = \frac{1000}{10000} = 0.1, \quad P(F|A_1) = 0.08$$

$$P(A_2) = \frac{2000}{10000} = 0.2, \quad P(F|A_2) = 0.05$$

$$P(A_3) = \frac{3000}{10000} = 0.3, \quad P(F|A_3) = 0.03$$

$$P(A_4) = \frac{4000}{10000} = 0.4, \quad P(F|A_4) = 0.02$$

Calculations in R:

```
P_Ai     <- Doorknobs_per_machine/Number_doorknobs
P_F_Ai  <- Rejects_per_machine
```

## Example 5.13: Doorknobs (3)

The probability for event  $F$  is composed as follows:

$$\begin{aligned} P(F) &= P(A_1) \cdot P(F|A_1) + P(A_2) \cdot P(F|A_2) + P(A_3) \cdot P(F|A_3) + \\ &\quad P(A_4) \cdot P(F|A_4) = \\ &\sum_{i=1}^4 P(A_i) \cdot P(F|A_i). \end{aligned}$$

Therefore, we have:

$$P(F) = 0.1 \cdot 0.08 + 0.2 \cdot 0.05 + 0.3 \cdot 0.03 + 0.4 \cdot 0.02 = 0.035.$$

Calculations in R:

```
P_F <- sum(P_Ai * P_F_Ai)  
P_F  
[1] 0.035
```

## Example 5.14: Spam mails (1)

We want to discuss the nuisance of spam mails. Let the following two events be given:

A: The mail is spam.

B: The mail client marks the mail as spam.

Past experience taught us that roughly 85% of mails are spam. 95% of spam is marked as such by the mail client. But 8% of mails are wrongly marked as spam.

What percentage of mails, which have been marked as spam, is indeed spam?

# Tree diagrams and probabilities

## Multiplication of probabilities

In a multi-stage experiment we get the probability of a single event by multiplying the probabilities on the respective path.

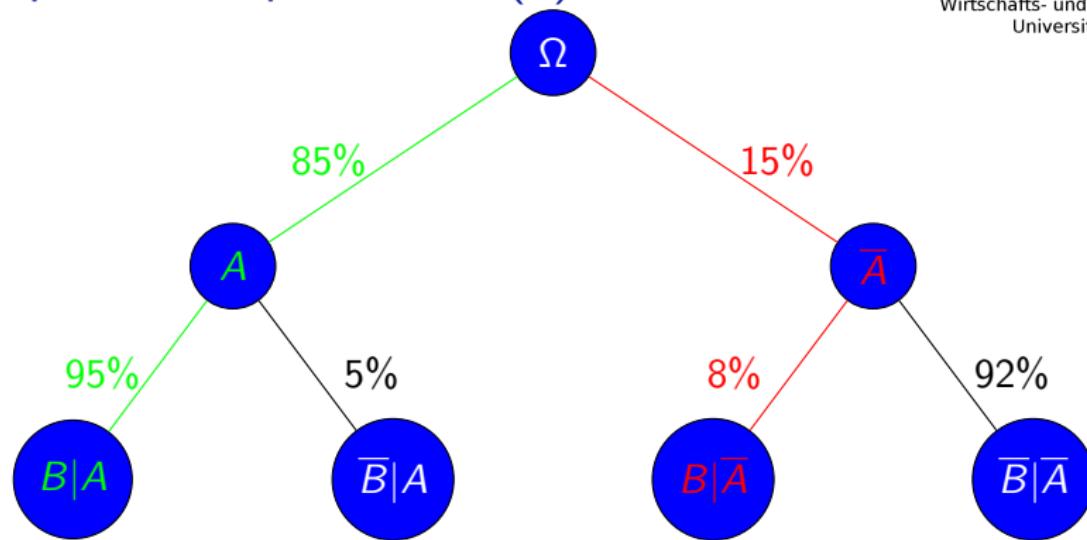
## Addition of probabilities

Probabilities from different paths of a multi-stage experiment are added.

## Total probability

The sum of probabilities at the *leaves* of a tree diagram is 1.

## Example 5.14: Spam mails (2)



We have:  $P(A \cap B) = P(A) \cdot P(B|A)$   
( $P(\overline{A} \cap B)$ ,  $P(A \cap \overline{B})$  and  $P(\overline{A} \cap \overline{B})$  analogously). Overall we have

$$P(A|B) = \frac{0.85 \cdot 0.95}{0.85 \cdot 0.95 + 0.15 \cdot 0.08} = 0.985.$$

## Theorem 5.2: Bayes' theorem

Let  $A_1, \dots, A_m$  be events of a sample space  $\Omega$ . Furthermore, let the events  $A_i$  ( $i = 1, \dots, m$ ) be pairwise disjoint and let

$$\Omega = \bigcup_{i=1}^m A_i. \quad (\text{Decomposition of } \Omega)$$

Now, let  $B$  be another event and the probability for all positive events. Then we have for all  $k = 1, \dots, m$ :

$$P(A_k|B) = \frac{P(A_k) \cdot P(B|A_k)}{\sum_{i=1}^m P(A_i) \cdot P(B|A_i)}.$$

We call  $P(A_i)$  the a-priori-probability and  $P(A_i|B)$  the a-posteriori-probability.

## Example 5.15: see Ex. 5.13 (1)

A quality control inspector of the supplier to the automobile industry has randomly sampled and inspected a doorknob where he detects a flaw. Now, he wants to know the probabilities for the flawed doorknob to be produced by machine  $M_1$ ,  $M_2$ ,  $M_3$  or  $M_4$ . According to *Bayes' theorem* we obtain the respective probabilities as follows:

$$P(A_1|F) = \frac{P(A_1) \cdot P(F|A_1)}{\sum_{i=1}^4 P(A_i) \cdot P(F|A_i)} = \frac{0.1 \cdot 0.08}{0.035} = 0.229$$

$$P(A_2|F) = \frac{P(A_2) \cdot P(F|A_2)}{\sum_{i=1}^4 P(A_i) \cdot P(F|A_i)} = \frac{0.2 \cdot 0.05}{0.035} = 0.286$$

## Example 5.15: see Ex. 5.13 (2)

According to *Bayes' theorem* we obtain the other searched-for probabilities:

$$P(A_3|F) = \frac{0.3 \cdot 0.03}{0.035} = 0.257$$

$$P(A_4|F) = \frac{0.4 \cdot 0.02}{0.035} = 0.229$$

Calculations in R:

```
P_Ai_F <- P_Ai * P_F_Ai / P_F
round(P_Ai_F, 3)
[1] 0.229 0.286 0.257 0.229
```

## Example 5.16: Missing probabilities (1)

$A_1$ ,  $A_2$  and  $A_3$  are three disjoint events that unite to  $\Omega$ . Let  $P(A_1) = 0.3$ ,  $P(A_2) = 0.5$ ,  $P(B|A_1) = 0.6$ ,  $P(B|A_2) = 0.5$  and  $P(B|A_3) = 0.1$ .

Data input in R:

```
P_Ai     <- c(0.3, 0.5, NA)
P_B_Ai  <- c(0.6, 0.5, 0.1)
```

We calculate  $P(A_1|B)$ .

First we have  $P(A_3) = 1 - 0.3 - 0.5 = 0.2$ .

Calculations in R:

```
P_Ai[3] <- 1 - P_Ai[1] - P_Ai[2]
P_Ai
[1] 0.3 0.5 0.2
```

## Example 5.16: Missing probabilities (2)

$$P(A_1|B) = \frac{0.3 \cdot 0.6}{0.3 \cdot 0.6 + 0.5 \cdot 0.5 + 0.2 \cdot 0.1} = \frac{0.18}{0.45} = \frac{2}{5} = 0.4$$

Calculations in R:

```
P_A1_B <- P_Ai[1] * P_B_Ai[1] / sum(P_Ai * P_B_Ai)  
P_A1_B
```

```
[1] 0.4
```

## Example 5.17: Accidents

We have the following notation:

W: Woman

M: Man

A: Accident

The probability for an accident is  $P(A) = 1/100,000$ . Furthermore  $P(W|A) = 1/8$  and  $P(M|A) = 7/8$ .

Can we conclude from this that women are better drivers because they cause less accidents?

# Urn model with replacement (WR)

In an urn there are  $N$  different balls.

From there we take  $n$  balls WR.

- ▶ Multiple draws possible
- ▶ **Independence** of draws

There are  $N^n$  variations.

Calculation of the number of variations in R:

```
# Attention: N and n must be defined
Number_of_variations <- N^n
```

Random draws following an urn model WR in R:

```
set.seed(123) # starting point for (pseudo-) random number
# generator
draws <- sample(x = 1:100, size = 3, replace = TRUE)
draws
[1] 31 79 51
```

## Example 5.18: WR

In an urn there are 100 balls with the numbers 1 to 100. We draw three times one ball, note the number and return it into the urn.

$$P(K_1 = 31; K_2 = 79; K_3 = 51) = \frac{1}{100} \cdot \frac{1}{100} \cdot \frac{1}{100} = 10^{-6}$$

# Urn model without replacement (WoR)

In an urn there are  $N$  different balls.

From this we draw  $n$  balls WoR.

- ▶ Multiple draws not possible (number of balls decreases)
- ▶ **Dependence** of the draws

There are  $N \cdot (N - 1) \cdot \dots \cdot (N - n + 1) = \frac{N!}{(N - n)!}$  variations.

Calculation of the number of variations in R:

```
# Attention: N and n must be defined
Number_of_variations <- factorial(N)/factorial(N-n)
```

Random draws following the urn model WoR in R:

```
set.seed(321) # new starting point
draws <- sample(x = 1:100, size = 3, replace = FALSE)
draws
[1] 54 77 88
```

## Example 5.19: WoR

In an urn there are 100 balls with the numbers 1 to 100. We draw three times one ball and note the number.

$$\begin{aligned} P(K_1 = 54; K_2 = 77; K_3 = 88) &= \\ W(K_1 = 54) \cdot P(K_2 = 77 | K_1 = 54) \cdot P(K_3 = 88 | K_1 = 54, K_2 = 77) &= \\ \frac{1}{100} \cdot \frac{1}{99} \cdot \frac{1}{98} \end{aligned}$$

For  $n/N < 0,05$  there is only a small numerical difference.

## More urn models

- WoR without ordering:

As before, but here  $n!$  orders of the balls are identical.

There are  $\frac{N!}{(N-n)! \cdot n!} = \binom{N}{n}$  combinations.

Lottery as example:  $\binom{49}{6} = 13,983,816$

Determination in R:

```
Number_of_variations <- choose(49, 6)
```

- WR without ordering:

There are  $\binom{N+n-1}{n}$  combinations

Determination in R:

```
# Attention: N and n must be defined
Number_of_variations <- choose(N+n-1, n)
```

# Elements of Statistics

## Chapter 6: Random variables

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

## Example 6.1: Triple coin toss (1)

Triple coin toss (H: Heads, T: Tails):

Let  $X = \text{Number of heads}$  be a random variable with  $x \in \{0, 1, 2, 3\}$ .

Toss	Result			Realisation			Number of heads
1	H	H	H	1	1	1	3
2	H	H	T	1	1	0	2
3	H	T	H	1	0	1	2
4	H	T	T	1	0	0	1
5	T	H	H	0	1	1	2
6	T	H	T	0	1	0	1
7	T	T	H	0	0	1	1
8	T	T	T	0	0	0	0

## Example 6.1: Triple coin toss (2)

Construction of the table in R:

```
result <- expand.grid(lapply(
    X = 1:3,
    FUN = function(x) c("H", "T")))
domain <- expand.grid(lapply(
    X = 1:3,
    FUN = function(x) c(1, 0)))
number_of_hats <- rowSums(domain)

Example6_1<- data.frame(result, domain, number_of_hats)
names(Example6_1)<- c("result", "", "", "domain", "", "", "",
                      "number_of_hats")

save(Example6_1, file="Example6-1.RData")
head(Example6_1, n = 4)
```

	result	domain	number_of_hats
1	H H H	1 1 1	3
2	T H H	0 1 1	2
3	H T H	1 0 1	2
4	T T H	0 0 1	1

# Definition of a random variable

## Definition 6.1:

Let a probability space  $(\Omega; \mathcal{S}; P)$  be given. A function

$$X : \Omega \rightarrow \mathbb{R}; \quad \omega \mapsto X(\omega)$$

is called *random variable*, if the set

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega | X(\omega) \leq x\}$$

belongs to the sigma algebra  $\mathcal{S}$  over  $\Omega$  for all  $x \in \mathbb{R}$ .

## Example 6.2: see Ex. 6.1 (1)

$$\begin{aligned} P(X = 1) &= P(\{(H, T, T), (T, H, T), (T, T, H)\}) \\ &= \frac{3}{8} \end{aligned}$$

---

```
load("Example6-1.RData")
sum(Example6_1$number_of_hats == 1) /
length(Example6_1$number_of_hats)
```

---

[1] 0.375

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= \frac{1}{8} + \frac{3}{8} = \frac{1}{2} \end{aligned}$$

---

```
sum(Example6_1$number_of_hats==0 |
Example6_1$number_of_hats==1) /
length(Example6_1$number_of_hats)
```

---

[1] 0.5

## Example 6.2: see Ex. 6.1 (2)

$$P(X > 1) = 1 - P(X \leq 1) = \frac{1}{2}$$

---

```
sum(Example6_1$number_of_hats > 1) /  
length(Example6_1$number_of_hats)
```

---

[1] 0.5

$$P(0 < X \leq 2) = P(X = 1) + P(X = 2) = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}$$

---

```
sum(Example6_1$number_of_hats>0 &  
Example6_1$number_of_hats<=2) /  
length(Example6_1$number_of_hats)
```

---

[1] 0.75

# Distribution function

## Definition 6.2:

The function  $F(x) := P(\{X \leq x\})$ , which assigns to each  $x \in \mathbb{R}$  the probability that the random variable  $X$  is less than or equal to  $x$ , is called *distribution function* of  $X$ .

We use the short hand  $P(X \leq x)$ .

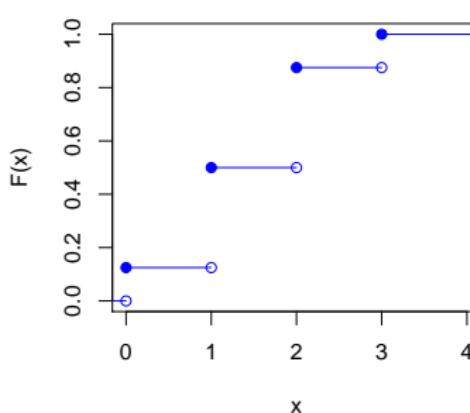
## Example 6.3: see Ex. 6.2 (1)

For the random variable  $X$  we have:

1. For  $x < 0$ :  $P(X \leq x) = 0$
2. For  $0 \leq x < 1$ :  $P(X \leq x) = P(X = 0) = 1/8$
3. For  $1 \leq x < 2$ :  $P(X \leq x) = P(X = 0) + P(X = 1) = 1/2$
4. For  $2 \leq x < 3$ :  $P(X \leq x) = 1 - P(X = 3) = 7/8$
5. For  $x \geq 3$ :  $P(X \leq x) = 1$

Therefore, we have:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1/8 & \text{for } 0 \leq x < 1 \\ 1/2 & \text{for } 1 \leq x < 2 \\ 7/8 & \text{for } 2 \leq x < 3 \\ 1 & \text{for } x \geq 3 \end{cases}$$



## Example 6.3: see Ex. 6.2 (2)

Calculation of  $F(x)$  and construction of the graphic in R:

```
load("Example6-1.RData")  
  
x6_3 <- Example6_1$number_of_hats  
F_x6_3 <- cumsum(prop.table(x = table(x = x6_3)))  
  
plot(ecdf(x6_3), col = "blue", xlab = "x", ylab = "F(x)",  
     main = "", xlim = c(0, 4))  
points(0:3, c(0, F_x6_3[-4]), col = "blue")  
  
F_x6_3
```

---

0	1	2	3
0.125	0.500	0.875	1.000

# Properties of distribution functions

We have:

1.  $0 \leq F(x) \leq 1$  for all  $x \in \mathbb{R}$
2.  $\lim_{x \rightarrow \infty} F(x) = 1$
3.  $\lim_{x \rightarrow -\infty} F(x) = 0$
4.  $F$  is monotonously increasing.
5.  $F$  has no more than a countable number of jump discontinuities.
6.  $F$  is right-continuous.

## Discrete random variables

### Definition 6.3:

A random variable  $X$  is called *discrete* if it cannot take more than a countable number of values (realisations) with a positive probability.

If  $x_1, \dots, x_i$  are the realisations of  $X$ , then the probabilities

$P(X = x_1), \dots, P(X = x_i)$  contain the complete information about this random variable.

### Definition 6.4:

The function  $f(x)$ , which is defined for all real  $x$  and given by

$$f(x) = \begin{cases} P(X = x) & \text{for all possible realisations of } X \\ 0 & \text{else} \end{cases},$$

is called *probability function* of the (discrete) random variable  $X$ .

## Example 6.4: Red and blue balls (1)

An urn contains 3 red and 7 black balls. 3 balls are drawn with replacement. Determine the probability table for the number of red balls drawn.

Furthermore, make suitable plots for the respective probability function and distribution function.

## Example 6.4: Red and blue balls (2)

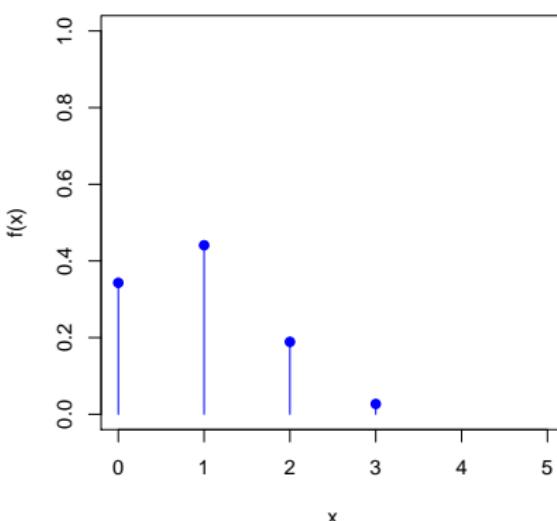
$x_i$	0	1	2	3
$f(x_i)$	0.343	0.441	0.189	0.027

```
x6_4 <- 0:3
```

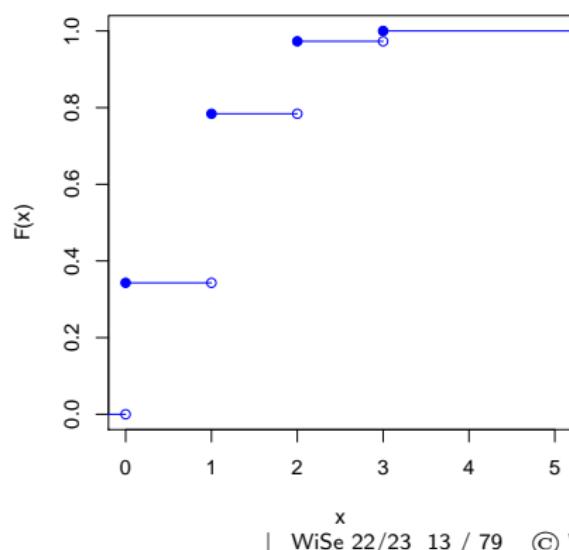
```
f_x6_4 <- c(0.343, 0.441, 0.189, 0.027)
```

```
F_x6_4 <- cumsum(f_x6_4)
```

Probability function



Distribution function



## Example 6.4: Red and blue balls (3)

Construction of the graphics in R:

```
plot(x = x6_4, y = f_x6_4, type = "h", lwd = 3,
      xlim = c(0,5), ylim = c(0,0.6), col = "blue", xlab="x",
      ylab = "f(x)", main = "Probability function")
points(x = x6_4, y = f_x6_4, col = "blue", pch = 19)

x_axis <- c(-1,sort(x6_4),4)
y_axis <- c(0,F_x6_4,1)

plot(x = c(0,4), y = c(0,1), main = "Distribution function",
      type = "n", col = "blue", xlab = "x", ylab = "F(x)")

lines(x = x_axis[1:2],y=rep(y_axis[1],2),col="blue",lwd=2)
lines(x = x_axis[2:3],y=rep(y_axis[2],2),col="blue",lwd=2)
lines(x = x_axis[3:4],y=rep(y_axis[3],2),col="blue",lwd=2)
lines(x = x_axis[4:5],y=rep(y_axis[4],2),col="blue",lwd=2)
lines(x = x_axis[5:6],y=rep(y_axis[5],2),col="blue",lwd=2)
points(x = x_axis[1:5],y=y_axis[1:5], col="blue", pch=19)
points(x = x_axis[2:5],y=y_axis[1:4], col="blue", pch=1)
```

## Example 6.5: Another random variable (1)

Let a discrete random variable have the following probability and distribution function, respectively:

$$f(x) = \begin{cases} 0.2 \cdot 0.8^x & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{else} \end{cases}$$

$$F(x) = \begin{cases} 1 - 0.8^{x+1} & \text{for } x \geq 0 \\ 0 & \text{else} \end{cases}$$

Definition of  $f(x)$  and  $F(x)$  in R:

---

```
x6_5 <- 0:10
```

```
# ATTENTION: here functions
f_x <- function(x) {0.2 * 0.8^x}

F_x <- function(x) {1 - 0.8^(x+1)}
```

---

## Example 6.5: Another random variable (2)

We get the following tabulated results:

$x_i$	0	1	2	3	4	5	6	7	8	9	10
$f(x_i)$	0.200	0.160	0.128	0.102	0.082	0.066	0.052	0.042	0.034	0.027	0.021
$F(x_i)$	0.200	0.360	0.488	0.590	0.672	0.738	0.790	0.832	0.866	0.893	0.914

```
round(f_x(x6_5), digits = 3)
```

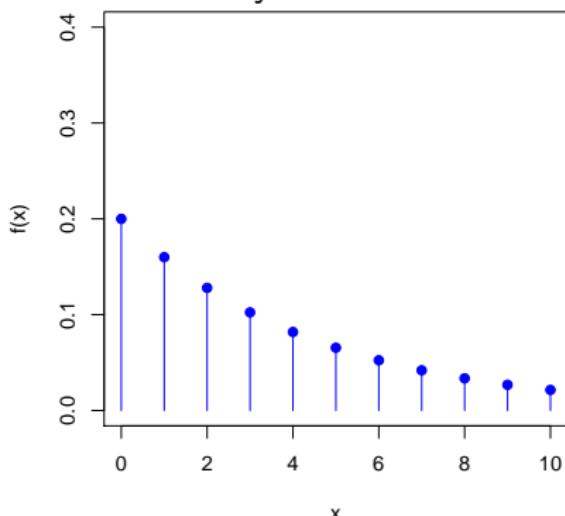
```
[1] 0.200 0.160 0.128 0.102 0.082 0.066 0.052 0.042 0.034 0.027 0.021
```

```
round(F_x(x6_5), digits = 3)
```

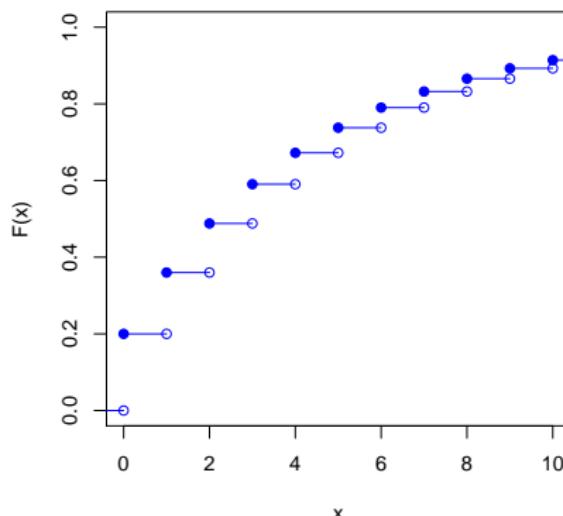
```
[1] 0.200 0.360 0.488 0.590 0.672 0.738 0.790 0.832 0.866 0.893 0.914
```

## Example 6.5: Another random variable (3)

Probability function



Distribution function



The random variable has a countably infinite number of realisations.  
Here, we are dealing with a geometric distribution with parameter  $p = 0.2$ .

# Continuous random variables

## Definition 6.5:

If there is a non-negative function  $f(x)$  for a random variable  $X$ , in such a way that the distribution function for all  $x$  can be described by

$$F(x) = \int_{-\infty}^x f(y) dy,$$

we call  $X$  a *continuous random variable*.

## Definition 6.6:

The function  $f(x)$  of Definition 7.5 is called the *density function* of the continuous random variable  $X$ .

# Properties of continuous random variables

1. The area between the density curve and the abscissa has to sum up to 1.
  1. Notice the analogy to the empirical relative frequency distribution (histogram).
2. The probability  $F(x)$  that  $X$  takes on a value which is less than or equal to  $x$  is expressed in terms of the measure of the area between the density curve and the abscissa on the interval  $(-\infty, x]$ . Notice the analogy to the empirical distribution function.
3.  $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

## Properties of continuous random variables (ctd.)

4.  $P(X = x) = 0$

Density values cannot be interpreted as probabilities!

5.  $X$  continuous  $\Rightarrow$

$$P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) =$$

$$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$$

6. Interpretation of densities:  $P(x_1 < X \leq x_2) \approx f(x) \cdot \underbrace{(x_2 - x_1)}_{\text{small}}$

7.  $f(x) > 1$  is possible!

8.  $F'(x) = f(x)$  for all  $x$ , for which  $F$  is differentiable.

## Example 6.6: A continuous random variable

Let the continuous random variable  $X$  have the following density function:

$$f(x) = \begin{cases} 0.5 & \text{for } 1 \leq x \leq 3 \\ 0 & \text{else} \end{cases} .$$

Then, we get the distribution function

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ 0.5x - 0.5 & \text{for } 1 \leq x \leq 3 \\ 1 & \text{for } x > 3 \end{cases} .$$

Application of  $f(x)$  and  $F(x)$  in R:

```
# Distinction from the latest functions
f_x6_6 <- function(x) {0.5}

F_x6_6 <- function(x) {0.5 * x - 0.5}
```

## Example 6.7: An exponentially distributed RV

Let the continuous random variable  $X$  have the following distribution function:

$$F(x) = \begin{cases} 1 - e^{-\frac{1}{2}x} & \text{for } x \geq 0 \\ 0 & \text{else} \end{cases}$$

(exponential distribution with parameter  $\lambda = \frac{1}{2}$ ). Then, differentiation yields the density function

$$f(x) = \begin{cases} \frac{1}{2}e^{-\frac{1}{2}x} & \text{for } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

Application of  $F(x)$  and  $f(x)$  in R:

```
F_x6_7 <- function(x) {1 - exp(-1/2 * x)}
```

```
f_x6_7 <- function(x) {1/2 * exp(-1/2 * x)}
```

## Definition 6.7:

Let the random variable  $X$  have the following probability or density function  $f(x)$ , respectively. If

$$\sum_i |f(x_i) \cdot x_i| < \infty \quad \text{or} \quad \int_{-\infty}^{\infty} |f(x) \cdot x| dx < \infty$$

holds, then

$$E(X) := \sum_i f(x_i) \cdot x_i \text{ or } E(X) := \int_{-\infty}^{\infty} f(x) \cdot x dx$$

is called the *expected value* of the discrete or continuous random variable  $X$ , respectively.

## Definition 6.8:

Let the random variable  $X$  have the following probability or density function  $f(x)$ , respectively. If

$$\sum_i |f(x_i) \cdot x_i^2| < \infty \quad \text{or} \quad \int_{-\infty}^{\infty} |f(x) \cdot x^2| dx < \infty$$

holds, then

$$\text{Var}(X) := \sum_i (x_i - \mathbb{E} X)^2 \cdot f(x_i) \text{ or}$$

$$\text{Var}(X) := \int_{-\infty}^{\infty} (x - \mathbb{E} X)^2 \cdot f(x) dx$$

is called the *variance* of the discrete or continuous random variable  $X$ , respectively.

## Example 6.8: see Ex. 6.4

$$\begin{aligned} E(X) &= \sum_{x=0}^3 x \cdot f(x) \\ &= 0 \cdot 0.343 + 1 \cdot 0.441 + 2 \cdot 0.189 + 3 \cdot 0.027 \\ &= 0.9 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \sum_{x=0}^3 x^2 \cdot f(x) - E(X)^2 \\ &= 0^2 \cdot 0.343 + 1^2 \cdot 0.441 + 2^2 \cdot 0.189 + 3^2 \cdot 0.027 - 0.9^2 \\ &= 1.44 - 0.9^2 = 0.63 \end{aligned}$$

Calculation of  $E(X)$  and  $\text{Var}(X)$  in R:

---

```
Mean_X6_8 <- weighted.mean(x = x6_4, w = f_x6_4)
```

---

```
Var_X6_8 <- sum(f_x6_4*(x6_4 - Mean_X6_8)^2)
```

---

```
Mean_X6_8
```

[1] 0.9

---

```
Var_X6_8
```

[1] 0.63

## Example 6.9: see Ex. 6.6 (1)

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_1^3 x \cdot 0.5 dx \\ &= \left[ \frac{1}{2} \cdot x^2 \cdot 0.5 \right]_1^3 = \left[ \frac{1}{4} \cdot x^2 \right]_1^3 \\ &= \frac{9}{4} - \frac{1}{4} = 2 \end{aligned}$$

Calculation of  $E(X)$  in R:

---

```
Mean_X6_9 <- integrate(f = function(x){0.5*x}, lower = 1,  
                           upper = 3)$value
```

```
Mean_X6_9
```

---

[1] 2

## Example 6.9: see Ex. 6.6 (2)

$$\begin{aligned}\text{Var}(X) &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - E(X)^2 = \int_1^3 x^2 \cdot 0.5 dx - 2^2 \\ &= \left[ \frac{1}{6}x^3 \right]_1^3 - 4 \\ &= \frac{27}{6} - \frac{1}{6} - 4 = \frac{1}{3}\end{aligned}$$

---

Calculation of  $\text{Var}(X)$  in R:

```
Var_X6_9 <- integrate(f = function(x){0.5*x^2}, lower = 1,  
                      upper = 3  
                     )$value - Mean_X6_9^2  
Var_X6_9
```

---

[1] 0.3333333

## Linear transformation of a random variable

Let  $Y = a + b \cdot X$ , then

$$\mathbb{E}(Y) = a + b \cdot \mathbb{E}(X)$$

$$\text{Var}(Y) = b^2 \cdot \text{Var}(X)$$

### Example 6.10:

X: Filling weight of a package of detergent in kg

Y: Deviation from targeted weight of 5 kg in g

Then  $Y = (X - 5) \cdot 1000 = -5000 + 1000 \cdot X$  and therefore  $a = -5000$  and  $b = 1000$ .

### Example 6.11: (see Example 6.4)

Now we are interested in the share of black balls

(earlier: number of red balls):

$$Y = \frac{n - X}{n} = 1 - \frac{1}{n} \cdot X \text{ with } n = 3$$

## Standard transformation

The special linear transformation

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = -\frac{E(X)}{\sqrt{\text{Var}(X)}} + \frac{1}{\sqrt{\text{Var}(X)}} \cdot X$$

is called standard transformation of random variable  $X$  (see Chapter 4).

We have  $E(Z) = 0$  and  $\text{Var}(Z) = 1$ .

## Example 6.12: see Ex. 6.4 and 6.8

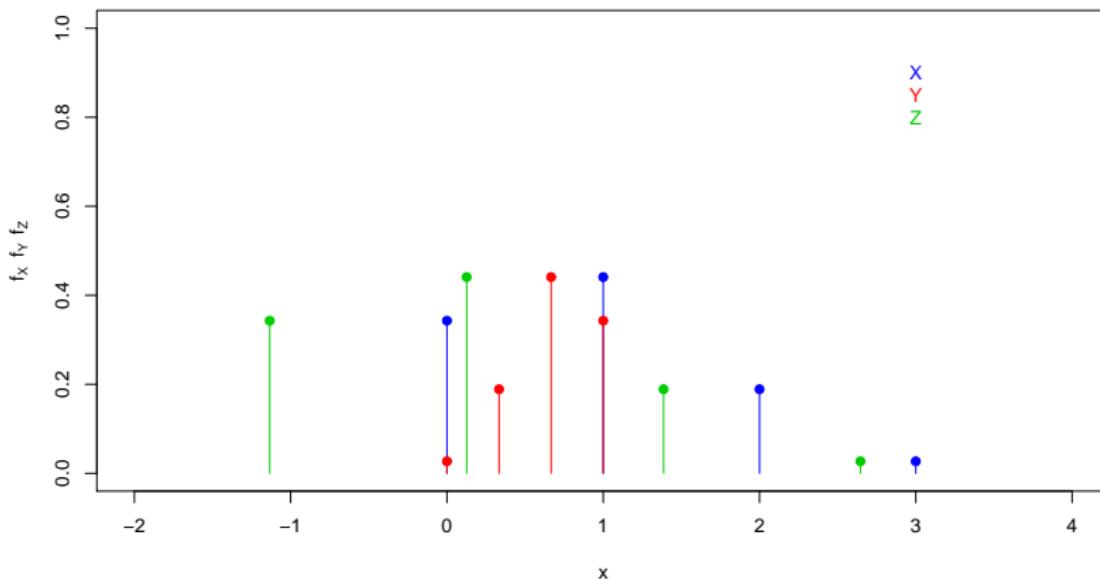
$X$	0	1	2	3
$Z$	$-0.9/\sqrt{0.63}$	$0.1/\sqrt{0.63}$	$1.1/\sqrt{0.63}$	$2.1/\sqrt{0.63}$
$f(x)$	0.343	0.441	0.189	0.027

Calculation of  $Z$  in R:

```
Z6_12 <- (x6_4 - Mean_X6_8) / sqrt(Var_X6_8)  
round(Z6_12, digits = 3)
```

```
[1] -1.134 0.126 1.386 2.646
```

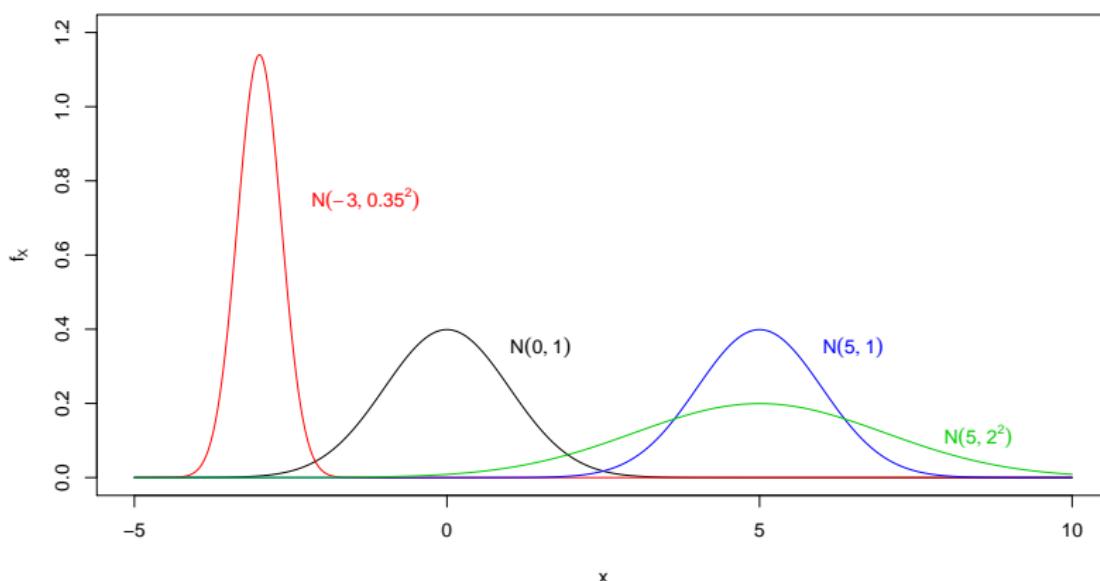
# Visualisation for Examples 6.11 and 6.12



$$Y = \frac{3 - X}{3}$$

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

# Different normal distributions $N(\mu; \sigma^2)$



$$f(x) = \varphi(x | \mu; \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

# Quantiles of distributions

Definition of a quantile (see Schaich and Münnich, 2001):

For a random variable  $X$ , a value  $x$ , which satisfies the inequalities

$$P(X \leq x) \geq p \quad \text{and} \quad P(X \geq x) \leq 1 - p$$

for  $0 < p < 1$ , is called its *quantile of order p* ( $p$ -quantile).

The median  $x_{0.5}$  (also called the 0.5-quantile), as well as the first and third quartile ( $p = 0.25$  and  $p = 0.75$ , respectively) are particularly interesting.

For continuous random variables (with a strictly monotonous distribution function) the  $p$ -quantile equals  $x_p = F^{-1}(p)$ .

Schaich, E. and Münnich, R. (2001): Mathematische Statistik für Ökonomen: Lehrbuch.  
Vahlen.

## Example 6.13: Quantiles of exp. distr. (1)

Let the random variable  $X$  follow an exponential distribution with parameter  $\lambda = \frac{1}{2}$ . The distribution function is:

$$F(x) = \begin{cases} 1 - e^{-\frac{1}{2} \cdot x} & \text{for } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

The  $p$ -quantile is derived as follows:

$$\begin{array}{l|l} p = 1 - e^{-\frac{1}{2} \cdot x} & | -p \\ e^{-\frac{1}{2} \cdot x} = 1 - p & | +e^{-\frac{1}{2} \cdot x} \\ -0.5 \cdot x = \ln(1 - p) & | \ln \\ x = -2 \ln(1 - p) & | : (-0.5) \end{array}$$

## Example 6.13: Quantiles of exp. distr. (2)

Therefore, the median is

$$\begin{aligned}x_{0.5} &= -2 \ln(1 - 0.5) = -2 \ln\left(\frac{1}{2}\right) = -2(\ln 1 - \ln 2) \\&= -2 \ln 1 + 2 \ln 2 = 2 \ln 2 \approx 1.3863\end{aligned}$$

and the first quantile is

$$x_{0.25} = -2 \ln \frac{3}{4} \approx 0.5754.$$

Calculation of  $x_{0.5}$  and  $x_{0.25}$  in R:

```
q_050 <- -2 * log(1 - 0.5)
q_025 <- -2 * log(1 - 0.25)
```

```
round(q_050, digits = 4)
```

```
[1] 1.3863
```

```
round(q_025, digits = 4)
```

```
[1] 0.5754
```

# Markov's and Tchebysheff's inequality

**Theorem 6.1 (Markov's inequality):** If a random variable  $X$  only takes on non-negative values and the expected value  $E(X)$  exists, the following approximation holds for every  $x^* > 0$ :

$$P(X \geq x^*) \leq \frac{E(X)}{x^*}.$$

**Theorem 6.2 (Tchebysheff's inequality):** If the variance  $\text{Var}(X)$  of a random variable  $X$  exists, the following holds for  $\varepsilon > 0$ :

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

Notice the special case where  $\varepsilon = k \cdot \sqrt{\text{Var}(X)}$ .

## Example 6.14: An inequality

Let a non-negative random variable  $X$  have the expected value  $E(X) = 10$  (applies for discrete as well as continuous variables).

We can approximate:

$$P(X \geq 25) \leq \frac{10}{25} = 0.4$$

$$P(X \geq 40) \leq \frac{10}{40} = 0.25$$

$$P(X \geq 5) \leq \frac{10}{5} = 2.$$

The third row is a *trivial* approximation as probabilities are bounded by 0 and 1.

## Example 6.15: Another inequality

Let the expected value  $E(X) = 2$  and the variance  $\text{Var}(X) = 36$  of a random variable  $X$  be known.

Then we have:

$$P(-8 < X < 12) = P(|X - 2| < 10) \geq 1 - \frac{36}{100} = 0.64$$

$$P(|X - 2| \geq 10) \leq \frac{36}{100} = 0.36$$

$$P(X \leq -3 \vee X \geq 7) = P(|X - 2| \geq 5) \leq \frac{36}{25} = 1.44 .$$

## Example 6.16: More dimensions (1)

We are looking at two- and multi-dimensional random variables.

a)

Random questioning of a person with replacement (income; age):  
The resulting observation is (1815; 25).

b)

Two rolls of a dice:

The resulting pair of number of pips is (4; 6).

We could as well be interested in the overall sum of pips or the product of the number of pips (10; 24).

## Example 6.16: More dimensions (2)

c) An urn contains  $N = 100$  balls, of which 30 are red (r), 20 are white (w) and 50 are black (s). How does the sample space of this experiment look like, if we draw 3 balls?  $X = (\text{number r, number w})$

$\omega$	$X(\omega)$	$\omega$	$X(\omega)$	$\omega$	$X(\omega)$
(s,s,s)	(0,0)	(r,s,s)	(1,0)	(w,s,s)	(0,1)
(s,s,r)	(1,0)	(r,s,r)	(2,0)	(w,s,r)	(1,1)
(s,s,w)	(0,1)	(r,s,w)	(1,1)	(w,s,w)	(0,2)
(s,r,s)	(1,0)	(r,r,s)	(2,0)	(w,r,s)	(1,1)
(s,r,r)	(2,0)	(r,r,r)	(3,0)	(w,r,r)	(2,1)
(s,r,w)	(1,1)	(r,r,w)	(2,1)	(w,r,w)	(1,2)
(s,w,s)	(0,1)	(r,w,s)	(1,1)	(w,w,s)	(0,2)
(s,w,r)	(1,1)	(r,w,r)	(2,1)	(w,w,r)	(1,2)
(s,w,w)	(0,2)	(r,w,w)	(1,2)	(w,w,w)	(0,3)

## Example 6.16: More dimensions (3)

Construction of the table in R:

```
omega <- expand.grid(lapply(X = 1:3,
                             FUN = function(x) c("s", "r", "w")))
Number_of_r <- rowSums(omega == "r")
Number_of_w <- rowSums(omega == "w")
X6_16 <- cbind(Number_of_r, Number_of_w)

Example6_16 <- data.frame(omega, X6_16)
names(Example6_16) <- c("Omega", "", "", 
                         "Number_of_r", "Number_of_w")

head(Example6_16)
```

	Omega	Number_of_r	Number_of_w
1	s s s	0	0
2	r s s	1	0
3	w s s	0	1
4	s r s	1	0
5	r r s	2	0
6	w r s	1	1

## Example 6.16: More dimensions (4)

The set of all possible realisations is

$$\{(x_1, x_2) \mid x_1, x_2 \in \mathbb{N}_0, 0 \leq x_1 + x_2 \leq 3\}.$$

For instance, we could determine  $P(X_1 = x_1; X_2 = x_2)$ ,  
 $P(X_1 \leq x_1; X_2 \leq x_2)$ ,  $P(X_1 \leq x_1 \vee X_2 \leq x_2)$  or  $P(X_1 \leq x_1 | X_2 = x_2)$ .

# Multi-dimensional random variables

Multi-dimensional random variables can be considered as a generalisation of one-dimensional random variables.

The inverse images of half-open  $n$ -intervals must again be part of the sigma algebra over  $\Omega$ .

# Distribution function of multi-dimensional random variables

The function

$$F_X(x_1, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \quad ,$$

which gives the probability that  $X_1$  is at most  $x_1$  and  $X_n$  is at most  $x_n$  for all real  $n$ -tuple is called distribution function of the random vector  $X_1, X_2, \dots, X_n$ .

Interval probabilities in the two-dimensional case:

$$\begin{aligned} P(x'_1 < X_1 \leq x''_1, x'_2 < X_2 \leq x''_2) &= \\ F(x''_1, x''_2) - F(x'_1, x''_2) - F(x''_1, x'_2) + F(x'_1, x'_2) &\quad . \end{aligned}$$

# Discrete random vectors

## Discrete random vector

A random vector  $\mathbf{X}$  is called a multi-dimensional discrete random variable if each of its components can take on at most a countable number of values.

## Probability function of a discrete random vector

The function  $f(x_1, x_2)$ , which is defined for all real pairs of numbers  $(x_1, x_2)$  and which is characterised by

$$f(x_1, x_2) = \begin{cases} P(X_1 = x_{1j}, X_2 = x_{2k}) & \text{for all } j, k \\ 0 & \text{else} \end{cases}$$

is called the probability function of the discrete random vector  $\mathbf{X}$ .

# Two-dimensional discrete random variables

$X_1 \backslash X_2$	$x_{21}$	$\dots$	$x_{2k}$	$\dots$	$x_{2r}$	$\sum$
$x_{11}$	$f(x_{11}, x_{21})$	$\dots$	$f(x_{11}, x_{2k})$	$\dots$	$f(x_{11}, x_{2r})$	$f_{X_1}(x_{11})$
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
$x_{1j}$	$f(x_{1j}, x_{21})$	$\dots$	$f(x_{1j}, x_{2k})$	$\dots$	$f(x_{1j}, x_{2r})$	$f_{X_1}(x_{1j})$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_{1m}$	$f(x_{1m}, x_{21})$	$\dots$	$f(x_{1m}, x_{2k})$	$\dots$	$f(x_{1m}, x_{2r})$	$f_{X_1}(x_{1m})$
$\sum$	$f_{X_2}(x_{21})$	$\dots$	$f_{X_2}(x_{2k})$	$\dots$	$f_{X_2}(x_{2r})$	1

# Properties of discrete random vectors

1. We have:

$$\sum_j \sum_k f(x_{1j}, x_{2k}) = 1.$$

2. Distribution function:

$$F(x_1, x_2) = \sum_{x_{1j} \leq x_1} \sum_{x_{2k} \leq x_2} f(x_{1j}, x_{2k})$$

3. Interval probabilities:

$$P(x'_1 < X_1 \leq x''_1, x'_2 < X_2 \leq x''_2) = \sum_{x'_1 < x_{1j} \leq x''_1} \sum_{x'_2 < x_{2k} \leq x''_2} f(x_{1j}, x_{2k})$$

## Marginal distributions of bivariate distributions

In addition to the joint distribution of the random vector  $(X_1, X_2)$  with the distribution function  $F(x_1, x_2)$ , the *marginal distributions*, ergo the univariate distributions of the random variables involved in the distribution functions  $F_{X_1}(x_1)$  and  $F_{X_2}(x_2)$ , may be considered as well. We obtain those by

$$F_{X_1}(x_1) = \sum_{x_{1j} \leq x_1} \sum_k f(x_{1j}, x_{2k}) \quad \text{or} \quad F_{X_2}(x_2) = \sum_j \sum_{x_{2k} \leq x_2} f(x_{1j}, x_{2k})$$

and thus by adding up all probabilities of the variable which is not of interest.

The indexation of the marginal distribution functions is used for unique identification.

## Example 6.17: see Ex. 6.16 c) (1)

We are interested in the random vector (number of red balls, number of white balls). For example, we have:

$$f(1, 0) = 3 \cdot 0.3^1 \cdot 0.2^0 \cdot 0.5^2 = 0.225 .$$

Calculation of  $f(1, 0)$  in R:

```
Number_of_s <- rowSums(omega == "s")
Example6_16 <- cbind(Example6_16, Number_of_s)
Probs <- 0.3^Example6_16$Number_of_r *
          0.2^Example6_16$Number_of_w *
          0.5^Example6_16$Number_of_s

Example6_16 <- cbind(Example6_16, Probs)

pos <- which(Example6_16$Number_of_r == 1 &
              Example6_16$Number_of_w == 0)
f_1_0 <- sum(Example6_16[pos, 7])
f_1_0
[1] 0.225
```

## Example 6.17: see Ex. 6.16 c) (2)

Finally, we get the following probability table:

$X_1 \mid X_2$	$x_{21} = 0$	$x_{22} = 1$	$x_{23} = 2$	$x_{24} = 3$	$\sum$
$x_{11} = 0$	0.125	0.150	0.060	0.008	0.343
$x_{12} = 1$	0.225	0.180	0.036	0.000	0.441
$x_{13} = 2$	0.135	0.054	0.000	0.000	0.189
$x_{14} = 3$	0.027	0.000	0.000	0.000	0.027
$\sum$	0.512	0.384	0.096	0.008	1.000

Probability table in R:

```
X1_6_17 <- 0:3 ; X2_6_17 <- 0:3
ProbTable6_17 <- matrix(c(0.125, 0.150, 0.060, 0.008,
                           0.225, 0.180, 0.036, 0.000,
                           0.135, 0.054, 0.000, 0.000,
                           0.027, 0.000, 0.000, 0.000),
                           ncol = length(X2_6_17),
                           byrow = TRUE)
dimnames(ProbTable6_17) <- list(X1_6_17, X2_6_17)
ProbTable_new6_17 <- addmargins(ProbTable6_17)
```

## Example 6.17: see Ex. 6.16 c) (3)

ProbTable\_new6\_17

	0	1	2	3	Sum
0	0.125	0.150	0.060	0.008	0.343
1	0.225	0.180	0.036	0.000	0.441
2	0.135	0.054	0.000	0.000	0.189
3	0.027	0.000	0.000	0.000	0.027
Sum	0.512	0.384	0.096	0.008	1.000

For  $F(1, 2)$  we have:

$X_1$	$X_2$	$x_{21} = 0$	$x_{22} = 1$	$x_{23} = 2$	$x_{24} = 3$	$\sum$
$x_{11} = 0$		0.125	0.150	0.060	0.008	0.343
$x_{12} = 1$		0.225	0.180	0.036	0.000	0.441
$x_{13} = 2$		0.135	0.054	0.000	0.000	0.189
$x_{14} = 3$		0.027	0.000	0.000	0.000	0.027
$\sum$		0.512	0.384	0.096	0.008	1.000

$$F(1, 2) = 0.125 + 0.150 + 0.060 + 0.225 + 0.180 + 0.036 = 0.776$$

## Example 6.17: see Ex. 6.16 c) (4)

Determination of the joint distribution function in R:

```
F_x1_x2 <- t(apply(X = apply(X = ProbTable6_17,
                           MARGIN = 2, FUN = cumsum),
                           MARGIN = 1, FUN = cumsum))
F_x1_x2
```

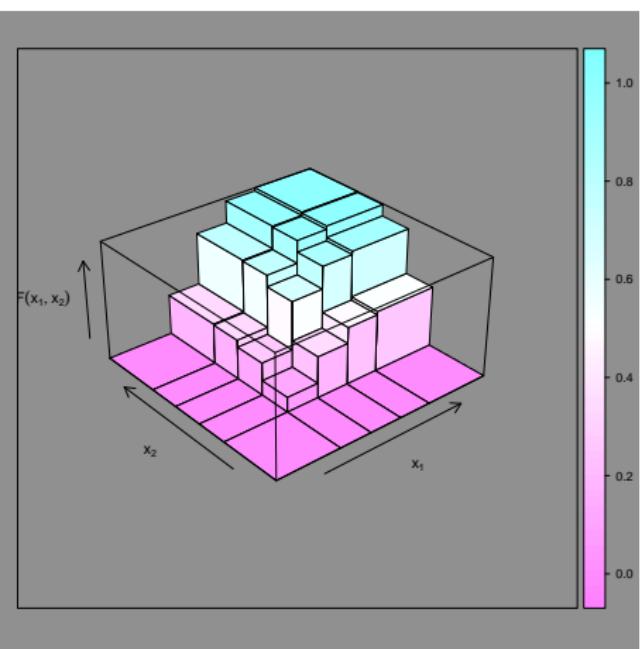
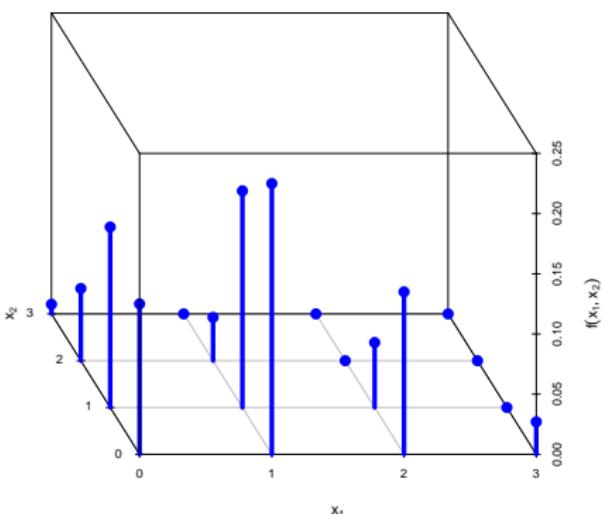
	0	1	2	3
0	0.125	0.275	0.335	0.343
1	0.350	0.680	0.776	0.784
2	0.485	0.869	0.965	0.973
3	0.512	0.896	0.992	1.000

---

```
F_x1_x2[rownames(F_x1_x2) == 1, colnames(F_x1_x2) == 2]
```

```
[1] 0.776
```

## Example 6.17: see Ex. 6.16 c) (5)



# Continuous random vectors

Continuous random vectors are defined analogously to continuous random variables. We have:

$$1. \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 dx_1 = 1$$

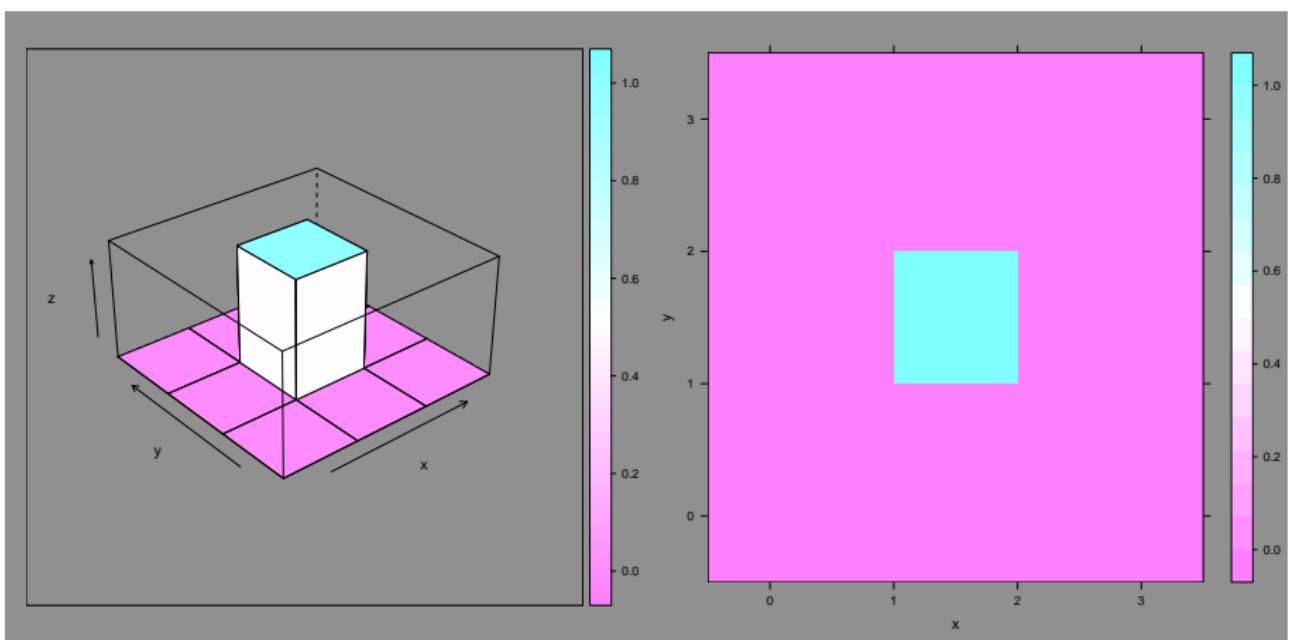
$$2. F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(y_1, y_2) dy_2 dy_1$$

$$3. P(x'_1 < X_1 \leq x''_1, x'_2 < X_2 \leq x''_2) = \int_{x'_1}^{x''_1} \int_{x'_2}^{x''_2} f(x_1, x_2) dx_2 dx_1$$

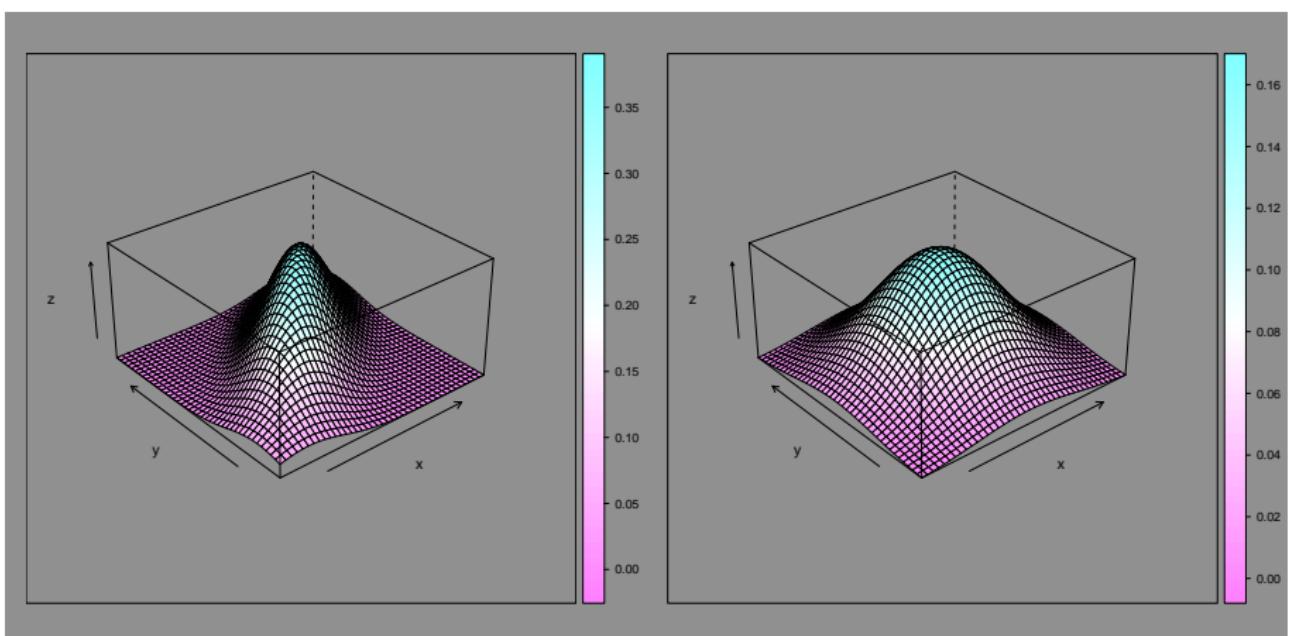
$$4. f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} \text{ (assuming differentiability)}$$

$$5. \text{Marg. distr.: } f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \text{ and } f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

# Rectangular distribution



# Bivariate normal distribution



## Example 6.18: A continuous random vector (1)

Let the density function of a continuous random vector be

$$f(x_1, x_2) = \begin{cases} 6 \cdot \exp(-2x_1) \cdot \exp(-3x_2) & \text{for } x_1, x_2 > 0 \\ 0 & \text{else} \end{cases} .$$

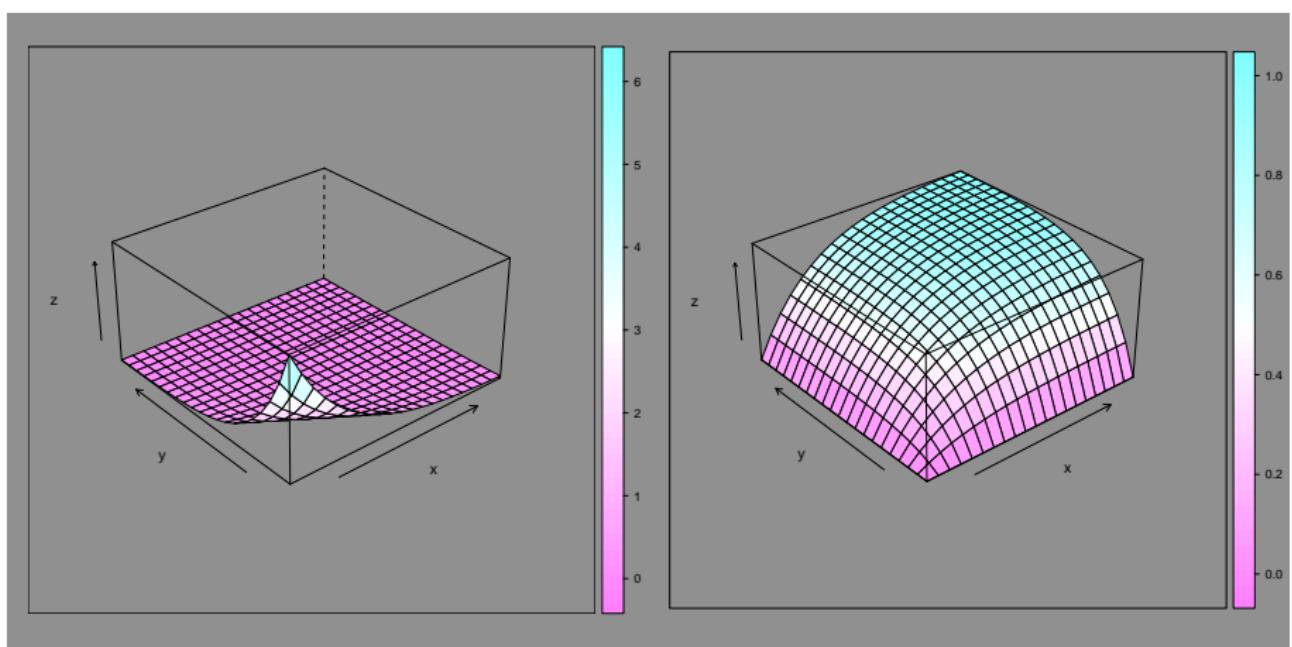
Using integration we get the following distribution function:

$$F(x_1, x_2) = \begin{cases} (1 - \exp(-2x_1)) \cdot (1 - \exp(-3x_2)) & \text{for } x_1, x_2 > 0 \\ 0 & \text{else} \end{cases} .$$

For  $x_2 > 0$  we get the following marginal density function for random variable  $X_2$ :

$$f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = \int_0^{\infty} 6 \cdot \exp(-2x_1 - 3x_2) dx_1 = 3 \cdot \exp(-3x_2) .$$

## Example 6.18: A continuous random vector (2)



# Stochastical independence

Let  $F(x_1, x_2)$  be the joint distribution function of the random vector  $\mathbf{X}$  and let  $F(x_1)$  and  $F(x_2)$  be the marginal distribution functions. Two random variables  $X_1$  and  $X_2$  are called stochastically independent if and only if we have

$$F(x_1, x_2) = F(x_1) \cdot F(x_2)$$

for all  $(x_1, x_2) \in \mathbb{R}^2$ . Otherwise, they are called stochastically dependent. Stochastical independence may be proven using probabilities or probability functions and density functions as well (see Schaich and Münnich, 2001).

### Example 6.19: see Ex. 6.17

$$f_{X_1}(0) \cdot f_{X_2}(0) = 0.343 \cdot 0.512 = 0.175616 \neq 0.125 = f(0, 0)$$

$X_1$  and  $X_2$  are stochastically dependent.

Checking in R:

```
(ProbTable_new6_17[1,5] * ProbTable_new6_17[5,1]) ==  
ProbTable_new6_17[1,1]
```

```
[1] FALSE
```

### Example 6.20 (1)

Let the random vector  $\mathbf{X}$  have the following probability table:

$X_1$	$X_2$	2	4	6	$\sum$
1		0.05	0.14	0.01	0.20
5		0.20	0.56	0.04	0.80
	$\sum$	0.25	0.70	0.05	1.00

## Example 6.20 (2)

Probability table in R:

```
ProbTable6_20 <- matrix(c(0.05,0.14,0.01,  
                           0.20,0.56,0.04),  
                           ncol = 3, byrow = TRUE)
```

```
X1_6_20 <- c(1, 5)  
X2_6_20 <- seq(2, 6, 2)
```

```
rownames(ProbTable6_20) <- X1_6_20  
colnames(ProbTable6_20) <- X2_6_20
```

```
ProbTable6_20 <- addmargins(ProbTable6_20)
```

```
ProbTable6_20
```

---

	2	4	6	Sum
1	0.05	0.14	0.01	0.2
5	0.20	0.56	0.04	0.8
Sum	0.25	0.70	0.05	1.0

## Example 6.20 (3)

We have  $f_{X_1}(x_{1i}) \cdot f_{X_2}(x_{2j}) = f(x_{1i}, x_{2j})$  for all  $i, j$ . Therefore,  $X_1$  and  $X_2$  are stochastically independent.

Checking of stochastic independence in R:

---

```
round(ProbTable6_20[3,] * ProbTable6_20[1,4],4) ==
ProbTable6_20[1,]
```

---

2	4	6	Sum
TRUE	TRUE	TRUE	TRUE

---

```
round(ProbTable6_20[3,] * ProbTable6_20[2,4],4) ==
ProbTable6_20[2,]
```

---

2	4	6	Sum
TRUE	TRUE	TRUE	TRUE

## Covariance of two random variables

The covariance of two random variables  $X_1$  and  $X_2$  is defined as:

$$\text{Cov}(X_1, X_2) = E((x_1 - E X_1) \cdot (x_2 - E X_2)) .$$

For discrete random variables we have:

$$\text{Cov}(X_1, X_2) = \sum_i \sum_j (x_{1i} - E X_1) \cdot (x_{2j} - E X_2) \cdot f(x_{1i}, x_{2j}) .$$

Analogously, for continuous random variables we have:

$$\text{Cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - E X_1) \cdot (x_2 - E X_2) f(x_1, x_2) dx_1 dx_2 .$$

Furthermore, the displacement law holds:

$$\text{Cov}(X_1, X_2) = E(X_1 \cdot X_2) - E X_1 \cdot E X_2 .$$

## Example 6.21: see Ex. 6.17 (1)

$$E X_1 = 0 \cdot 0.343 + 1 \cdot 0.441 + 2 \cdot 0.189 + 3 \cdot 0.027 = 0.9$$

$$E X_2 = 0 \cdot 0.512 + 1 \cdot 0.384 + 2 \cdot 0.096 + 3 \cdot 0.008 = 0.6$$

Calculation of  $E X_1$  and  $E X_2$  in R:

```
f_X1_6_17 <- rowSums(ProbTable6_17)
f_X2_6_17 <- colSums(ProbTable6_17)
```

```
Mean_X1 <- sum(f_X1_6_17 * X1_6_17)
Mean_X2 <- sum(f_X2_6_17 * X2_6_17)
```

```
Mean_X1_old <- Mean_X1
Mean_X2_old <- Mean_X2
```

Mean\_X1

[1] 0.9

Mean\_X2

[1] 0.6

## Example 6.21: see Ex. 6.17 (2)

$$\begin{aligned}
 \text{Cov}(X_1, X_2) &= 0 \cdot 0 \cdot 0.125 + 0 \cdot 1 \cdot 0.150 + 0 \cdot 2 \cdot 0.060 + 0 \cdot 3 \cdot 0.008 \\
 &\quad + 1 \cdot 0 \cdot 0.225 + 1 \cdot 1 \cdot 0.180 + 1 \cdot 2 \cdot 0.036 + 1 \cdot 3 \cdot 0.000 \\
 &\quad + 2 \cdot 0 \cdot 0.135 + 2 \cdot 1 \cdot 0.054 + 2 \cdot 2 \cdot 0.000 + 2 \cdot 3 \cdot 0.000 \\
 &\quad + 3 \cdot 0 \cdot 0.027 + 3 \cdot 1 \cdot 0.000 + 3 \cdot 2 \cdot 0.000 + 3 \cdot 3 \cdot 0.000 \\
 &\quad - 0.9 \cdot 0.6 \\
 &= -0.18
 \end{aligned}$$

Calculation of  $\text{Cov}(X_1, X_2)$  in R:

```

Intermed_matrix6_21 <- matrix(
  rep(x = X2_6_17, times = length(X1_6_17)),
  ncol = length(X2_6_17), byrow = TRUE)

Cov_X1_X2 <- sum(Intermed_matrix6_21 * X2_6_17 *
  ProbTable6_17) - Mean_X1 * Mean_X2
Cov_X1_X2_old <- Cov_X1_X2
Cov_X1_X2

[1] -0.18
  
```

## Example 6.22: see Ex. 6.20

$$\begin{aligned}\text{Cov}(X_1, X_2) &= 1 \cdot 2 \cdot 0.05 + 1 \cdot 4 \cdot 0.14 + 1 \cdot 6 \cdot 0.01 + 5 \cdot 2 \cdot 0.2 + 5 \cdot 4 \cdot 0.56 \\ &\quad + 5 \cdot 6 \cdot 0.04 - (0.2 + 5 \cdot 0.8) \cdot (2 \cdot 0.25 + 4 \cdot 0.7 + 6 \cdot 0.05) \\ &= 15.12 - 15.12 = 0\end{aligned}$$

Notice that  $X_1$  and  $X_2$  are stochastically independent!

Calculation of  $\text{Cov}(X_1, X_2)$  in R:

---

```
i <- 1:3 ; ProbTable6_20 <- ProbTable6_20[-3,-4]
Mean_X1 <- weighted.mean(x = X1_6_20,
                           w = prop.table(ProbTable6_20[,2]))
Mean_X2 <- weighted.mean(x = X2_6_20,
                           w = prop.table(ProbTable6_20[2,]))

Cov_X1_X2 <- (sum(X1_6_20[1] * X2_6_20[i] *
                     ProbTable6_20[1,i]) +
                  sum(X1_6_20[2] * X2_6_20[i] *
                     ProbTable6_20[2,i])) - Mean_X1 * Mean_X2

Cov_X1_X2
[1] 0
```

---

# Independence and uncorrelatedness

If  $\text{Cov}(X_1, X_2) = 0$ , then the random variables  $X_1$  and  $X_2$  are called uncorrelated.

We have:

$$\begin{array}{ccc} \text{Independence} & \xrightarrow{\quad} & \text{Uncorrelatedness} \\ & \Leftarrow & \end{array}$$

Example 6.23:

$X_1$	$X_2$	-2	0	1	$\sum$
0		0.125	0.000	0.250	0.375
1		0.125	0.250	0.250	0.625
$\sum$		0.250	0.250	0.500	1.000

We have  $\text{Cov}(X_1, X_2) = 0$  but  $f(0, 0) \neq f_{X_1}(0) \cdot f_{X_2}(0)$  as well. Therefore,  $X_1$  and  $X_2$  are uncorrelated but not independent.

## Correlation of two random variables

The correlation coefficient of Bravais-Pearson for two random variables  $X_1$  and  $X_2$  is defined as:

$$\rho_{X_1, X_2} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var } X_1 \cdot \text{Var } X_2}}$$

## Example 6.24: see Ex. 6.21

By using  $\text{Cov}(X_1, X_2) = -0.18$  as well as  $\text{Var } X_1 = 0.630$  and  $\text{Var } X_2 = 0.480$ , we finally have:

$$\rho_{X_1, X_2} = \frac{-0.18}{\sqrt{0.630 \cdot 0.480}} = -0.3273.$$

Calculation of  $\rho_{X_1, X_2}$  in R:

```
Var_X1 <- sum(f_X1_6_17 * (X1_6_17 - Mean_X1_old)^2)
Var_X2 <- sum(f_X2_6_17 * (X2_6_17 - Mean_X2_old)^2)
Cor_X1_X2 <- Cov_X1_X2_old / (sqrt(Var_X1 * Var_X2))
round(Cor_X1_X2, digits = 4)
[1] -0.3273
```

## Properties of the correlation coefficient

- Let  $Z_1$  and  $Z_2$  be the standardised random variables of the random variables  $X_1$  and  $X_2$ . We then have:

$$\rho_{X_1, X_2} = \text{Cov}(Z_1, Z_2).$$

- Generally  $-1 \leq \rho_{X_1, X_2} \leq 1$ .
- If  $X_2 = a_0 + a_1 \cdot X_1$  and  $a_1 \neq 0$ , it follows that  $|\rho_{X_1, X_2}| = 1$  (where the reverse holds as well).
- If

$$\begin{aligned} U_1 &= a_0 + a_1 \cdot X_1 & (a_1 \neq 0) \\ U_2 &= b_0 + b_1 \cdot X_2 & (b_1 \neq 0) \end{aligned}$$

are linear transformations of the random variables  $X_1$  and  $X_2$ , we have

$$\rho_{U_1, U_2} = \text{sgn}(a_1 \cdot b_1) \cdot \rho_{X_1, X_2}.$$

## Example 6.25: (see Example 6.24)

$X_1$ : Number of red balls

$X_2$ : Number of white balls

$Y_1$ : Share of red balls

$Y_2$ : Share of white balls

We have  $Y_1 = X_1/3$  and  $Y_2 = X_2/3$ . Furthermore, we already know that  $\rho_{X_1, X_2} = -0.3273$ .

Finally, we get

a)  $\rho_{Y_1, Y_2} = -0.3273$ ,

b)  $\rho_{X_1, Y_1} = 1$ ,

c)  $\rho_{X_1, Y_2} = -0.3273$ .

## More than two random variables (1)

1. The  $n$  random variables  $X_1, \dots, X_n$  are called collectively stochastically independent, if

$$F(x_1, \dots, x_n) = F(x_1) \cdot \dots \cdot F(x_n)$$

(analogously for density and probability functions).

2. The  $n$  random variables  $X_1, \dots, X_n$  are called pairwise stochastically independent, if for two arbitrary but different random variables  $X_i$  and  $X_j$  we have:

$$F(x_i, x_j) = F(x_i) \cdot F(x_j)$$

(analogously for density and probability functions).

3. We have:  
collectively stochastically independent  $\Rightarrow$   
pairwise stochastically independent  $\Rightarrow$  pairwise uncorrelated

## More than two random variables (2)

### 4. Variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \text{Var } X_1 & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \text{Cov}(X_{n-1}, X_n) \\ \text{Cov}(X_n, X_1) & \cdots & \text{Cov}(X_n, X_{n-1}) & \text{Var } X_n \end{pmatrix}$$

If  $\Sigma$  is a diagonal matrix, then the  $n$  random variables are pairwise uncorrelated.

# Functions of random variables

## Example 6.26:

- a) Two rolls of a dice: We are interested in the overall number of pips  
 $Y = X_1 + X_2$ .

- b)  $N = 101$  balls ( $0, \dots, 100$ ):  $n$  balls are drawn with replacement.

$$Y_1 = \frac{1}{2}(X_1 + X_2)$$

$$Y_2 = \frac{1}{20}(X_1 + \dots + X_{20})$$

- c) Construction of cylindric components (technical QC):  
 $X_1$  is the component's diameter and  $X_2$  is its length. Then

$$Y = \frac{\pi}{4} \cdot X_1^2 \cdot X_2$$

is its volume.

## Expected value and variance of linearly transformed random variables

If  $Y = a_0 + \sum_{i=1}^n a_i X_i$  is a general linear transformation of  $n$  random variables, then

$$\mathbb{E} Y = a_0 + \sum_{i=1}^n a_i \mathbb{E} X_i$$

is the expected value of the transformed random variable  $Y$ . We call  $\mathbb{E}$  a linear operator! Furthermore

$$\begin{aligned}\text{Var } Y &= \sum_{i=1}^n \sum_{j=1}^n a_i \cdot a_j \cdot \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2 \cdot \text{Var } X_i + 2 \cdot \sum_{i < j} a_i \cdot a_j \cdot \text{Cov}(X_i, X_j)\end{aligned}$$

is the variance of the transformed random variable  $Y$ .

## Example 6.27: see Ex. 6.26 b) (1)

An urn contains  $N = 100$  balls (1,...,100).  $n = 3$  balls are drawn with replacement, where  $X_i$  is the number drawn in the  $i$ -th draw. Then we have:

$$\mathbb{E} X_i = \frac{1}{100}(1 + \dots + 100) = 50.5$$

$$\text{Var } X_i = \frac{1}{100}(1^2 + \dots + 100^2) - 50.5^2 = 833.25 \quad .$$

Calculation of  $\mathbb{E} X_i$  and  $\text{Var } X_i$  in R:

---

```
X6_27 <- 1:100
f_x6_27 <- rep(x = 1/100, times = 100)

Mean_X <- sum(f_x6_27 * X6_27)
Var_X <- sum(f_x6_27 * (X6_27 - Mean_X)^2)
```

---

**Mean\_X**

[1] 50.5

**Var\_X**

[1] 833.25

## Example 6.27: see Ex. 6.26 b) (2)

Now we are interested in the sample mean  $\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$ . We get:

$$E \bar{X} = \frac{1}{3}(E X_1 + E X_2 + E X_3) = 50.5$$

$$\text{Var } \bar{X} = \left(\frac{1}{3}\right)^2 (\text{Var } X_1 + \text{Var } X_2 + \text{Var } X_3) = \frac{1}{3} \cdot 833.25 = 277.75 .$$

Notice that the draws are stochastically independent (with replacement).  
In the model without replacement we would have  $E \bar{X} = 50.5$  and  
 $\text{Var } \bar{X} = 272.139$ .

## Example 6.28: see Ex. 6.21 (1)

We are now interested in  $Y = 2X_1 + 4X_2 - 1$ . We get:

$$\begin{aligned}\mathbb{E} Y &= 2\mathbb{E} X_1 + 4\mathbb{E} X_2 - 1 \\ &= 2 \cdot 0.9 + 4 \cdot 0.6 - 1 = 3.2\end{aligned}$$

and

$$\begin{aligned}\text{Var } Y &= \sum_{i=1}^2 \sum_{j=1}^2 a_i a_j \cdot \text{Cov}(X_i, X_j) \\ &= 2^2 \cdot \text{Var } X_1 + 2 \cdot 2 \cdot 4 \cdot \text{Cov}(X_1, X_2) + 4^2 \cdot \text{Var } X_2 \\ &= 4 \cdot 0.63 - 16 \cdot 0.18 + 16 \cdot 0.48 \\ &= 2.52 + 4.8 = 7.32\end{aligned}.$$

## Example 6.28: see Ex. 6.21 (2)

Calculation of  $E Y$  and  $\text{Var } Y$  in R

```
a0 <- -1  
a1 <- 2  
a2 <- 4
```

```
# ATTENTION: Means etc. from Ex. 6.21!
```

```
Mean_Y <- a0 + a1 * Mean_X1_old + a2 * Mean_X2_old
```

```
Var_Y <- a1^2 * Var_X1 + 2 * a1 * a2 * Cov_X1_X2_old +  
a2^2 * Var_X2
```

Mean\_Y

```
[1] 3.2
```

Var\_Y

```
[1] 7.32
```

# Elements of Statistics

## Chapter 7: Selected distributions

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# The Bernoulli distribution

A discrete random variable  $X$  is said to be bernoulli-distributed with parameter  $p$  ( $0 < p < 1$ ) if its probability or distribution function, respectively, satisfies:

$$be(x|p) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \\ 0 & \text{else} \end{cases} \quad \text{or} \quad Be(x|p) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - p & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

Hence, it is valid:

$$\mathbb{E} X = 0 \cdot (1 - p) + 1 \cdot p = p$$

$$\text{Var } X = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2 = p \cdot (1 - p).$$

Access on the Bernoulli distribution in R:

```
dbinom(x = c(1,0), size = 1, prob = p)
```

# The Binomial distribution (1)

A discrete random variable  $X$  is said to be distributed binomially with parameters  $n$  and  $\theta$  ( $n \in \mathbb{N}, 0 < \theta < 1$ ) if its probability function satisfies:

$$b(x|n, \theta) = \begin{cases} \binom{n}{x} \cdot \theta^x \cdot (1 - \theta)^{n-x} & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{else} \end{cases}$$

or its distribution function complies with

$$B(x|n, \theta) = \begin{cases} 0 & x < 0 \\ \sum_{\nu=0}^{[x]} \binom{n}{\nu} \cdot \theta^\nu \cdot (1 - \theta)^{n-\nu} & \text{for } x \geq 0 . \end{cases}$$

$n$  balls are drawn with replacement (WR). The probability that  $x$  balls of the type of interest are sampled is given by  $b(x|n, \theta)$  where  $\theta$  is the probability associated with each ball.

## The Binomial distribution (2)

Since the draws of  $x$  balls of the type of interest and  $n - x$  balls of the remaining type are independent of each other, we say:

1. There are  $\binom{n}{x}$  possible combinations to draw the  $x$  or rather  $n - x$  balls, given the  $n$  draws.
2. Each and every combination has a probability of  $W(x|n, \theta) = \theta^x \cdot (1 - \theta)^{n-x}$ .

A binomially distributed random variable is the sum of  $n$  bernoulli distributed random variables. Since the experiment is performed with replacement, the draws are stochastically independent ( $X = X_1 + \dots + X_n$ ). Hence:

$$\mathbb{E} X = \sum_{i=1}^n \mathbb{E} X_i = \sum_{i=1}^n \theta = n \cdot \theta$$

and

$$\text{Var } X = \sum_{i=1}^n \text{Var } X_i = \sum_{i=1}^n \theta \cdot (1 - \theta) = n \cdot \theta \cdot (1 - \theta) .$$

## Example 7.1: see ex. 6.4 (1)

Let  $\theta = \frac{3}{10} = 0.3$ . We then have for  $n = 3$

$$b(0|3; 0.3) = \binom{3}{0} \cdot 0.3^0 \cdot 0.7^3 = 0.343$$

$$b(1|3; 0.3) = \binom{3}{1} \cdot 0.3^1 \cdot 0.7^2 = 0.441$$

$$b(2|3; 0.3) = \binom{3}{2} \cdot 0.3^2 \cdot 0.7^1 = 0.189$$

$$b(3|3; 0.3) = \binom{3}{3} \cdot 0.3^3 \cdot 0.7^0 = 0.027$$

Those are the very same probabilities as reported in the previous chapter.

## Example 7.1: see ex. 6.4 (2)

The Binomial distribution in R:

```
x7_1 <- 0:3
Theta7_1 <- 0.3 ; n <- 3
f_x7_1 <- dbinom(x7_1, size = n, prob = Theta7_1)
names(f_x7_1) <- x7_1
f_x7_1
```

0	1	2	3
0.343	0.441	0.189	0.027

Furthermore, we get:

$$EX = 3 \cdot 0.3 = 0.9$$

EX in R:

```
Mean_X <- weighted.mean(x = x7_1, w = f_x7_1)
Mean_X
```

[1] 0.9

Alternatively:

```
Mean_X <- n * Theta7_1
```

## Example 7.1: see ex. 6.4 (3)

... as well as

$$\text{Var}X = 3 \cdot 0.3 \cdot 0.7 = 0.63.$$

$\text{Var}X$  in R:

```
Var <- sum(f_x7_1 * (x7_1 - Mean_X)^2)
Var
```

```
[1] 0.63
```

Alternatively:

```
Var <- n * Theta7_1 * (1 - Theta7_1)
```

# The multinomial distribution

A discrete random variable  $X$  is said to follow a multinomial distribution with parameters  $n$  and  $\theta_1, \dots, \theta_k$  ( $n \in \mathbb{N}, 0 < \theta < 1$  and  $\theta_1 + \dots + \theta_k = 1$ ) if its probability function complies to:

$$m(x_1, \dots, x_k | n, \theta_1, \dots, \theta_k) = \begin{cases} \frac{n!}{x_1! \cdot \dots \cdot x_k!} \cdot \theta_1^{x_1} \cdot \dots \cdot \theta_k^{x_k} & \text{for } x_i \in \mathbb{N}_0 \text{ and} \\ & x_1 + \dots + x_k = n \\ 0 & \text{else.} \end{cases}$$

In practice, the parameter  $n$  of the function  $m$  is mostly neglected due to the relation  $x_1 + \dots + x_k = n$ .

The binomial distribution is a special case of the multinomial distribution with probabilities  $(\theta, 1 - \theta)$  associated with the two possible outcomes  $(x, n - x)$ .

## Example 7.2: see Ex. 6.16

Let now  $\theta_1 = 0.3$ ,  $\theta_2 = 0.2$  and  $\theta_3 = 0.5$  (balls' colours). Then

$$f(1, 0) = m(\underbrace{1, 0, 2}_{=3} \mid \underbrace{0.3; 0.2; 0.5}_{=1}) = \frac{3!}{1! \cdot 0! \cdot 2!} \cdot 0.3^1 \cdot 0.2^0 \cdot 0.5^2 = 0.225.$$

$f(1, 0)$  (Probability for one red and no white ball) in R:

```
x7_2 <- c(1, 0, 2); n <- 3
Theta7_2 <- c(0.3, 0.2, 0.5)
f_1_0_2 <- dmultinom(x7_2, size = n, prob = Theta7_2)
f_1_0_2
```

[1] 0.225

The probability  $f(2, 2) = m(2, 2, 0 | 0.3, 0.2, 0.5)$  can be derived from the *else* node of the probability function of the multinomial distribution because of  $2 + 2 + 0 > n$ .

# The hypergeometric distribution

A discrete random variable  $X$  is said to follow a hypergeometric distribution with parameters  $n$ ,  $M$  and  $N$  ( $n, M, N \in \mathbb{N}$  with  $n \leq N$  and  $M < N$ ) if its probability function complies to:

$$h(x|n, N, M) = \begin{cases} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} & \text{for } x = 0, \dots, n \text{ as well as} \\ & \max(0, M+n-N) \leq x \leq \min(n, M) \\ 0 & \text{else} \end{cases}$$

$M$  equals the number of balls of the type of interest sampled from an urn containing  $N$  balls in total ( $\theta = \frac{M}{N}$ ). Furthermore,  $n$  balls are drawn without replacement (WOR). Thus:

$$\mathbb{E} X = n \cdot \frac{M}{N} \quad \text{and} \quad \text{Var } X = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}.$$

## Example 7.3: Another urn example (1)

Given an urn containing  $N = 100$  balls of which  $M = 30$  are white.  $n = 3$  balls are drawn from the urn **without replacement**. The resulting probabilities are:

$$h(0|3; 0.3) = \frac{\binom{30}{0} \cdot \binom{70}{3}}{\binom{100}{3}} = 0.3385 \quad b(0|3; 0.3) = 0.343$$

$$h(1|3; 0.3) = \frac{\binom{30}{1} \cdot \binom{70}{2}}{\binom{100}{3}} = 0.4481 \quad b(1|3; 0.3) = 0.441$$

$$h(2|3; 0.3) = \frac{\binom{30}{2} \cdot \binom{70}{1}}{\binom{100}{3}} = 0.1883 \quad b(2|3; 0.3) = 0.189$$

$$h(3|3; 0.3) = \frac{\binom{30}{3} \cdot \binom{70}{0}}{\binom{100}{3}} = 0.0251 \quad b(3|3; 0.3) = 0.027$$

## Example 7.3: Another urn example (2)

The hypergeometric distribution in R

```
x7_3 <- 0:3
n <- 3
N7_3 <- 100
M7_3 <- 30
Theta7_3 <- M7_3 / N7_3

f_x7_3 <- dhyper(x7_3, m = M7_3, n = N7_3 - M7_3, k = n)
names(f_x7_3) <- x7_3
round(f_x7_3, digits = 3)
```

	0	1	2	3
0.339	0.339	0.448	0.188	0.025

# The Poisson distribution

A discrete random variable  $X$  is said to be poisson distributed with parameter  $\lambda (\lambda > 0)$  if its probability function satisfies:

$$po(x|\lambda) = \begin{cases} \frac{e^{-\lambda} \cdot \lambda^x}{x!} & \text{for } x = 0, 1, \dots \\ 0 & \text{else} \end{cases}$$

We have:

$$EX = \lambda \quad \text{and} \quad \text{Var } X = \lambda \quad .$$

The poisson distribution is used to model rare events, e.g. natural disasters. In addition to that, it is the limit distribution of the binomial distribution.

## Example 7.4: Earthquake (1)

Within a certain region an earthquake occurs more or less every ten years. We are interested in the probability that this specific region experiences three earthquakes within one single year.

The information 'once in every ten years' translates to  $\lambda = 0.1$ , hence it follows:

$$\begin{aligned} P(X \geq 3) &= 1 - W(X \leq 2) = 1 - Po(2|0.1) \\ &= 1 - \left( \frac{0.1^0}{0!} \cdot e^{-0.1} + \frac{0.1^1}{1!} \cdot e^{-0.1} + \frac{0.1^2}{2!} \cdot e^{-0.1} \right) \\ &= 1 - \left( \frac{0.1^0}{0!} + \frac{0.1^1}{1!} + \frac{0.1^2}{2!} \right) \cdot e^{-0.1} \\ &= 1 - 1.105 \cdot e^{-0.1} \\ &\approx 0.000155. \end{aligned}$$

## Example 7.4: Earthquake (2)

Calculation of  $W(X \geq 3)$  in R:

```
x7_4 <- 0:5
Lambda7_4 <- 0.1
f_x7_4 <- dpois(x = x7_4, lambda = Lambda7_4)
names(f_x7_4) <- x7_4
F_x7_4 <- ppois(q = x7_4, lambda = Lambda7_4)
names(F_x7_4) <- x7_4
round(f_x7_4, digits = 6)
```

0	1	2	3	4	5
0.904837	0.090484	0.004524	0.000151	0.000004	0.000000

F\_x7\_4

0	1	2	3	4	5
0.9048374	0.9953212	0.9998453	0.9999962	0.9999999	1.0000000

```
Prob <- 1 - F_x7_4["2"]
Prob_old <- sum(f_x7_4[4:6])
```

```
round(Prob, digits=6)
[1] 0.000155
```

```
Prob_old
[1] 0.000155
```

# Interrelation between poisson and binomial distribution

We have:

$$\lim_{\substack{n \rightarrow \infty \\ \theta \rightarrow 0}} b(x|n, \theta) = po(x|\lambda) \quad ,$$

where  $\lambda = n \cdot \theta$ . Thus the expected values are identical and the variances are approximatively identical..

## Example 7.5:

Suppose  $\lambda = 2 = n \cdot \theta$ :

Distribution	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Var X
$b(x 10; 0.2)$	0.107	0.268	0.302	0.201	0.088	0.026	0.006	1.6
$b(x 50; 0.04)$	0.130	0.271	0.276	0.184	0.090	0.035	0.011	1.92
$b(x 100; 0.02)$	0.133	0.271	0.273	0.182	0.090	0.035	0.011	1.96
$b(x 200; 0.01)$	0.134	0.271	0.272	0.181	0.090	0.036	0.012	1.98
$b(x 500; 0.004)$	0.135	0.271	0.271	0.181	0.090	0.036	0.012	1.992
$b(x 1000; 0.002)$	0.135	0.271	0.271	0.181	0.090	0.036	0.012	1.996
$p(x 2)$	0.135	0.271	0.271	0.180	0.090	0.036	0.012	2

# Interrelation between binomial and hypergeometric distribution

The distributions differ due to drawing with/without replacement. For smaller samples ( $n/N$ ) we thus have negligible differences.

## Example 7.6:

For  $n = 6$  and  $M/N = 0.2$  we get:

Distribution	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	$\text{Var } X$
$h(x 6, 25, 5)$	0.219	0.438	0.274	0.064	0.005	0.000	0.000	0.76
$h(x 6, 50, 10)$	0.242	0.414	0.259	0.075	0.010	0.001	0.000	0.862
$h(x 6, 100, 20)$	0.252	0.403	0.252	0.079	0.013	0.001	0.000	0.912
$h(x 6, 200, 40)$	0.257	0.398	0.249	0.080	0.014	0.001	0.000	0.936
$h(x 6, 500, 100)$	0.260	0.395	0.247	0.081	0.015	0.001	0.000	0.950
$h(x 6, 100, 200)$	0.261	0.394	0.246	0.082	0.015	0.001	0.000	0.955
$b(x 6, 0.2)$	0.262	0.393	0.246	0.082	0.015	0.002	0.000	0.96

An approximation of the hypergeometric distribution with the Poisson distribution can be done in two steps using the binomial distribution.

## The exponential distribution

A continuous random variable  $X$  is said to be exponentially distributed with parameter  $\lambda$  ( $\lambda > 0$ ) if its density function satisfies:

$$f(x|\lambda) = \begin{cases} \lambda \cdot \exp(-\lambda \cdot x) & \text{for } x > 0 \\ 0 & \text{else} \end{cases}$$

We have:

$$\mathbb{E} X = \frac{1}{\lambda} \quad \text{and} \quad \text{Var } X = \frac{1}{\lambda^2} \quad .$$

The exponential distribution is used to model lifespans, e.g. the one of light bulbs. It is a distribution *without* memory (see Schira 2005, p. 365.)

Access on the exponential distribution in R:

```
dexp(x, rate = Lambda)
```

# The normal distribution

A continuous random variable  $X$  is called normally distributed with parameters  $\mu$  and  $\sigma^2$  ( $\mu \in \mathbb{R}$  and  $\sigma > 0$ ), if it exhibits the following density function:

$$\varphi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for all } x \in \mathbb{R}.$$

We use the short hand  $N(\mu, \sigma^2)$ .

We have:

$$\mathbb{E} X = \mu \quad \text{and} \quad \text{Var } X = \sigma^2 \quad .$$

The expected value and the variance are the explicit parameters of the normal distribution.

## Example 7.7: Normal distributions (1)

In the following, we will illustrate the normal distributions  $N(0, 1)$ ,  $N(5, 1)$ ,  $N(-3, 0.35^2)$  and  $N(5, 2^2)$  graphically.

Density function of the normal distribution in R:

```
x7_7 <- seq(from = -5, to = 10, by = 0.00001)

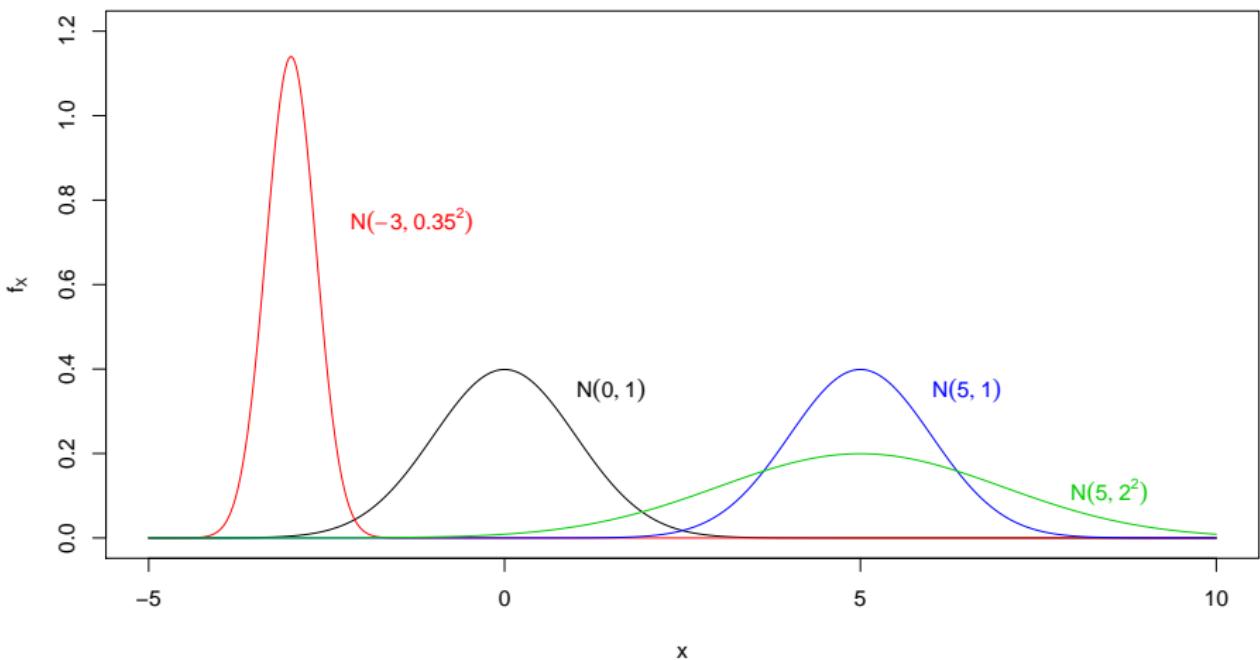
f_x7_7_1 <- dnorm(x = x7_7, mean = 0, sd = 1)
f_x7_7_2 <- dnorm(x = x7_7, mean = 5, sd = 1)
f_x7_7_3 <- dnorm(x = x7_7, mean = -3, sd = 0.35)
f_x7_7_4 <- dnorm(x = x7_7, mean = 5, sd = 2)
```

Illustration in R:

```
plot(x = x7_7, y = f_x7_7_1, type = "l", xlab = "x",
      ylab = "f(x)", ylim = c(0, 1.2))
lines(x = x7_7, y = f_x7_7_2, type = "l", col = "blue")
lines(x = x7_7, y = f_x7_7_3, type = "l", col = "red")
lines(x = x7_7, y = f_x7_7_4, type = "l", col = "green")
```

## Example 7.7: Normal distributions (2)

Normal distributions  $N(0, 1)$ ,  $N(5, 1)$ ,  $N(-3, 0.35^2)$  and  $N(5, 2^2)$



# Properties of the normal distribution

1. The normal distribution is symmetric to the axis  $x = \mu$ .
2. The graph of the normal distribution reaches its maximum at the point  $\left(\mu, \frac{1}{\sqrt{2\pi}\sigma}\right)$ .
3. The inflection points of the normal distribution are at  $x = \mu - \sigma$  and  $x = \mu + \sigma$ .
4. The distribution function of the normal distribution

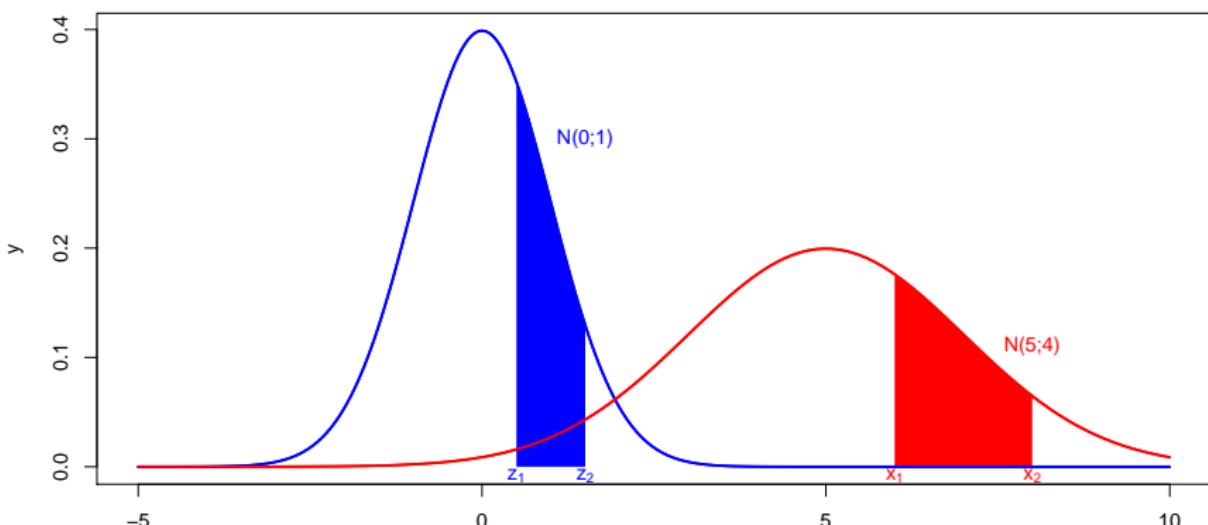
$$\Phi(x|\mu, \sigma^2) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right) ds$$

cannot be given in a closed-form expression.

# Standard normal distribution and standard transformation

If  $X \sim N(\mu, \sigma^2)$ , then  $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ , where  $N(0, 1)$  is the standard normal distribution.

Particularly  $P(x_1 < X \leq x_2) = P(z_1 < Z \leq z_2) = \Phi(z_2) - \Phi(z_1)$ , with  $z_1 = \frac{x_1-\mu}{\sigma}$  and  $z_2 = \frac{x_2-\mu}{\sigma}$ .



## Example 7.8: Standard normal distribution (1)

Using tabulated values of the standard normal distribution or R, we reach:

$$\begin{aligned}\Phi(1) &= 0.8413 & \Phi(1.96) &= 0.9750 & \Phi(3.29) &= 0.9995 \\ \Phi(-1) &= 1 - \Phi(1) = 0.1587 & \Phi(-1.96) &= 0.0250 & \Phi(-3.29) &= 0.0005\end{aligned}$$

### Calculation of the values of the distribution function in R

```
F_x7_8 <- pnorm(q = c(1, 1.96, 3.29, -1, -1.96, -3.29),  
                  mean = 0, sd = 1)  
round(F_x7_8, digits = 4)
```

```
[1] 0.8413 0.9750 0.9995 0.1587 0.0250 0.0005
```

## Example 7.8: Standard normal distribution (2)

From this, we obtain the following probabilities:

$$\begin{aligned}P(1 \leq Z \leq 2) &= \Phi(2) - \Phi(1) \\&= 0.9772 - 0.8413 = 0.1359\end{aligned}$$

$$\begin{aligned}P(-2 \leq Z \leq -1) &= \Phi(-1) - \Phi(-2) = 1 - \Phi(1) - (1 - \Phi(2)) \\&= 0.1359\end{aligned}$$

$$\begin{aligned}P(-2 \leq Z \leq 1) &= \Phi(1) - (1 - \Phi(2)) = \Phi(1) + \Phi(2) - 1 \\&= 0.8185\end{aligned}$$

$$\begin{aligned}P(Z \geq -1) &= 1 - \Phi(-1) = 1 - (1 - \Phi(1)) = \Phi(1) \\&= 0.8413\end{aligned}$$

Calculation of  $P(1 \leq Z \leq 2)$  in R:

```
Prob7_8 <- pnorm(q=2, mean=0, sd=1) - pnorm(q=1, mean=0, sd=1)
round(Prob7_8, digits=4)
```

```
[1] 0.1359
```

## Example 7.9: Simplifications (1)

Let a random variable  $X$  be  $N(20; 64)$  distributed.

$$\Phi(28|20; 64) = \Phi\left(\frac{28 - 20}{8}\right) = \Phi(1) = 0.8413$$

$$\Phi(12|20; 64) = \Phi\left(\frac{12 - 20}{8}\right) = \Phi(-1) = 0.1587$$

$$\Phi(35.68|20; 64) = \Phi\left(\frac{35.68 - 20}{8}\right) = \Phi(1.96) = 0.9750$$

$$\Phi(4.32|20; 64) = \Phi(-1.96) = 0.0250$$

$$\Phi(46.32|20; 64) = \Phi(3.29) = 0.9995$$

$$\Phi(-6.32|20; 64) = \Phi(-3.29) = 0.0005$$

## Example 7.9: Simplifications (2)

Calculations in R:

```
Phi7_9 <- pnorm(q = c(28, 12, 35.68, 4.32, 46.32, -6.32),  
                  mean = 20, sd = sqrt(64))  
round(Phi7_9, digits=4)  
[1] 0.8413 0.1587 0.9750 0.0250 0.9995 0.0005
```

Alternative calculation with the standard normal distribution in R:

```
Phi7_9 <- pnorm(q = c(1, -1, 1.96, -1.96, 3.29, -3.29),  
                  mean = 0, sd = 1)
```

$$P(28 \leq X \leq 36) = 0.1359$$

$$P(4 \leq X \leq 12) = 0.1359$$

$$P(4 \leq X \leq 28) = 0.8185$$

$$P(X \geq 12) = 0.8413$$

Calculation of  $P(28 \leq X \leq 36)$  in R:

```
Prob7_9 <- pnorm(36, mean=20, sd=8) - pnorm(28, mean=20, sd=8)  
round(Prob7_9, digits=4)  
[1] 0.1359
```

## Example 7.10: Quantiles of the normal distribution (1)

Find the  $p$ -quantiles of  $N(0, 1)$  and  $N(6, 2^2)$  for  $p = 0.8; 0.9; 0.95; 0.15; 0.01$ .

As  $x(p) = \mu + \sigma \cdot z(p)$  we have:

$$z(0.8) = 0.84$$

$$z(0.9) = 1.28$$

$$z(0.95) = 1.645$$

$$z(0.15) = -z(0.85) = -1.04$$

$$z(0.01) = -z(0.99) = -2.33$$

$$x(0.8) = 6 + 2 \cdot 0.84 = 7.68$$

$$x(0.9) = 6 + 2 \cdot 1.28 = 8.56$$

$$x(0.95) = 6 + 2 \cdot 1.645 = 9.29$$

$$x(0.15) = 6 + 2 \cdot (-1.04) = 3.92$$

$$x(0.01) = 6 + 2 \cdot (-2.33) = 1.34$$

## Example 7.10: Quantiles of the normal distribution (2)

Determination of the quantiles in R:

```
Quantile_z7_10 <- qnorm(p = c(0.8,0.9,0.95,0.15,0.01),  
                         mean = 0, sd = 1)  
Quantile_x7_10 <- qnorm(p = c(0.8,0.9,0.95,0.15,0.01),  
                         mean = 6, sd= 2)  
round(Quantile_z7_10, digits=2)  
[1] 0.84 1.28 1.64 -1.04 -2.33
```

```
round(Quantile_x7_10, digits=2)  
[1] 7.68 8.56 9.29 3.93 1.35
```

Alternative calculation for  $N(6, 2^2)$ :

```
Quantile_x7_10_alternative <- 6 + 2 * Quantile_z7_10  
round(Quantile_x7_10_alternative, digits=2)  
[1] 7.68 8.56 9.29 3.93 1.35
```

## Linear functions of normally distributed random variables

Let the random variables  $X_1, \dots, X_n$  all be  $N(\mu_i, \sigma_i^2)$  ( $i = 1, \dots, n$ ) distributed and overall stochastically independent. Then the linear combination

$$Y = a_0 + a_1 \cdot X_1 + \dots + a_n \cdot X_n$$

is normally distributed with

$$N\left(a_0 + \sum_{i=1}^n a_i \cdot \mu_i; \quad \sum_{i=1}^n a_i^2 \cdot \sigma_i^2\right) \quad .$$

Particularly, the arithmetic mean of identically distributed variables ( $X_i \sim N(\mu; \sigma^2)$ ) is

$$N\left(\mu; \frac{\sigma^2}{n}\right)$$

distributed.

## Example 7.11: Linear combination

Let the random variables  $X_1, X_2, X_3$  be  $N(2; 1^2)$ ,  $N(5; 2^2)$  and  $N(8; 3^2)$  distributed and stochastically independent. Furthermore, let

$$Y = 30 + 5X_1 + 2X_2 - 5X_3 \quad .$$

We get

$$\mathbb{E} Y = 30 + 5 \cdot 2 + 2 \cdot 5 - 5 \cdot 8 = 10$$

and

$$\text{Var } Y = 5^2 \cdot 1^2 + 2^2 \cdot 2^2 + (-5)^2 \cdot 3^2 = 266 \quad .$$

Therefore,  $Y$  is  $N(10; 266)$  distributed.

Notice the difference:

An interviewer questions 10 people or he questions one person and replicates the result 10 times.

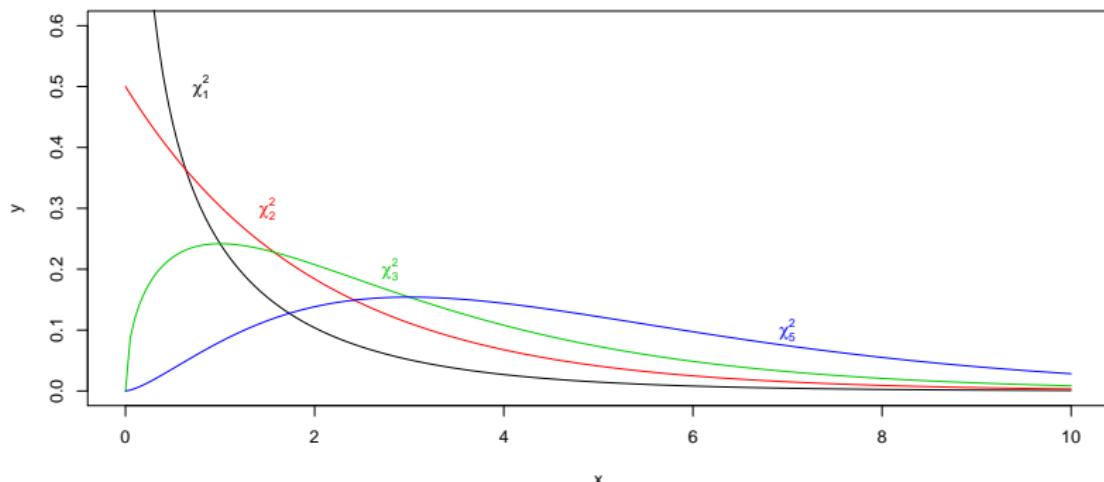
# The $\chi^2$ distribution

The sum of  $k$  squared, stochastically independent random variables  $Z_i$  ( $i = 1, \dots, k$ ) following a standard normal distribution

$$Y = Z_1^2 + Z_2^2 + \dots + Z_k^2$$

is  $\chi^2$ -distributed with  $k$  degrees of freedom.

We have  $E Y = k$  and  $\text{Var } Y = 2 \cdot k$ . In R: `dchisq(x, df = k)`



## Example 7.12: Probabilities with the $\chi^2$ distribution

Let the random variable  $Y$  be  $\chi^2$ -distributed with  $k = 10$  degrees of freedom. We have:

$$P(Y \leq 3.94) = F(3.94|10) = 0.05$$

```
round(pchisq(q = 3.94, df = 10), digits = 2)
[1] 0.05
```

$$P(Y \geq 15.987) = 1 - F(15.987|10) = 0.1$$

```
round(1 - pchisq(q = 15.987, df = 10), digits = 2)
[1] 0.1
```

$$P(4.87 \leq Y \leq 20.48) = F(20.48|10) - F(4.87|10) = 0.975 - 0.1 = 0.875$$

```
round(pchisq(q = 20.48, df = 10) - pchisq(q = 4.87, df = 10),
      digits = 3)
[1] 0.875
```

# Approximation using the normal distribution

1. If  $Y \sim \chi^2_k$ , then

$$\sqrt{2 \cdot Y} - \sqrt{2 \cdot k - 1} \sim N(0, 1) \quad \text{for } k > 30$$

2.  $\chi^2_k \sim N(k; 2 \cdot k)$  for  $k > 100$

## Example 7.13: Approximation (1)

1. Let  $Y \sim \chi^2_{40}$ . Then we have

$$P(Y \leq 55.758) = F(55.758|40) = 0.95 \quad \text{or}$$

$$F(55.758|40) = \Phi(\sqrt{2 \cdot 55.758} - \sqrt{2 \cdot 40 - 1}) = \Phi(1.67) = 0.9525.$$

2. Let  $Y \sim \chi^2_{200}$ . Then we have

$$P(Y \leq 220) = F(220|200) = 0.8417$$

Calculation of  $P(Y \leq 220)$  in R:

```
Prob7_13 <- pchisq(q = 220, df = 200)
round(Prob7_13, digits = 4)
```

[1] 0.8417

## Example 7.13: Approximation (2)

Alternative 1:

$$F(220|200) = \Phi\left(\sqrt{2 \cdot 220} - \sqrt{2 \cdot 200 - 1}\right) = \Phi(1,00) = 0.8416$$

```
Prob7_13_1 <- pnorm(q = sqrt(2*220)-sqrt(2*200-1),  
                      mean = 0, sd = 1)  
round(Prob7_13_1, digits = 4)
```

[1] 0.8416

Alternative 2:

$$F(220|200) = \Phi(220|200; 400) = \Phi\left(\frac{220 - 200}{\sqrt{400}}\right) = \Phi(1,00) = 0.8413$$

```
Prob7_13_2 <- pnorm((220-200)/sqrt(400), mean = 0, sd = 1)  
round(Prob7_13_2, digits=4)
```

[1] 0.8413

# The $t$ distribution (1)

Let  $Z$  follow a standard normal distribution and let  $Y$  be  $\chi_k^2$ -distributed. Furthermore, let  $Z$  and  $Y$  be stochastically independent. Then, the random variable

$$T = \frac{Z}{\sqrt{Y/k}}$$

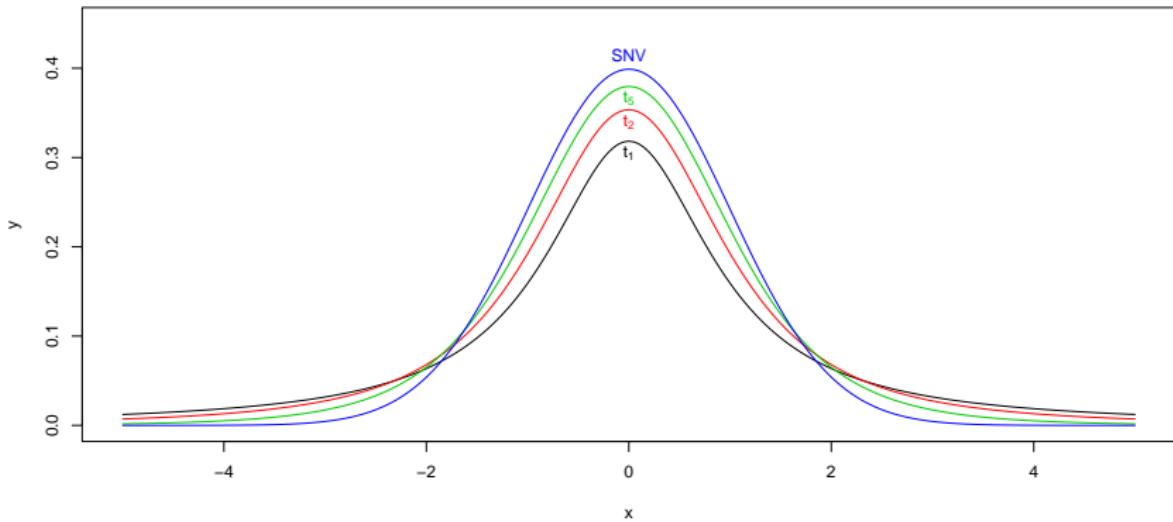
follows a  $t$ -distribution with  $k$  degrees of freedom.

We have  $E T = 0$  ( $k > 1$ ) and  $\text{Var } T = \frac{k}{k-2}$  ( $k > 2$ ) as well as  $t(1-p, k) = -t(p, k)$ .

Access on the  $t$  distribution in R:

```
dt(x, df = k)
```

# The $t$ distribution (2)

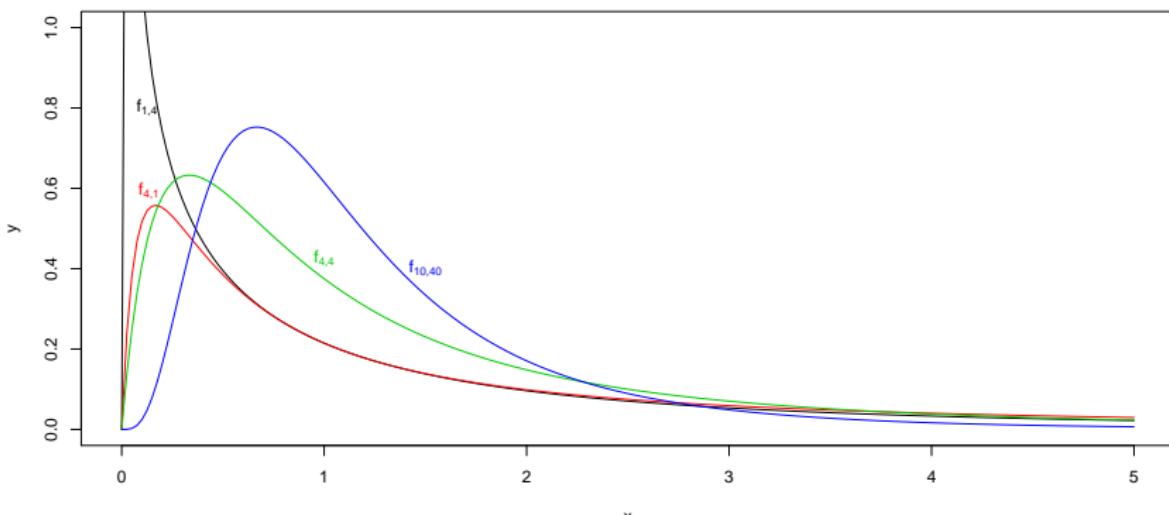


# The $F$ distribution (1)

Let the random variables  $Y_1$  and  $Y_2$  be  $\chi^2$ -distributed with  $k_1$  and  $k_2$  degrees of freedom, respectively. Then, the random variable

$$X = \frac{Y_1/k_1}{Y_2/k_2}$$

follows an  $F$ -distribution with  $(k_1, k_2)$  degrees of freedom.



## The $F$ distribution (2)

We have:

$$\begin{aligned} E X &= \frac{k_2}{k_2 - 2} && \text{for } k_2 > 2 \\ \text{Var } X &= \frac{2(k_1 + k_2 - 2)}{k_1(k_2 - 4)} \left( \frac{k_2}{k_2 - 2} \right)^2 && \text{for } k_2 > 4. \end{aligned}$$

Furthermore, we have

$$f(p; k_1, k_2) = \frac{1}{f(1-p; k_2, k_1)}$$

for the  $p$ - and  $(1-p)$ -quantile of the  $F$  distribution, respectively.

Access on the  $F$  distribution in R:

```
df(x, df1 = k_1, df2 = k_2)
```

# Law of large numbers

## Bernoulli's law of large numbers

Assume  $\{X_n\}$  to be a sequence of binomially distributed random variables with parameters  $n$  and  $\theta$  ( $0 < \theta < 1$ ). This sequence  $\{X_n\}$  satisfies the *weak law of large numbers* if

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_n}{n} - \theta\right| > \varepsilon\right) = 0$$

is valid for each  $\varepsilon > 0$ .

- ▶ The law of large numbers constitutes the foundation of the statistical concept of probability.
- ▶ Practical application of the law: Suppose the very same experiment is conducted with a sufficiently larger number of repetitions. With increasing  $n$ , the sample mean of this repetitions will approach the theoretical mean.

## Central limit theorem

Assume  $\{X_n\}$  to be a sequence of independent random variables with existing and positive variances.

In addition,

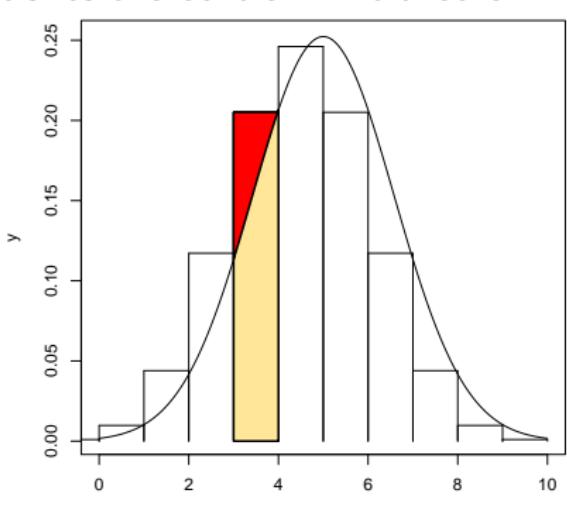
$$Y_n := \frac{\sum_{i=1}^n (X_i - \mathbb{E} X_i)}{\sqrt{\sum_{i=1}^n \text{Var } X_i}}$$

is the standardized sum variable and the random variable  $Y$  follows a standard normal distribution. The sequence of random variables  $\{X_n\}$  satisfies the *central limit theorem* if

$$\lim_{n \rightarrow \infty} F_{Y_n}(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot s^2\right) ds .$$

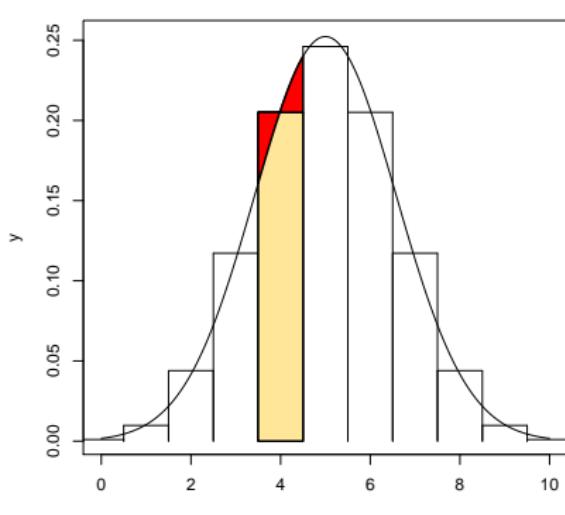
# Theorem of De Moivre and Laplace

The random variables  $X_i$  are bernoulli-distributed with parameter  $\theta$  ( $0 < \theta < 1$ ) and stochastically independent. The resulting sequence  $\{Y_n\}$  satisfies the central limit theorem.



$$b_X(4|10; 0.5) = 0.2051$$

$$\Phi_X(4.5|5; 2.5) - \Phi_X(3.5|5; 2.5) = 0.1606$$
$$\Phi_X(4.5|5; 2.5) - \Phi_X(3.5|5; 2.5) = 0.2045$$

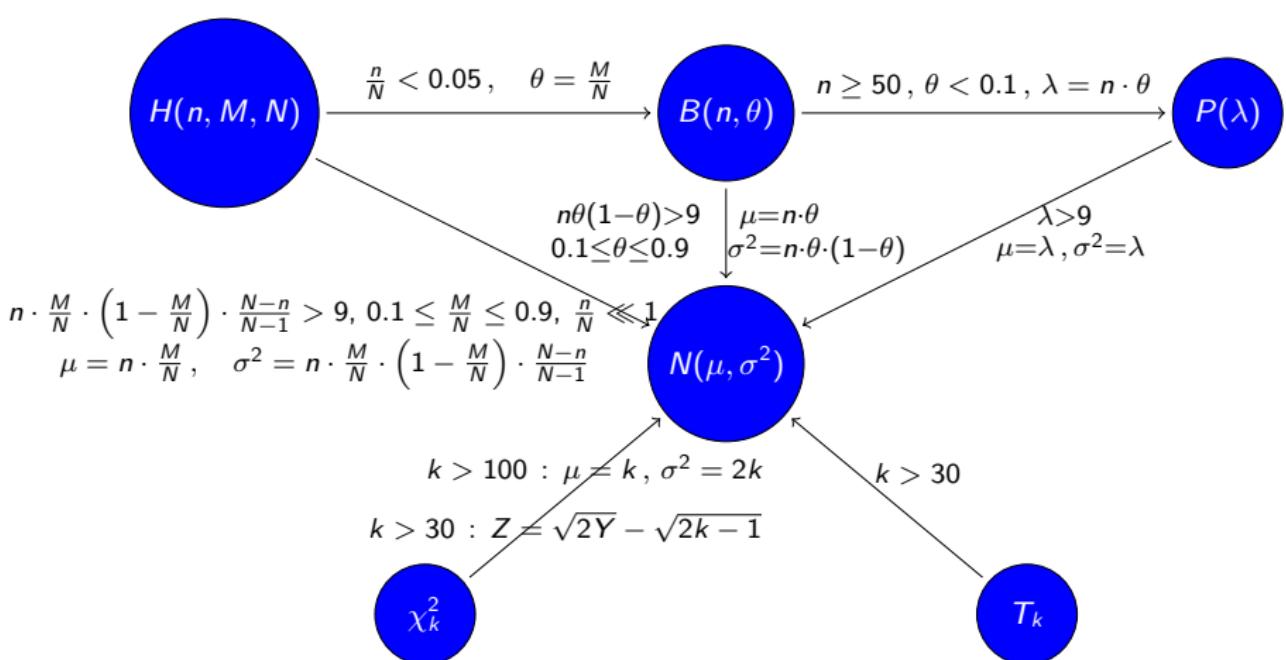


## Theorem of Lindeberg and Lévy

$\{X_n\}$  being a sequence of independent identically distributed random variables with existing and positive variances. Thus the sequence  $\{X_n\}$  satisfies the central limit theorem.

- ▶ In contrast to the theorem of de Moivre and Laplace, there is no specific distribution required
- ▶ In practice, approximations for  $n$  sufficiently large
- ▶ Convergence is reached despite of very different distributions
- ▶ *Technically* only applicable to sampling models **with replacement**
- ▶ There are numerous other central limit theorems

# Options of approximations between distributions



## Continuity correction binomial approximation

$$\begin{aligned} P(X \leq x) &= B(x|n, \theta) \\ &= \Phi(x + 0.5|n\theta, n\theta(1 - \theta)) \end{aligned}$$

$$\begin{aligned} P(x_1 < X \leq x_2) &= B(x_2|n, \theta) - B(x_1|n, \theta) \\ &= \Phi(x_2 + 0.5|n\theta, n\theta(1 - \theta)) - \Phi(x_1 + 0.5|n\theta, n\theta(1 - \theta)) \end{aligned}$$

$$\begin{aligned} P(X \geq x) &= 1 - B(x - 1|n, \theta) \\ &= 1 - \Phi(x - 1 + 0.5|n\theta, n\theta(1 - \theta)) \\ &= 1 - \Phi(x - 0.5|n\theta, n\theta(1 - \theta)) \end{aligned}$$

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= B(x_2|n, \theta) - B(x_1 - 1|n, \theta) \\ &= \Phi(x_2 + 0.5|n\theta, n\theta(1 - \theta)) - \Phi(x_1 - 0.5|n\theta, n\theta(1 - \theta)) \end{aligned}$$

## Example 7.14: Continuity correction (1)

The sample size  $n = 100$  and the parameter  $\theta = 0.36$  are given. Initially, the conditions of approximation have to be checked.

$$100 \cdot 0.36 \cdot 0.64 = 23.04 > 9 \quad \checkmark$$

Now we have  $\mu = 100 \cdot 0.36 = 36$  and  $\sigma^2 = 23.04 = 4.8^2$ .

Calculation of  $\mu$  and  $\sigma^2$  in R:

```
n <- 100
Theta7_14 <- 0.36

Mean7_14 <- n * Theta7_14
Var <- n * Theta7_14 * (1 - Theta7_14)
```

Mean7\_14

[1] 36

Var

[1] 23.04

## Example 7.14: Continuity correction (2)

With this, we get

$$\begin{aligned} P(X \leq 40) &= B(40|100; 0.36) \\ &\approx \Phi(40.5|36; 4.8^2) = \Phi(0.94) = 0.8257, \end{aligned}$$

Three possible ways to calculate  $P(X \leq 40)$  in R:

```
Prob7_14_1 <- pbinom(q = 40, size = n, prob = Theta7_14)
Prob7_14_2 <- pnorm(q = (40.5-Mean7_14)/sqrt(Var),
                      mean = 0, sd = 1)
Prob7_14_3 <- pnorm(q = 40.5, mean = Mean7_14, sd = sqrt(Var))
```

```
round(Prob7_14_1, 4)
[1] 0.8261
```

```
round(Prob7_14_3, 4)
[1] 0.8257
```

## Example 7.14: Continuity correction (3)

... as well as

$$\begin{aligned}P(X \geq 40) &= 1 - B(39|100; 0.36) \\&\approx 1 - \Phi(39.5|36; 4.8^2) = 1 - \Phi(0.73) = 0.2329.\end{aligned}$$

Three possible ways to calculate  $P(X \geq 40)$  in R:

```
Prob7_14_4 <- 1 - pbinom(q=39, size=n, prob=Theta7_14)
Prob7_14_5 <- 1 - pnorm(q=(39.5-Mean7_14)/sqrt(Var),
                           mean=0, sd=1)
Prob7_14_6 <- 1 - pnorm(q=39.5, mean=Mean7_14, sd=sqrt(Var))

round(Prob7_14_4, 4)                         round(Prob7_14_6, 4)
[1] 0.2316                                     [1] 0.2329
```

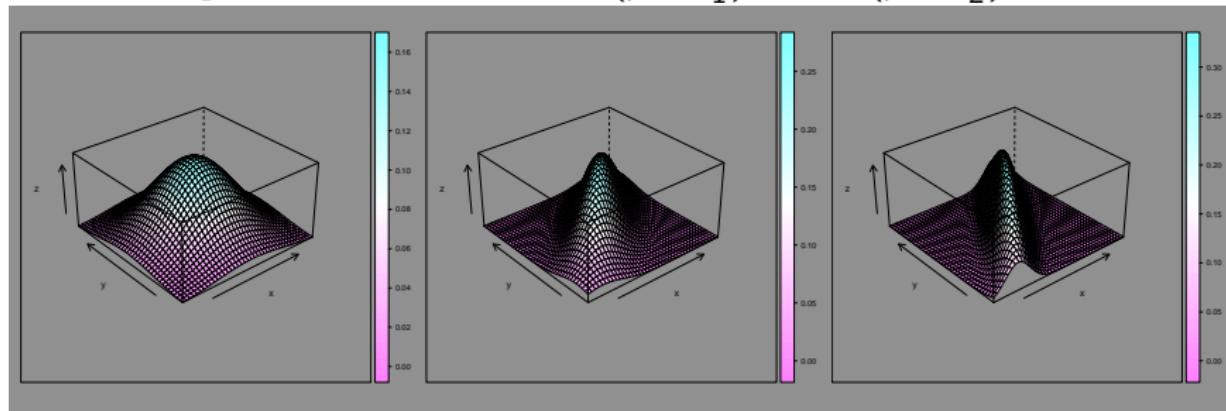
Suppose a hypergeometric distribution (model **without replacement**,  $N = 1000$ , hence  $0.1 \ll 1$ ) is used instead, the additional correction factor would lead to an approximation by  $N(36; 20.76)$ .

# Two-dimensional normal distribution

The two-dimensional normal distribution has the density function ( $|\rho| < 1$ )

$$\varphi(x_1, x_2 | \mu_1; \mu_2; \sigma_1^2; \sigma_2^2; \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right]\right).$$

The two marginal distributions are  $N(\mu_1; \sigma_1^2)$  and  $N(\mu_2; \sigma_2^2)$ .



# Representativeness of samples

A sample is called **representative** with respect to a population if its result can be transferred to the population in a suitable manner, e.g. it represents the same properties and proportions as that of the population of interest.

- ▶ **Small image** of the population
- ▶ Structures of the characteristics of interest have to be borne in mind
- ▶ **Sampling frame** has to be known and accessible
- ▶ **Random sampling** with known **inclusion probabilities**
- ▶ One should be aware of non-sampling errors
- ▶ A suitable **estimation methodology** is required

# Probability sampling procedures

- ▶ Simple random sampling
  - ▶ Model with replacement
  - ▶ Model without replacement
- ▶ Two-stage sampling processes
  - ▶ Stratified sampling
  - ▶ Cluster sampling
  - ▶ Special two-stage processes
- ▶ Methods with unequal sampling probabilities

Please note:

- ▶ Methods with fixed sampling sizes
- ▶ Methods with alterable sampling sizes

# Non-probability sampling procedures

- ▶ Conscious sampling
- ▶ Quota sampling
  - e.g.: income and consumption sample
    - ▶ Demographic features used for adjustments:  
gender, age, regional affiliation
    - ▶ Correlation between study- and quota characteristics
    - ▶ Non-random sampling within quota cells
- ▶ Concentration sample
  - Elements with high concentration
  - Projection / distribution
- ▶ Snowball sampling

# Simple random sample

## Random sample

A process that draws  $n$  elements from a population of size  $N$  ( $n < N$ ) in a random and successive manner is called **random sampling procedure**.

The result of such a method is a **random sample** or **probability sample**.

## Simple random sample

If the random experiment corresponds to the urn model with or without replacement it is called a **simple random sample** (SRS).

# Sampling with and without replacement

## Sampling with replacement

There are  $N^n$  possible different samples when drawing elements with replacement (WR). Each sample might be drawn with identical probability. Each element can be drawn 0 up to  $n$  times. The individual draws are **stochastically independent**.

## Sampling without replacement

There are  $N!/(N - n)!$  possible different samples when drawing elements without replacement (WOR). Each sample might be drawn with identical probability. Each element can be drawn 0 or 1 times. The individual draws are **stochastically dependent**.

# Sampling functions and distributions

A sample function is a function

$$u(x_1, x_2, \dots, x_n),$$

which evaluates the realisations of a variable of interest observed in a sample.

Examples:

- ▶ Sample mean:  $\bar{x}$
- ▶ Sample proportion:  $p$
- ▶ Sampling variance:  $s^2$

Which distributions do these sample functions follow?

Distributions of the random variables  $\bar{X}$ ,  $P$  and  $S^2$ , respectively.

## Sample mean under SRS

We observe  $n$  realisations  $x_1, \dots, x_n$  of study variable  $X$ .

The sample mean is calculated as:  $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$

Expected value (WR and WOR):

$$\mathbb{E} \bar{X} = \mathbb{E} \left( \frac{1}{n} \cdot (X_1 + \dots + X_n) \right) = \frac{1}{n} \cdot n \cdot \mathbb{E} X = \mathbb{E} X$$

Variance (WR):

$$\text{Var } \bar{X} = \frac{1}{n} \cdot \text{Var } X$$

Variance (WOR):

$$\text{Var } \bar{X} = \frac{1}{n} \cdot \text{Var } X \cdot \frac{N-n}{N-1}$$

## Sample total under SRS

The sample total is given by  $\sum_{i=1}^n x_i$ . Since the relation of population to sample size  $N/n$  has to be taken into account, a representative value of the total is given by the sample function  $\frac{N}{n} \cdot \sum_{i=1}^n = N \cdot \bar{X}$ .

Expected value (WR and WOR):

$$E N \cdot \bar{X} = E \left( N \cdot \frac{1}{n} \cdot (X_1 + \dots + X_n) \right) = \frac{N}{n} \cdot n \cdot E X = N \cdot E X$$

Variance (WR):

$$\text{Var } N \cdot \bar{X} = N^2 \cdot \frac{1}{n} \cdot \text{Var } X$$

Variance (WOR):

$$\text{Var } N \cdot \bar{X} = N^2 \cdot \frac{1}{n} \cdot \text{Var } X \cdot \frac{N-n}{N-1}$$

## Sample proportion under SRS

A dichotomous population with  $W(X = 1) = \theta$  and  $W(X = 0) = 1 - \theta$  as well as  $\text{E } X = \theta$  and  $\text{Var } X = \theta \cdot (1 - \theta)$  is considered.

The sample proportion is calculated as:  $p = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ .

Expected value (WR and WOR):

$$\text{E } P = \text{E} \left( \frac{1}{n} \cdot (X_1 + \dots + X_n) \right) = \theta$$

Variance (WR):

$$\text{Var } P = \frac{1}{n} \cdot \theta \cdot (1 - \theta)$$

Variance (WOR):

$$\text{Var } P = \frac{1}{n} \cdot \theta \cdot (1 - \theta) \cdot \frac{N - n}{N - 1}$$

## Sampling variance under SRS

The sampling variance is calculated as:  $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$

Expected value (WR):

$$\mathbb{E} S^2 = \text{Var } X$$

Expected value (WOR):

$$\mathbb{E} S^2 = \frac{N}{N-1} \cdot \text{Var } X$$

Variance (WR):

$$\text{Var } S^2 = \frac{2}{n-1} \cdot (\text{Var } X)^2$$

## General remarks on sample functions

- ▶ Aim: estimation of unknown parameters of the underlying population.  
e.g. the true, unknown mean of the population  $\mu$

### Estimation using $\hat{\mu}$ .

- ▶ Is the sample function  $\bar{x}$  applicable to this problem?
- ▶ Which properties does  $\bar{x}$  have?
- ▶ Is it possible to make statements concerning the distributions of the sample functions?  
Partially yes. See special functions of the normal distribution

In practice, only the results - and not the true values - are available!

# Elements of Statistics

## Chapter 8: Estimation

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

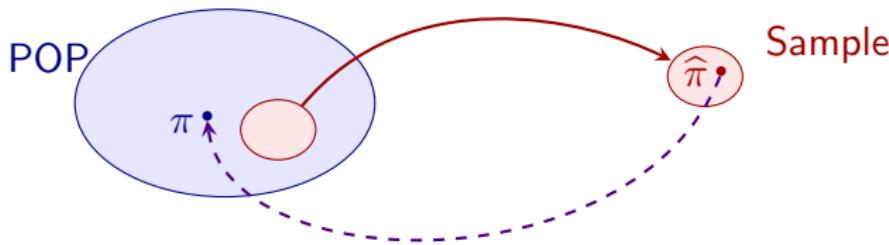
© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

## General idea

We are interested in population parameters which are generally unknown (here:  $\pi$ ).

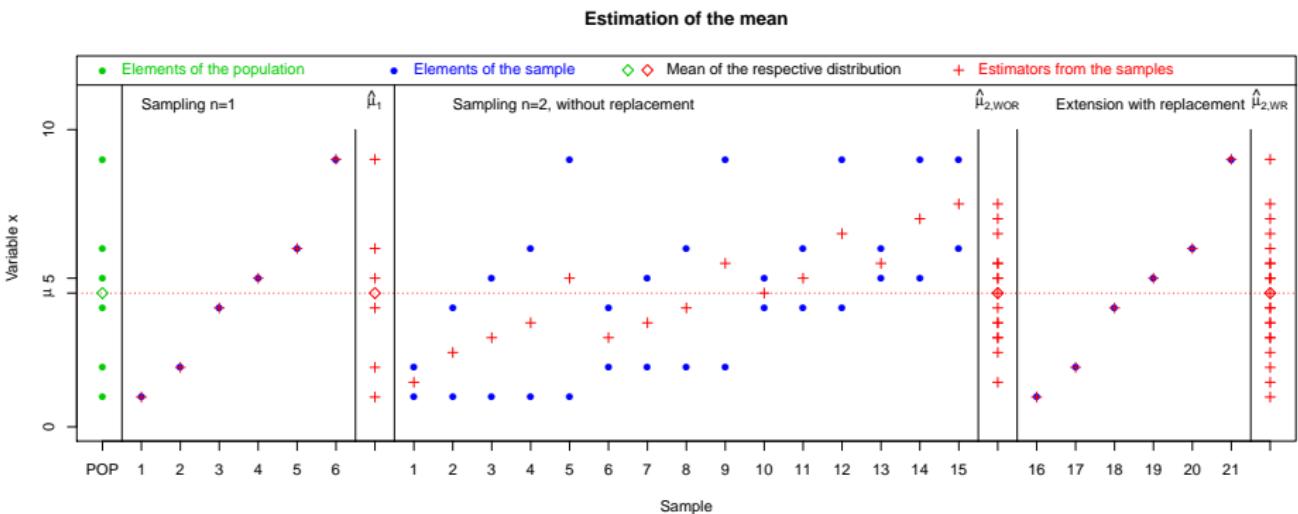
Before, we analysed populations using methods of descriptive statistics. Now, we draw a sample of the population and analyse this sample. The aim is to transfer results to the population ( $\rightarrow$  point estimation).



Additionally, we want to specify an interval of *plausible* values ( $\rightarrow$  interval estimation).

# Distribution of the sample mean

We draw all possible samples of size  $n = 1$  and  $n = 2$ , respectively, out of a population of  $N = 6$  elements. We have:



# Sample function and estimating function

## Specification of an estimating function

An estimating function for an unknown population parameter  $\pi$  is a sample function which qualifies to be used to estimate the parameter  $\pi$  by virtue of its properties. It is labelled  $u_\pi(x_1, \dots, x_n)$ . The realisation of the estimating function is the estimate  $\hat{\pi} = u_\pi(x_1, \dots, x_n)$ .

Attention: We distinguish between the parameter to be estimated  $\pi$ , the estimate  $\hat{\pi}$  and the distribution of the latter or the corresponding random variable. The latter results when we substitute the sample variables  $X_1, \dots, X_n$  for the corresponding realisations  $x_1, \dots, x_n$ . To be concrete, e. g. when estimating the mean of the population, we have  $\mu$  and  $\hat{\mu} = \bar{x}$ . We label the distribution of the estimator  $U_\pi(X_1, \dots, X_n)$  or  $U$ , and in this case  $\bar{X}$ . We may write  $U(X_1, \dots, X_n | \pi)$  as well.

## Example 8.1: Four estimating functions (1)

We want to estimate the mean  $\mu$  of the population. With a sample size of  $n$  we have four estimating functions at our disposal:

$$\hat{\mu}_1 = U_1(X_1, \dots, X_n | \pi) = \frac{1}{n} \sum_{i=1}^n X_i$$

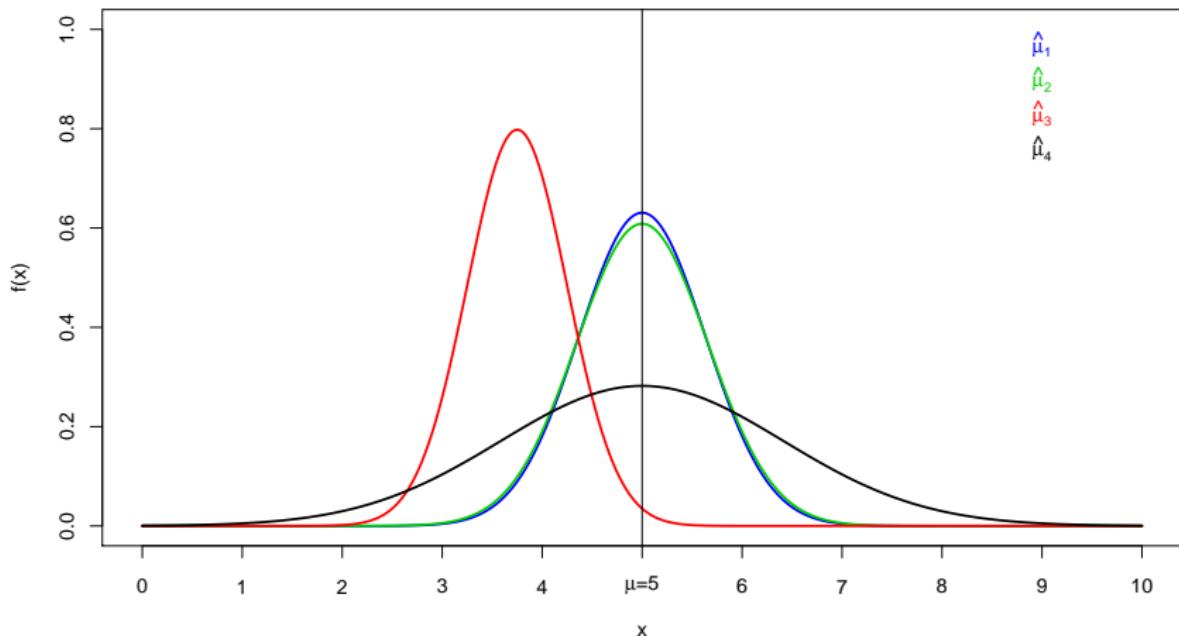
$$\hat{\mu}_2 = U_2(X_1, \dots, X_n | \pi) = \frac{1}{n+1} \cdot \left( 2 \cdot X_1 + \sum_{i=2}^n X_i \right)$$

$$\hat{\mu}_3 = U_3(X_1, \dots, X_n | \pi) = \frac{1}{n+6} \cdot \left( 2 \cdot X_1 + 2 \cdot X_n + \sum_{i=2}^{n-1} X_i \right)$$

$$\hat{\mu}_4 = U_4(X_1, \dots, X_n | \pi) = \frac{1}{2} \cdot (X_1 + X_n)$$

## Example 8.1: Four estimating functions (2)

Let the population be normally distributed with parameters  $\mu = 5$  and  $\sigma^2 = 4$ . We draw a sample of size  $n = 10$  with replacement.



## Example 8.1: Four estimating functions (3)

Calculation of  $\hat{\mu}_1$ ,  $\hat{\mu}_2$ ,  $\hat{\mu}_3$  and  $\hat{\mu}_4$  in R:

```
mu <- 5
sigma <- sqrt(4)
n <- 10

Mean_U1 <- 1/n * (n * mu)
Mean_U2 <- 1/(n + 1) * (2 * mu + 9 * mu)
Mean_U3 <- 1/(n + 6) * (2 * mu + 2 * mu + 8 * mu)
Mean_U4 <- 1/2 * (2 * mu)

Means<-cbind(Mean_U1 ,Mean_U2 ,Mean_U3 ,Mean_U4)

Means
```

	Mean_U1	Mean_U2	Mean_U3	Mean_U4
[1,]	5	5	3.75	5

## Example 8.1: Four estimating functions (4)

Calculation of  $\hat{\sigma}_1^2$ ,  $\hat{\sigma}_2^2$ ,  $\hat{\sigma}_3^2$  and  $\hat{\sigma}_4^2$  in R:

```
Var_U1 <- (1/n)^2 * (n * sigma^2)
Var_U2 <- (2/(n + 1))^2 * sigma^2 + (1/(n + 1))^2 *
            (9 * sigma^2)
Var_U3 <- (2/(n + 6))^2 * (2 * sigma^2) +
            (1/(n + 6))^2 * (8 * sigma^2)
Var_U4 <- (1/2)^2 * (2 * sigma^2)

Variances <- cbind(Var_U1, Var_U2, Var_U3, Var_U4)
```

Variances

	Var_U1	Var_U2	Var_U3	Var_U4
[1,]	0.4	0.4297521	0.25	2

## Example 8.1: Four estimating functions (5)

Creation of the graphics in R:

```
x8_1 <- seq(from = 0, to = 10, length.out = 1000)

f_x8_1_1 <- dnorm(x = x8_1, mean = Mean_U1,
                    sd = sqrt(Var_U1))
f_x8_1_2 <- dnorm(x = x8_1, mean = Mean_U2,
                    sd = sqrt(Var_U2))
f_x8_1_3 <- dnorm(x = x8_1, mean = Mean_U3,
                    sd = sqrt(Var_U3))
f_x8_1_4 <- dnorm(x = x8_1, mean = Mean_U4,
                    sd = sqrt(Var_U4))

plot(x = x8_1, y = f_x8_1_1, type = "l", xlab = "x",
      ylab = "f(x)", ylim = c(0,1), lwd = 2, col = "blue")
lines(x=x8_1,y=f_x8_1_2,type="l",lwd=2,col="green")
lines(x=x8_1,y=f_x8_1_3,type="l",lwd=2,col="red")
lines(x=x8_1,y=f_x8_1_4,type="l",lwd=2,col="black")
abline(v = 5)
```

# Properties of estimating functions (1)

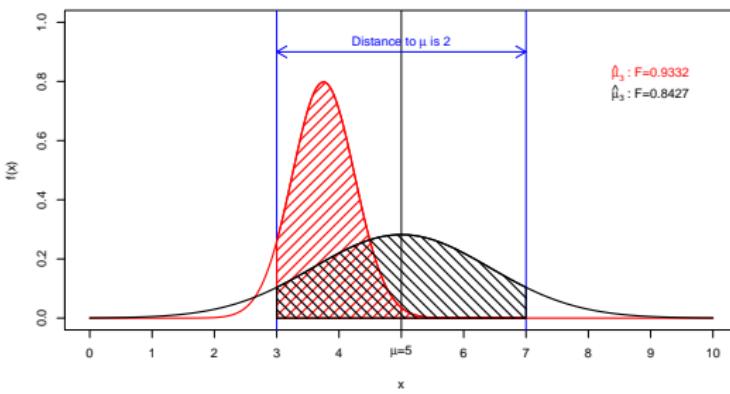
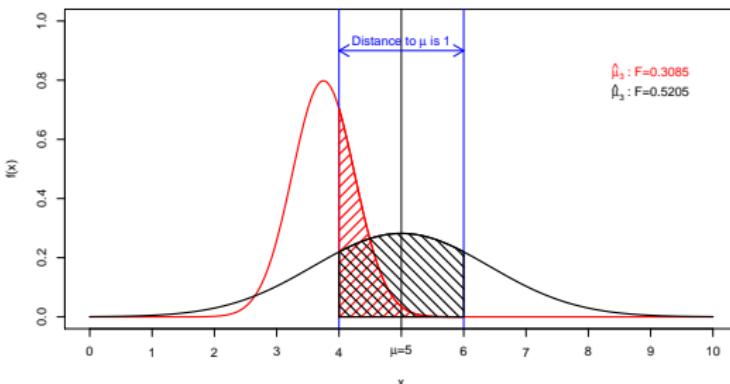
## Definition 8.1 (Estimation error):

An estimation error  $e$  is the actual error resulting from an estimation:

$$e = \hat{\mu} - \mu \quad .$$

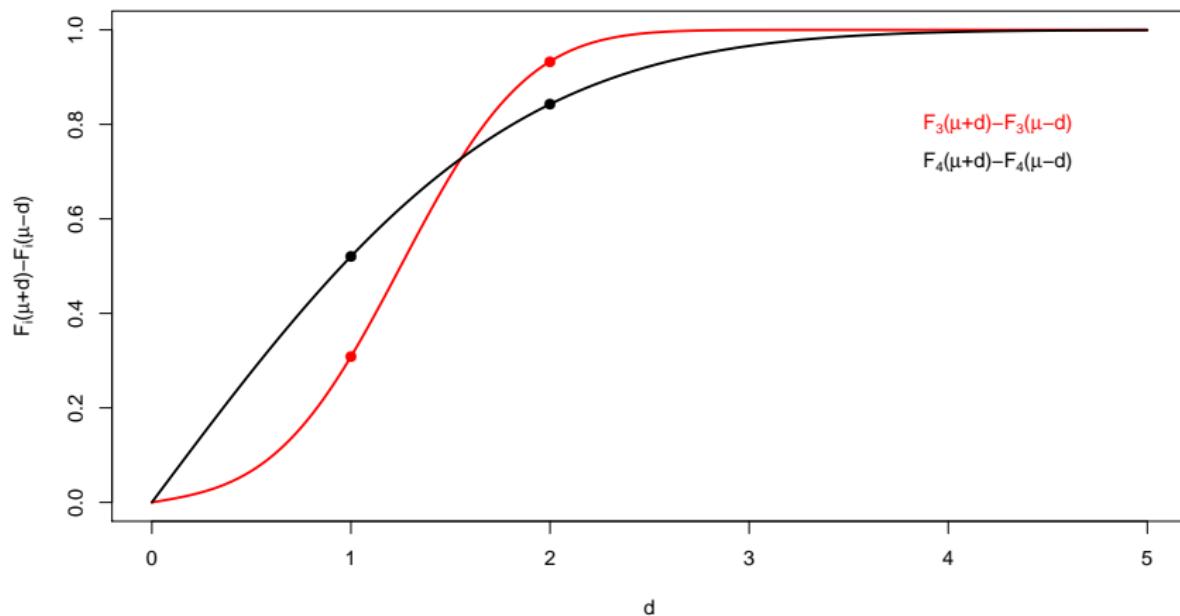
- ▶ Sampling is *random*. Therefore, the results of the different estimating functions will most likely lead to different evaluations for different samples.
- ▶ How can we compare estimating functions with regard to differing sample realisations?
- ▶ How should estimating functions behave for large samples ( $n \rightarrow \infty$ )?
- ▶ To what extent are such considerations useful in practice?

## Example 8.2: see Ex. 8.1 (1)



## Example 8.2: see Ex. 8.1 (2)

Probability for the interval  $[\mu - d; \mu + d]$  of the distributions of  $\hat{\mu}_3$  and  $\hat{\mu}_4$ :



## Properties of estimating functions (2)

### Definition 8.2 (Unbiasedness):

An estimating function  $U_\pi(X_1, \dots, X_n)$  (short hand:  $U$ ) is called unbiased for parameter  $\pi$  if we have

$$\mathbb{E}(U) = \pi.$$

The average estimate is equal to the parameter to be estimated  $\pi$ .

Otherwise it is called biased. The extent of the bias may be quantified as follows:

$$\text{Bias}(U) = \mathbb{E}(U) - \pi .$$

We speak of asymptotical unbiasedness, if the following holds:

$$\lim_{n \rightarrow \infty} \mathbb{E}(U_n) = \pi .$$

## Example 8.3: see Ex. 8.1 (1)

The estimating function  $U = \sum_i \gamma_i X_i$  with  $\sum_i \gamma_i = 1$  is unbiased because

$$\mathbb{E} U = \sum_{i=1}^n \gamma_i \cdot \underbrace{\mathbb{E} X_i}_{=\mu} = \mu \sum_{i=1}^n \gamma_i = \mu \quad .$$

Therefore,  $\hat{\mu}_1$ ,  $\hat{\mu}_2$  and  $\hat{\mu}_4$  are unbiased as their weights are  $\gamma_i = 1/n$  for  $\hat{\mu}_1$ ,  $\gamma_1 = 2/(n+1)$  and  $\gamma_i = 1/(n+1)$  ( $i > 1$ ) for  $\hat{\mu}_2$  as well as  $\gamma_1 = \gamma_n = 1/2$  and  $\gamma_i = 0$  ( $i \neq 1, n$ ) for  $\hat{\mu}_4$ .

For  $\hat{\mu}_3$  follows:

$$\begin{aligned}\mathbb{E} U_3 &= \mathbb{E} \left( \frac{1}{n+6} \cdot \left( 2 \cdot X_1 + 2 \cdot X_n + \sum_{i=2}^{n-1} X_i \right) \right) \\ &= \frac{1}{n+6} \cdot \left( 2 \cdot \mathbb{E} X_1 + 2 \cdot \mathbb{E} X_n + \sum_{i=2}^{n-1} \mathbb{E} X_i \right) = \frac{n+2}{n+6} \cdot \mu \quad .\end{aligned}$$

## Example 8.3: see Ex. 8.1 (2)

Calculations for  $\hat{\mu}_3$  in R:

```
Mean_U3 <- (n + 2)/(n + 6) * mu
Mean_U3
```

```
[1] 3.75
```

$U_3$  is biased but asymptotically unbiased as  $\lim_{n \rightarrow \infty} \frac{n+2}{n+6} \cdot \mu = \mu$ .

Calculation of the bias of  $\hat{\mu}_3$  in R:

```
Bias_U3 <- Mean_U3 - mu
Bias_U3
```

```
[1] -1.25
```

Calculation of the bias with  $n = 10,000$  in R:

```
n_new <- 10000
Bias_U3_new <- (n_new + 2)/(n_new + 6) * mu - mu
round(Bias_U3_new, digits = 4)
```

```
[1] -0.002
```

### Example 8.4:

The estimating function  $p = \hat{\theta}$  is unbiased for the proportion  $\theta$  of a certain type of interest in the population. This follows immediately from an application of the arithmetic mean in Example 8.3 on dichotomous variables.

### Example 8.5:

The sample variance

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$$

is unbiased for the population variance  $\sigma^2$ . Therefore,

$$S^{*2} = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$$

must be biased. Nevertheless,  $S^{*2}$  is asymptotically unbiased.

## Example 8.6: Two estimating functions

For the estimation of population parameter  $\pi$  we have two different unbiased estimating functions  $U_1$  and  $U_2$  at our disposal. We only know that  $\text{Var } U_1 = 0.9 \cdot \text{Var } U_2$ . Using Tchebysheff's inequality (theorem 7.2) we have:

$$P(|U_1 - \pi| \geq \varepsilon) \leq \frac{\text{Var } U_1}{\varepsilon^2} = 0.9 \cdot \frac{\text{Var } U_2}{\varepsilon^2}$$

$$P(|U_2 - \pi| \geq \varepsilon) \leq \frac{\text{Var } U_2}{\varepsilon^2} .$$

The probability of *committing* an estimation error of at least  $\varepsilon$  is smaller for  $U_1$  and depends on the variance of the estimating functions.

In case of biased estimating functions we may use the extended version of Tchebysheff's inequality (see Schaich and Münnich, 2001, p. 21):

$$P(|U_2 - \pi| \geq \varepsilon) \leq \frac{\text{Var } U_2 + (\mathbb{E } U_2 - \pi)^2}{\varepsilon^2} = \underbrace{\frac{\text{Var } U_2 + \text{Bias}^2(U_2)}{\varepsilon^2}}_{:=\text{MSE}(U_2)} .$$

## Properties of estimating functions (3)

### Definition 8.3 (Efficiency):

An unbiased estimating function  $U$  is called efficient (best) estimating function for parameter  $\pi$  if there is no other unbiased estimating function  $U'$  for  $\pi$  with  $\text{Var}(U') \leq \text{Var}(U)$ .

Out of a number of unbiased estimating functions we choose the one with the smallest variance.

In practice, it's far from easy to find the best estimating function. With the aid of sufficient estimating functions (estimating functions that use all information of a sample about the parameter that one wants to estimate) and the Rao-Blackwell theorem, one can construct *better* estimating functions (see lecture *Elements of Statistics and Econometrics* in the masters program *M.Sc. Applied Statistics*).

## Example 8.7: Arithmetic mean

Out of the linear unbiased estimating functions, the arithmetic mean is the best estimating function for  $\mu$ . Using the Lagrange multiplier we get:

$$\begin{aligned} \frac{\partial \left[ \text{Var} \left( \sum_{i=1}^n \gamma_i X_i \right) + \lambda \left( 1 - \sum_{i=1}^n \gamma_i \right) \right]}{\partial \gamma_i} &= \\ \frac{\partial}{\partial \gamma_i} \left[ \sum_{i=1}^n \gamma_i^2 \text{Var} X_i + \lambda \left( 1 - \sum_{i=1}^n \gamma_i \right) \right] &= \\ \frac{\partial}{\partial \gamma_i} \left[ \sigma^2 \cdot \sum_{i=1}^n \gamma_i^2 + \lambda \left( 1 - \sum_{i=1}^n \gamma_i \right) \right] &= \\ \sigma^2 \cdot 2 \cdot \gamma_i - \lambda &\stackrel{!}{=} 0 \quad . \end{aligned}$$

Finally, after equating we get  $\gamma_i = \gamma_j$  for all  $i, j = 1, \dots, n$  and therefore the proposition.

We say that the arithmetic mean is the *best linear unbiased estimator* (BLUE) for  $\mu$ .

## Properties of estimating functions (4)

### Definition 8.4 (Consistency):

An estimating function  $U(X_1, \dots, X_n | \pi)$  is called consistent for the estimation of the population parameter  $\pi$  if

$$\lim_{n \rightarrow \infty} P(|U_n - \pi| > \varepsilon) = 0$$

for any arbitrarily small  $\varepsilon > 0$ .

We say that  $U_n$  converges stochastically to the parameter to be estimated  $\pi$ .

## Example 8.8: Consistency of $\bar{X}$

$\bar{X}_n$  is the arithmetic mean for sample size  $n$  (with replacement). Using Tchebysheff's inequality and  $\text{Var } \bar{X}_n = \text{Var } X/n$  ( $\bar{X}_n$  is unbiased) we get

$$\begin{aligned} P(|\bar{X}_n - E \bar{X}_n| > \varepsilon) &\leq P(|\bar{X}_n - \mu| \geq \varepsilon) \\ &\leq \frac{\text{Var } \bar{X}_n}{\varepsilon^2} = \frac{\text{Var } X}{n \cdot \varepsilon^2} \end{aligned}$$

for every  $\varepsilon > 0$ . Finally, we then have

$$0 \leq \lim_{n \rightarrow \infty} P(|\bar{X}_n - E \bar{X}_n| > \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\text{Var } X}{n \cdot \varepsilon^2} = 0 \quad .$$

$\bar{X}_n$  is consistent.

# Methods to gain estimating functions

## ► Ordinary least squares (OLS):

The sum of the squared errors is minimised. Examples are the OLS regression (see Chapter 4) or  $\hat{\mu}_{KQ}$ :

$$\sum_i (x_i - \hat{\mu}_{KQ})^2 \rightarrow \min \text{ leads to } \hat{\mu}_{KQ} = \bar{x}.$$

## ► Method of moments:

The empirical moments  $\frac{1}{n} \sum_i x_i^k$  are made equal to the theoretical moments  $E(X^k)$ . From this, one obtains the estimates. With  $\hat{\mu} = \bar{x}$  and  $\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n} \sum_i x_i^2$  for  $k = 1, 2$ , one would finally  $\hat{\sigma}^2 = s^*{}^2$  with unknown  $\mu$ .

## ► Maximum Likelihood method (ML)

## ► Bayesian estimation

# Maximum Likelihood method

Given the  $n$  stochastically independent realisations of a random sample, the explicit parameters of a known distribution have to be estimated. From the set of all possible estimates, those estimates are selected which have the highest probability or probability density given the available sample result. Hence:

$$\begin{aligned} L(x_1, \dots, x_n | \hat{\pi}_1, \dots, \hat{\pi}_r) &= \max_{\pi_1, \dots, \pi_r} L(x_1, \dots, x_n | \pi_1, \dots, \pi_r) \\ &= \max_{\pi_1, \dots, \pi_r} \prod_{i=1}^n f(x_i | \pi_1, \dots, \pi_r) . \end{aligned}$$

In most cases, the log likelihood function  $\ln L$  is maximized instead of the likelihood function  $L$ , whereby a sum instead of a product is maximized.

# Properties of the Maximum Likelihood method

- ▶ Given there is an efficient estimate for a parameter  $\pi$ , the ML method yields it
- ▶ ML estimation functions are consistent, but generally not unbiased
- ▶ ML estimators are asymptotically normal distributed for  $n \rightarrow \infty$
- ▶ If  $U$  is an ML estimation function for  $\pi$ , then  $\tau(U)$  is also an ML estimation function for a wide class of functions  $\tau$

## Example 8.9: One urn, two colours (1)

An urn contains  $N = 50$  balls. Those balls are either black or yellow but the respective proportions  $\theta$  are unknown. A sample of size  $n = 10$  (WR) yields four black balls. We are looking for the  $\theta$  which maximizes  $b(4|10; \theta)$ . Because of  $N = 50$ ,  $\theta$  can only be a multiple of 0.02. Resulting in:

$\theta$	0.34	0.36	0.38	0.40	0.42	0.44	0.46
$b(4 10; \theta)$	0.2320	0.2424	0.2487	0.2508	0.2488	0.2427	0.2331

Creation of the table in R:

```
x8_9 <- 4
n <- 10
theta <- seq(from = 0.34, to = 0.46, by = 0.02)
theta
[1] 0.34 0.36 0.38 0.40 0.42 0.44 0.46
f_x8_9 <- dbinom(x = x8_9, size = n, prob = theta)
round(f_x8_9, digits = 4)
[1] 0.2320 0.2424 0.2487 0.2508 0.2488 0.2427 0.2331
```

## Example 8.9: One urn, two colours (2)

Thus,  $\hat{\theta} = 0.4$  is used as the ML estimate in this case.

Determination of  $\hat{\theta} = 0.4$  in R:

```
theta_hat <- theta[which.max(f_x8_9)]  
theta_hat
```

```
[1] 0.4
```

## Example 8.10:

### ML estimation of $\theta$ - Binomial distribution (1)

In a sample of size  $n$ , the outcome 1 results  $n \cdot p$  times whereas the outcome 0 results  $n \cdot (1 - p)$  times. Thus, the likelihood function is given by:

$$L(x_1, \dots, x_n | \theta) = \prod_{i=1}^n \theta^{x_i} \cdot (1 - \theta)^{1-x_i} = \theta^{np} \cdot (1 - \theta)^{n(1-p)} .$$

Taking the logarithm results in:

$$\ln L(x_1, \dots, x_n | \theta) = np \ln \theta + n(1 - p) \ln(1 - \theta)$$

finally, differentiation yields

$$\frac{\partial \ln L(x_1, \dots, x_n | \theta)}{\partial \theta} = \frac{np}{\theta} - \frac{n(1 - p)}{1 - \theta} \stackrel{!}{=} 0 .$$

Thus the necessary criterion for a maximum finally results in  $\hat{\theta} = p$ .  
 Sufficient criterion still has to be checked!

## Example 8.10: ML estimation of $\theta$ - Binomial distribution (2)

Log likelihood and partial derivative in R:

```
Log_Likelihood <- expression(n * p * log(Theta) +  
                           n * (1 - p) * log(1 - Theta))  
  
D_Log_Likelihood <- D(expr = Log_Likelihood, name = "Theta")  
  
Log_Likelihood  
  
expression(n * p * log(Theta) + n * (1 - p) * log(1 - Theta))  
  
D_Log_Likelihood  
  
n * p * (1/Theta) - n * (1 - p) * (1/(1 - Theta))
```

## Example 8.11:

### ML estimation of $\mu$ and $\sigma^2$ of a normal distribution I

The following is valid:

$$\begin{aligned} L(x_1, \dots, x_n | \mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

or:

$$\ln L(x_1, \dots, x_n | \mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 .$$

## Example 8.11:

### ML estimation of $\mu$ and $\sigma^2$ of a normal distribution II

Finally, partial derivation with respect to the parameters  $\mu$  and  $\sigma^2$

$$\frac{\ln L(x_1, \dots, x_n | \mu; \sigma^2)}{\partial \mu} = \frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^n (x_i - \mu) \stackrel{!}{=} 0 \quad \text{and}$$

$$\frac{\ln L(x_1, \dots, x_n | \mu; \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \stackrel{!}{=} 0$$

yields the estimators  $\hat{\mu} = \bar{x}$  and  $\hat{\sigma}^2 = s^{*2}$ .

# Bayesian estimation (see Fahrmeir et al., 2016)

- ▶  $x_1, \dots, x_n$  are  $n$  independent realisations of a random variable  $X$  which follows a distribution  $F$  with parameter  $\theta$
- ▶  $\theta$  is a realisation of a random variable  $\Theta$
- ▶  $f(x, \theta)$  is the joint density;  $f(x|\theta)$  is the conditional and  $f(x)$  the boundary distribution of  $X$
- ▶  $f(\theta)$  is the a-priori distribution of the parameter  $\Theta$
- ▶  $f(\theta|x)$  is the a-posteriori distribution of  $\Theta$

## Bayesian inference

Let  $f(x|\theta)$  be the density of  $X$  given  $\theta$  and  $L(\theta) = f(x_1, \dots, x_n|\theta)$  constitutes the corresponding likelihood function. Then, the a-priori density  $f(\theta)$  can be used to derive the a-posteriori density of  $\theta$

$$f(\theta|x_1, \dots, x_n) = \frac{f(x_1|\theta) \dots f(x_n|\theta) \cdot f(\theta)}{\int f(x_1|\theta) \dots f(x_n|\theta) \cdot f(\theta) d\theta} = \frac{L(\theta)f(\theta)}{\int L(\theta)f(\theta) d\theta}$$

(discrete distributions and multidimensional  $\Theta$  are also possible).

# Bayesian estimator und Bayesian learning

## A-posteriori expected value

$$\hat{\theta}_E = E(\theta|x_1, \dots, x_n) = \int \theta f(\theta|x_1, \dots, x_n) d\theta$$

## Maximum a-posteriori estimator (MAP)

$$\hat{\theta}_{MAP} = \arg \max_{\theta} L(\theta)f(\theta) \quad \text{or} \quad \hat{\theta}_{MAP} = \arg \max_{\theta} (\ln L(\theta) + \ln f(\theta))$$

The calculation of the a-posteriori density of  $\theta$  is often no longer analytically feasible → numerical or Monte Carlo integration or MCMC.

If the a-priori distribution of  $\Theta$  is *very flat* (non-informative prior), then one obtains the Maximum Likelihood estimation. Otherwise, the subjective conceptions of the a-priori distribution is used in the estimation.

## Example 8.12: see Fahrmeir et al., 2016

Let  $x_1, \dots, x_n$  be independent realisations from  $X \sim N(\mu, \sigma^2)$  with known  $\sigma^2$ . We want to estimate the parameter  $\mu$ . We use  $N(\mu_0, \sigma_0^2)$  as a-priori density for the parameter we want to estimate.  $\sigma_0^2$  controls the precision of the a-priori information.

With some effort, we can show that the a-posteriori distribution of  $\mu$  is

$$\mu | x_1, \dots, x_n \sim N\left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0; \frac{\sigma^2}{n + \sigma^2/\sigma_0^2}\right)$$

The *trust parameter*  $\sigma_0^2$  controls the evaluation of the sample information. If  $\sigma_0^2$  is very large ( $\rightarrow \infty$ ), then we obtain the classical MLE. If on the other hand  $\sigma_0^2$  is very small, then the a-priori information changes little with the sample information.

## General idea of interval estimation

Besides the point estimate derived from the sample, we need some kind of *quality criterion* for this point estimate. Some options:

- ▶ Variance of the estimator  
(requires an approximate normal distribution)
- ▶ Standard error (standard deviation of estimator)
- ▶ Coefficient of variation of estimate

The problem of each of those options is that missing information regarding the population forces us to *estimate* their respective values using the sample.

Another option is to state a certain *range of variation* around the point estimate. We would like to state an interval based on quantiles of an estimator's distribution, like  $[x_{0.025}; x_{0.975}]$ .

## Example 8.13: Random interval (1)

Let the random variable  $X$  be normally distributed with known variance  $\sigma^2 = 900$ . To estimate the population mean  $\mu$  we draw a sample of size  $n = 36$  with replacement.

We use the estimator  $\bar{X}$ . We use

- ▶  $\bar{X}_l = \bar{X} + z(0.025) \cdot \frac{\sigma}{\sqrt{n}} = \bar{X} - 1.96 \cdot 5 = \bar{X} - 9.8$
- ▶  $\bar{X}_u = \bar{X} + z(0.975) \cdot \frac{\sigma}{\sqrt{n}} = \bar{X} + 1.96 \cdot 5 = \bar{X} + 9.8$

as the limits of the interval motivated above. We get the following random interval  $[\bar{X}_l, \bar{X}_u] = [\bar{X} - 9.8; \bar{X} + 9.8]$ .

What is the probability that the parameter to be estimated  $\mu$  lies within the limits of this random interval?

## Example 8.13: Random interval (2)

$$\begin{aligned}P(\bar{X}_l \leq \mu \leq \bar{X}_u) &= P(\bar{X} - 9.8 \leq \mu \leq \bar{X} + 9.8) \\&= P(-9.8 \leq \bar{X} - \mu \leq 9.8) \\&= P\left(-1.96 \leq \frac{\bar{X} - \mu}{\frac{5}{\sqrt{n}}} \leq 1.96\right) \\&\quad \text{~$\sim$SND} \\&= 0.975 - 0.025 = 0.95\end{aligned}$$

A random interval constructed in this fashion *covers* the true parameter  $\mu$  with a probability of 95%.

Data input of relevant parameters in R:

```
alpha <- 0.05
sigma <- 30
n <- 36
```

## Example 8.13: Random interval (3)

Attention:

Such a statement may only be given in terms of probabilities and therefore only **before** an experiment is carried out. As soon as a concrete interval  $[\bar{x}_l, \bar{x}_u]$  is determined, we can only state if the true parameter is covered by the interval or not. But in reality this information will not be available in most cases.

# Confidence intervals (1)

- ▶ As we assume that the probability of the interval  $[\bar{X}_l, \bar{X}_u]$  covering the true parameter  $\mu$  is 0.95 before the experiment is carried out,
  - ▶ we have a respective level of *confidence* that
  - ▶ the true parameter actually lies within the limits of the confidence interval after the experiment has been carried out.
- ▶ Therefore, the interval  $[\bar{X}_l, \bar{X}_u]$  is called 95% confidence interval for  $\mu$ .
- ▶ Generally, depending on the question at hand, we use values of 0.95, 0.99 or 0.90.

## Confidence intervals (2)

### Definition 8.5 (Confidence interval):

Let the confidence level  $(1 - \alpha)$  be given. The interval  $[\pi_l, \pi_u]$  with  $\pi_l = f(X_1, \dots, X_n)$  and  $\pi_u = f(X_1, \dots, X_n)$  ( $\pi_l \leq \pi_u$ ) is called  $(1 - \alpha)$  confidence interval for  $\pi$ , if we have  $P(\pi_l \leq \pi \leq \pi_u) = 1 - \alpha$ .

Questions about the properties of such a confidence interval, like its symmetry or its minimal length, immediately arise.

## CI for $\mu$ , POP is normally distributed, $\sigma^2$ is known

The random variable

$$Z = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$$

follows the standard normal distribution. The resulting  $(1 - \alpha)$  confidence interval is

$$\left[ \bar{X} - z(1 - \alpha/2) \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + z(1 - \alpha/2) \cdot \frac{\sigma}{\sqrt{n}} \right].$$

- ▶ This  $(1 - \alpha)$  confidence interval is as short as possible and is symmetric to  $\bar{X}$ .
- ▶ The larger  $\sigma$ , the longer the CI
- ▶ The larger  $n$ , the shorter the CI
- ▶ The larger  $(1 - \alpha)$ , the longer the CI

## Example 8.14: see Ex. 8.13

The evaluation of the sample yielded  $\bar{x} = 72$ . Therefore, the 95% confidence interval is

$$\left[ 72 - 1.96 \cdot \frac{30}{\sqrt{36}}, 72 + 1.96 \cdot \frac{30}{\sqrt{36}} \right]$$

and finally

$$[62.2; 81.8] .$$

95% confidence interval in R:

```
SpMean <- 72
CI <- vector()
CI[1] <- SpMean - qnorm(p = 1 - (alpha/2))*(sigma/sqrt(n))
CI[2] <- SpMean + qnorm(p = 1 - (alpha/2))*(sigma/sqrt(n))

CI_lower_alternative <- SpMean + qnorm(p = alpha/2) *
                           (sigma/sqrt(n))
round(CI, digits = 1)
[1] 62.2 81.8
```

# CI for $\mu$ , POP is normally distributed, $\sigma^2$ is unknown

The random variable

$$T = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

follows the  $t$  distribution with  $n - 1$  degrees of freedom. The resulting  $(1 - \alpha)$  confidence interval is

$$\left[ \bar{X} - t(1 - \frac{\alpha}{2}, n - 1) \cdot \sqrt{\frac{S^2}{n}}; \bar{X} + t(1 - \frac{\alpha}{2}, n - 1) \cdot \sqrt{\frac{S^2}{n}} \right].$$

- ▶ We have  $\frac{n-1}{\sigma^2} \cdot S^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$ .
- ▶ Cochran's theorem holds and therefore  $\frac{1}{\sigma^2} S^2(n-1)$  and  $\frac{1}{\sigma} \cdot (\bar{X} - \mu) \cdot \sqrt{n}$  are stochastically independent.
- ▶  $\frac{1}{\sigma} \cdot (\bar{X} - \mu) \cdot \sqrt{n} / \sqrt{\left( \frac{1}{\sigma^2} S^2(n-1) \right) / (n-1)} = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$

## Example 8.15: Unknown variance (1)

Now, let  $\sigma^2$  be unknown. As an estimate of  $\sigma^2$  we use  $s^2 = 33^2$  which is derived from the sample. We get the 95% confidence interval

$$\left[ 72 - 2.0315 \cdot \frac{33}{\sqrt{36}}; 72 + 2.0315 \cdot \frac{33}{\sqrt{36}} \right]$$

and finally

$$\left[ 72 - 11.173; 72 + 11.173 \right] = [60.8268; 83.1733] .$$

Attention:

$t(0.975; 35)$  is not tabulated. We used the arithmetic mean of the tabulated values  $t(0.975; 30)$  and  $t(0.975; 40)$  as the normal approximation would still yield inexact values (small  $n$ ).

Thanks to R, this is not a problem anymore (see next slide).

## Example 8.15: Unknown variance (2)

95% Confidence interval in R:

```
alpha <- 0.05
SpMean <- 72
SpVar <- 33^2
n <- 36

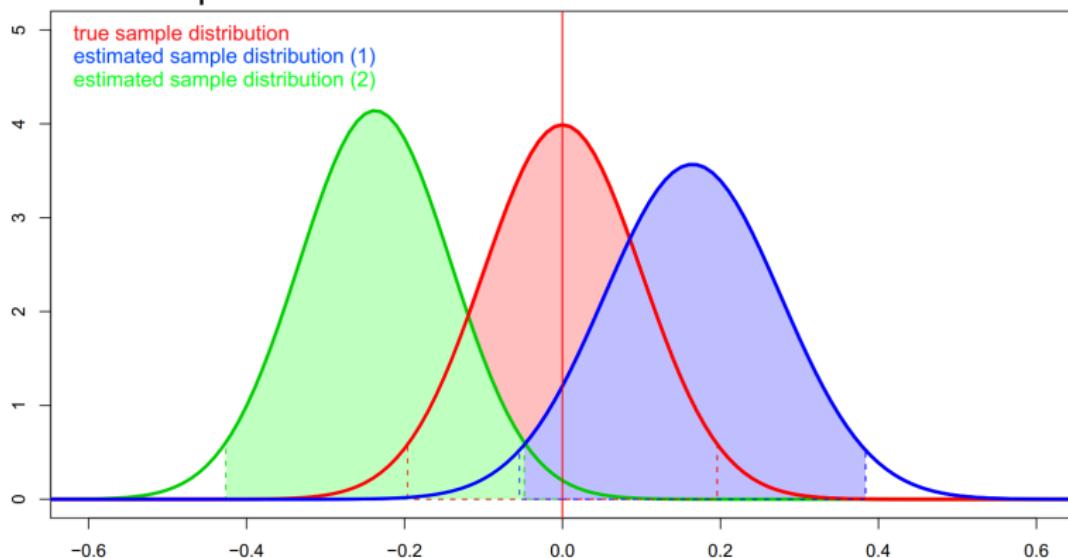
CI <- vector()
CI[1] <- SpMean - qt(p = 1 - (alpha/2), df = n - 1) *
            sqrt(SpVar/n)
CI[2] <- SpMean + qt(p = 1 - (alpha/2), df = n - 1) *
            sqrt(SpVar/n)

round(CI, digits = 1)
```

```
[1] 60.8 83.2
```

## Example 8.16: Sample distributions (1)

Let the population be normally distributed with unknown variance  $\sigma^2$ . A sample of size  $n = 10$  is drawn. We can compare the true but unknown sample distribution as well as two estimated distributions resulting from two different samples.



## Example 8.16: Sample distributions (2)

Confidence interval simulations (see next slide)

Point vs. variance estimates (upper left)

→ Cochran's theorem (given normal distribution)

Estimated distributions (lower left)

True distribution and estimated distributions of  $\bar{X}$

Confidence intervals (upper right)

For  $R = 100$  simulation runs;

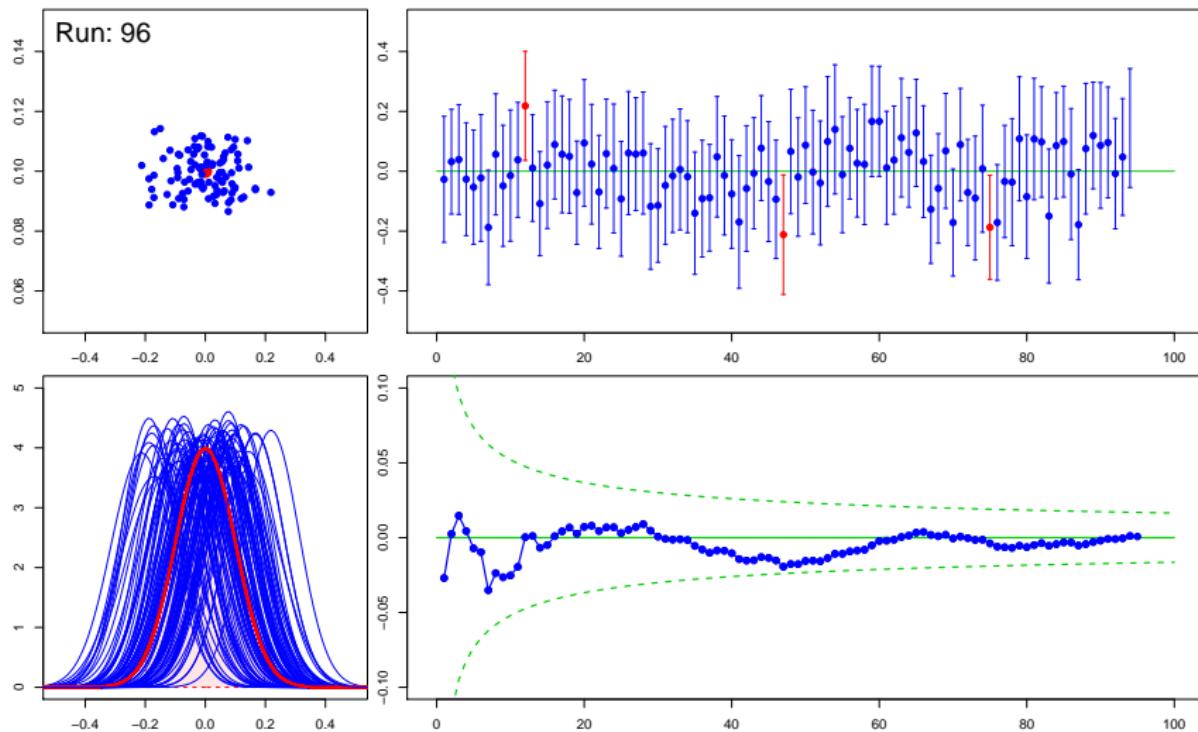
red intervals do not cover true value

Convergence of  $E(\bar{X})$  (lower right)

For  $r$ -th simulation run;

check of law of large numbers

# Simulation of standard normal distribution



# CI for $\sigma^2$ , POP is normally distributed

The random variable

$$\frac{n-1}{\sigma^2} \cdot S^2$$

follows a  $\chi^2$  distribution with  $n - 1$  degrees of freedom. We get the  $(1 - \alpha)$  confidence interval

$$\left[ \frac{(n-1)S^2}{\chi^2(1 - \frac{\alpha}{2}; n-1)}, \frac{(n-1)S^2}{\chi^2(\frac{\alpha}{2}; n-1)} \right]$$

- ▶ The CI follows from a rearrangement of
$$\chi^2(\frac{\alpha}{2}; n-1) \leq \frac{n-1}{\sigma^2} \cdot S^2 \leq \chi^2(1 - \frac{\alpha}{2}; n-1).$$
- ▶ The CI does not have a minimal length. For very large  $n$  the normal approximation ensures the property of symmetry and minimal length.

## Example 8.17: CI for variance (1)

Let a population be normally distributed. A sample of size  $n = 25$  yields  $s^2 = 7.244$ . We search the 90% confidence interval for  $\sigma^2$ .

We have  $\chi^2(0.05; 24) = 13.848$  and  $\chi^2(0.95; 24) = 36.415$ . Therefore, we get the 90% confidence interval

$$\left[ \frac{24 \cdot 7.244}{36.415}; \frac{24 \cdot 7.244}{13.848} \right]$$

and finally

$$\left[ 4.774; 12.555 \right].$$

## Example 8.17: CI for variance (2)

90% confidence interval in R:

```
alpha <- 0.1
n <- 25
SpVar <- 7.244

CI <- vector()

CI [1] <- ((n - 1) * SpVar) /
    qchisq(p = 1 - alpha/2, df = n-1)

CI [2] <- ((n - 1) * SpVar) /
    qchisq(p = alpha/2, df = n-1)

round(CI, digits = 3)

[1] 4.774 12.554
```

# CI for $E X$ , arbitrary distribution, $\text{Var } X$ known

The random variable

$$Z = \frac{\bar{X} - E X}{\sqrt{\text{Var } X}} \cdot \sqrt{n}$$

does approximatively follow a standard normal distribution. For  $n > 30$ , following the central limit theorem of Lindeberg and Lévy, the  $(1 - \alpha)$  confidence interval is

$$\left[ \bar{X} - z(1 - \alpha/2) \cdot \sqrt{\frac{\text{Var } X}{n}}; \bar{X} + z(1 - \alpha/2) \cdot \sqrt{\frac{\text{Var } X}{n}} \right].$$

## Example 8.18: see Ex. 8.13

Now, let a normal distribution of the population be questionable. As the sample size is  $n = 36$ , we again have the 95% confidence interval  $[62.2; 81.8]$ , but now it is not exact but approximative.

95% confidence interval in R:

```
n <- 36
alpha <- 0.05
VarX <- 30^2

CI_new <- vector()
CI_new[1] <- SpMean - qnorm(p=1-alpha/2) * sqrt(VarX/n)
CI_new[2] <- SpMean + qnorm(p=1-alpha/2) * sqrt(VarX/n)

round(CI_new, digits = 1)
```

```
[1] 62.2 81.8
```

# CI for $\text{E } X$ , arbitrary distribution, $\text{Var } X$ unknown

The random variable

$$Z = \frac{\bar{X} - \text{E } X}{\sqrt{S^2}} \cdot \sqrt{n}$$

does approximately follow a standard normal distribution. For  $n > 30$ , following the central limit theorem of Lindeberg and Lévy, the  $(1 - \alpha)$  confidence interval is

$$\left[ \bar{X} - z(1 - \alpha/2) \cdot \sqrt{\frac{S^2}{n}}; \bar{X} + z(1 - \alpha/2) \cdot \sqrt{\frac{S^2}{n}} \right].$$

## Example 8.19: see Ex. 8.15 (1)

Analogously to Example 8.18 as an approximative 95% confidence interval we have

$$\left[ 72 - 1.96 \cdot \frac{33}{\sqrt{36}}; 72 + 1.96 \cdot \frac{33}{\sqrt{36}} \right]$$

and therefore

$$\left[ 72 - 10.78; 72 + 10.78 \right] = [61.22; 82.78] .$$

Check of approximation conditions in R:

```
SpVar <- 33^2
n > 30
[1] TRUE
```

95% confidence interval in R:

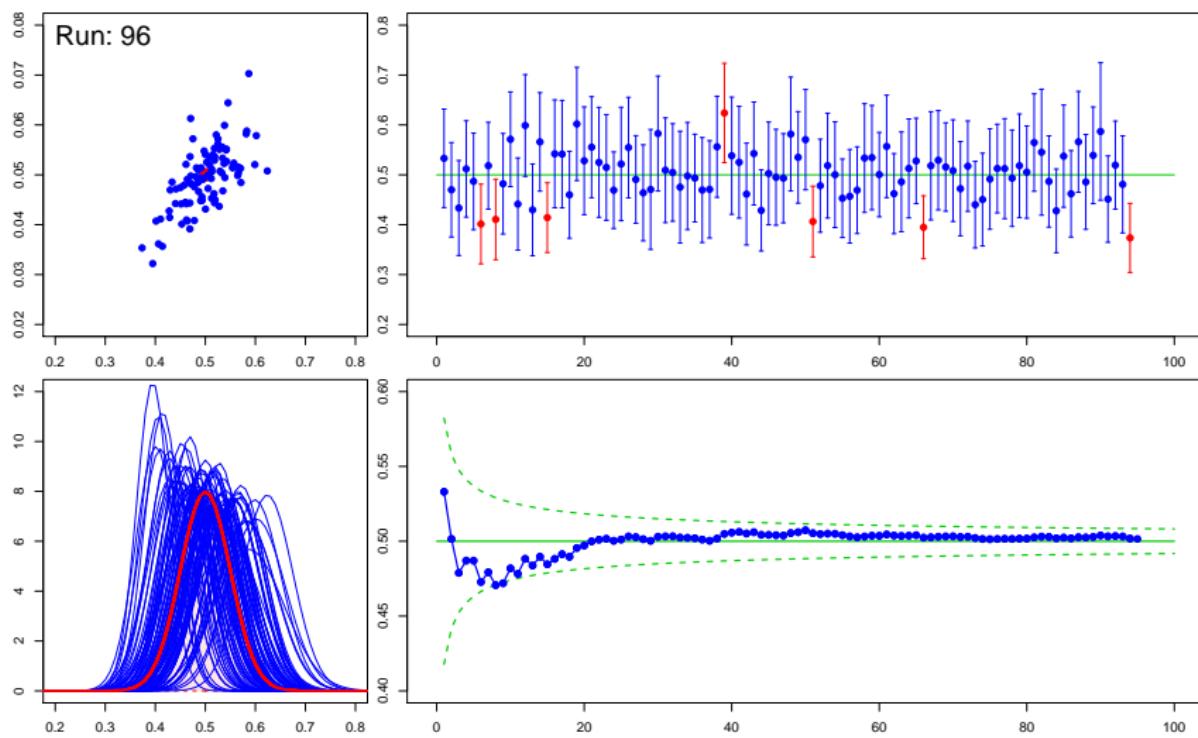
```
CI_new <- vector()
CI_new[1] <- SpMean - qnorm(p = 1 - alpha/2) *
  sqrt(SpVar/n)
CI_new[2] <- SpMean + qnorm(p = 1 - alpha/2) *
  sqrt(SpVar/n)
round(CI_new, digits = 2)
[1] 61.22 82.78
```

## Example 8.19: see Ex. 8.15 (2)

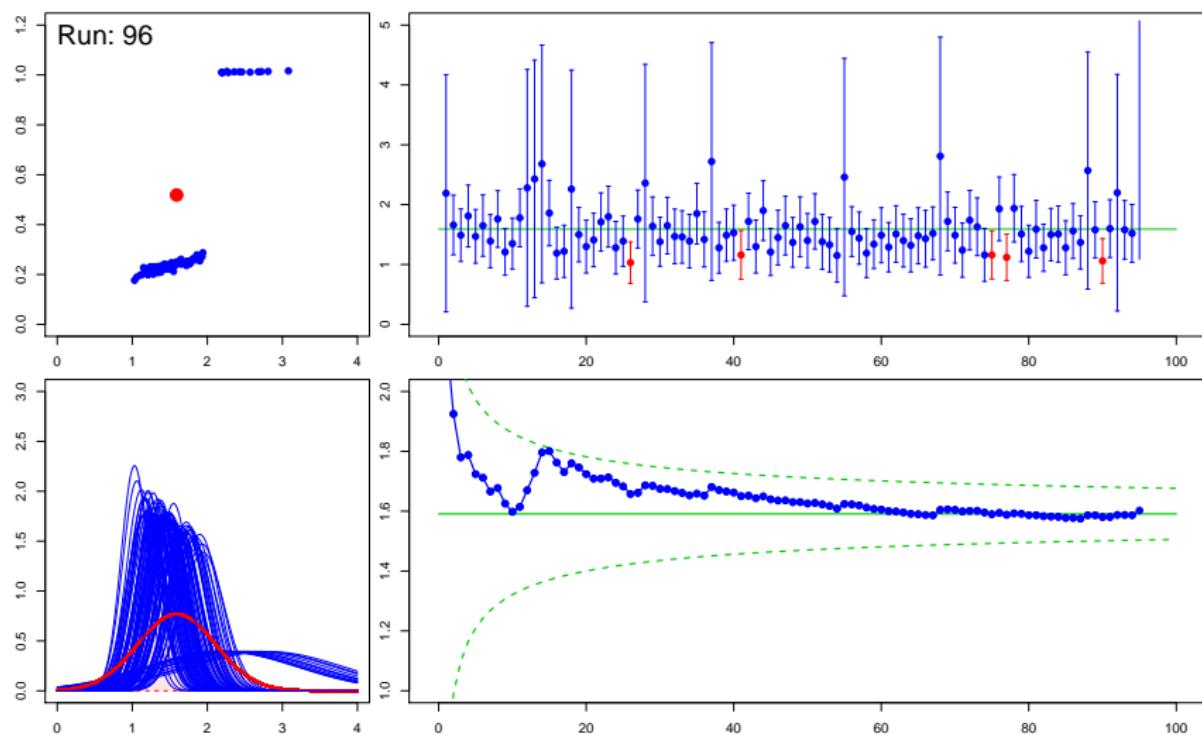
This approximative CI is shorter than the respective CI using the  $t$  distribution:  $[60.8268; 83.1733]$ . Notice the problems which may arise when using approximations.

The following examples illustrate this effect, e.g. that approximations may not always be used without concern.

# Simulation using the exponential distribution $(\lambda = 2)$



# Simulation using discrete distribution with outlier



# CI for $E X$ , arbitrary distribution, $\text{Var } X$ unknown, without replacement

The random variable

$$Z = \frac{\bar{X} - E X}{\sqrt{\frac{S^2}{n} \cdot \frac{N-n}{N-1}}}$$

does approximately follow a standard normal distribution. For  $n > 30$ , following the central limit theorem of Lindeberg and Lévy, the  $(1 - \alpha)$  confidence interval is

$$\left[ \bar{X} - z(1 - \alpha/2) \cdot \sqrt{\frac{S^2}{n} \cdot \frac{N-n}{N-1}}, \bar{X} + z(1 - \alpha/2) \cdot \sqrt{\frac{S^2}{n} \cdot \frac{N-n}{N-1}} \right]$$

- ▶ In case  $\text{Var } X$  is known, we substitute  $\text{Var } X$  for  $S^2$ .
- ▶ Mind the approximation conditions:  $n$  large and  $n$  not close to  $N$

## CI for proportions, variance unknown

Instead of  $\bar{X}$  we use the sample proportion  $P$ . We estimate the population variance using  $P \cdot (1 - P)$ . As the estimator distribution, using de Moivre and Laplace's theorem, the standard normal distribution is used. Mind the approximation conditions. We forego the continuity correction. The  $(1 - \alpha)$  confidence interval is

$$\left[ P - z\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{P(1 - P)}{n}}; P + z\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{P(1 - P)}{n}} \right].$$

## Example 8.20: CI for proportions (1)

A survey of  $n = 100$  students yielded a number of 15 students having a job. We get the 99% confidence interval

$$\left[ 0.15 - 2.575 \cdot \sqrt{\frac{0.15 \cdot 0.85}{100}}; 0.15 + 2.575 \cdot \sqrt{\frac{0.15 \cdot 0.85}{100}} \right] = [0.058; 0.242].$$

Data input in R:

```
alpha <- 0.01
n <- 100
p <- 15/100
```

Check of approximation conditions in R:

```
n * p * (1 - p) > 9
```

```
[1] TRUE
```

```
0.1 <= p & p <= 0.9
```

```
[1] TRUE
```

## Example 8.20: CI for proportions (2)

99% confidence interval in R:

```
CI <- vector()
CI[1] <- p - qnorm(p = 1 - alpha/2)*sqrt((p * (1 - p))/n)
CI[2] <- p + qnorm(p = 1 - alpha/2)*sqrt((p * (1 - p))/n)

round(CI, digits = 3)
```

```
[1] 0.058 0.242
```

## Example 8.21: see Ex. 8.14 (1)

Determination of needed sample size

We search the sample size for which the 95% CI is at most 5 units long.

We have

$$\left[ 72 - 1.96 \cdot \frac{30}{\sqrt{n}}; 72 + 1.96 \cdot \frac{30}{\sqrt{n}} \right].$$

This yields a length of  $d = 2 \cdot 1.96 \cdot 30 / \sqrt{n}$ . Using

$$2 \cdot 1.96 \cdot \frac{30}{\sqrt{n}} \leq 5$$

we finally get

$$n \geq \left( 2 \cdot 1.96 \cdot \frac{30}{5} \right)^2 = 553.1904 \quad .$$

We need a sample size of  $n \geq 554$ .

## Example 8.21: see Ex. 8.14 (2)

Calculation of  $n$  in R:

```
alpha <- 0.05
Quantile <- qnorm(p = 1 - alpha/2)
sigma <- 30
d <- 5

n_min <- ceiling((2 * Quantile * sigma/d)^2)

n_min
[1] 554
```

# Elements of Statistics

## Chapter 9: Hypothesis testing

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

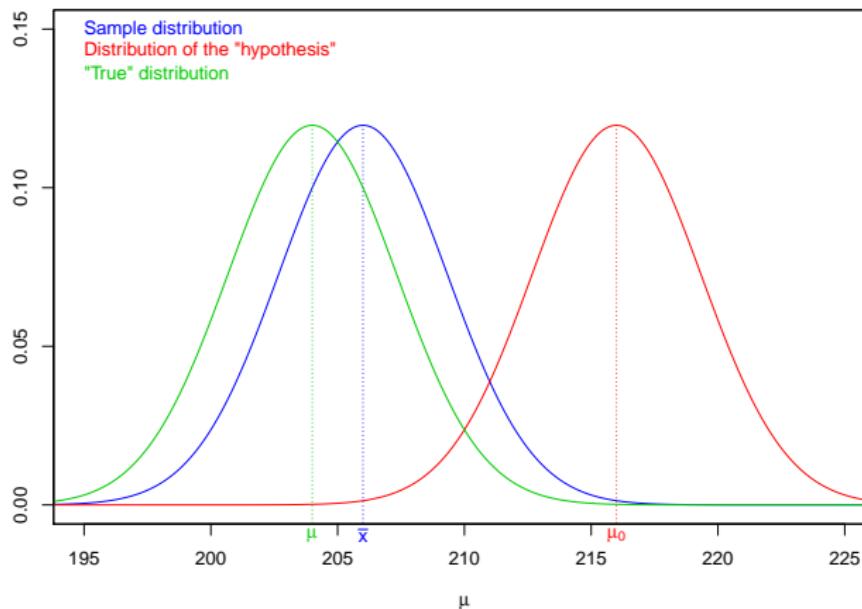
University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# General idea



- ▶ Analysis of sample (estimation, e.g. using  $\bar{x}$ )
- ▶ Hypothesis for population (e.g.  $\mu_0$  for  $\mu$ )
- ▶ True distribution *unknown* in reality

# Hypothesis testing

- ▶ Formulation of a *belief* for the population  
Example: Population is  $N(210; 30^2)$ -distributed  
→ Working hypothesis
- ▶ Check if sample results are *compatible* with working hypothesis
- ▶ Formulation of a decision rule:
  - ▶ right decision regarding working hypothesis
  - ▶ wrong decision regarding working hypothesis
- ▶ It can only be checked if the observations in the sample are realisations of the distribution postulated in the working hypothesis.
  - ▶ Is the sample result very unlikely given the working hypothesis?
  - ▶ Is the sample result *plausible* given the working hypothesis?

## Example 9.1: Typewriting

The long-time experience in a two-year typewriting class is that the final 10 minute exam yields test results which are roughly  $N(210; 30^2)$ -distributed. A new teaching method has been used for the last class and for a sample of  $n = 81$  pupils the following values have been recorded:  $\bar{x} = 218$  and  $s^*{}^2 = 28^2$ .

The working hypothesis is:

The new method does actually improve results.

The null hypothesis is:

The new method does not improve results.

Formally:

$$\bar{X} \sim N(210; 30^2)$$

We check if the  $n = 81$  sample values can be realisations of this distribution.

## Example 9.2: Surgery technique

The long-time experience is that in 22.8% of cases complications arise from a severe type of surgery. In a specialised hospital a new technique is studied. Out of the  $n = 22$  surgeries already performed, 4 complications resulted ( $4/22 = 18.18\%$ ).

Can we infer that the new technique is actually to be considered a progress?

Or do we have to consider this result as *random*?

## Example 9.3: Product batches

A certain batch of a product is considered acceptable if the share of scrap is at most 5%. A simple random sample of size  $n = 100$  yielded 7 pieces of scrap.

Should we dismiss this batch?

How many elements should be checked to avoid a wrong decision?

# One-sample tests

1. Test of localisation
2. Test of deviation
3. Combination of 1. and 2.
4. Test of distribution
  - ▶ Distribution law, e.g. normal distribution
  - ▶ Distribution including parameters

## Example 9.4: Readiness for school

It is examined if children are ready for school. The  $n = 400$  children examined are divided into two subgroups of size  $n_1 = 240$  and  $n_2 = 160$ , respectively, depending on the mother's working status. The following values are observed for a metric test score:  $\bar{x}_1 = 66.3$  and  $\bar{x}_2 = 78.4$  as well as  $s_1^* = 12.8$  and  $s_2^* = 13.2$ .

Working hypothesis:

- a) Working mothers' children reach a different score than non-working mothers' children *on average* (higher/lower).
- b) The former group's members' score is *10 points higher* on average.

## Two-sample tests

- ▶ Comparison of localisation
- ▶ Comparison of deviation
- ▶ Comparison of distribution

In Example 9.4 we may divide children into  $k$  subgroups. We would then speak of a  $k$ -sample case from which we abstract here.

## Example 9.5: Statistics exam

$n = 200$  students have to take two statistics exams, a written and an oral one.

In both exams students can reach a maximum of 100 points. The sample results are  $\bar{x}_w = 64.3$  with  $s_w = 9.8$  for the written exam and  $\bar{x}_o = 69.3$  with  $s_o = 7.6$  for the oral exam.

It should be checked if the two results are not independent of each other or the results of the oral exam are systematically better than the results of the written exam.

In contrast to Example 9.4 there are not two groups of observations here, but there is rather one group of students with two observations per student which may be presented as a tuple  $(x_{i,w}; x_{i,o})$ . Therefore, we speak of connected samples here, instead of unconnected samples like in Example 9.4.

# Types of hypotheses (1)

Every claim about the distribution of a variable or the relation of variables is called a statistical hypothesis.

We distinguish between working hypotheses and null hypotheses. The user's interest typically lies in the working hypothesis.

In statistical hypothesis testing the null hypothesis takes center stage. A statistical procedure to reach a decision regarding the *compatibility* of sample result and null hypothesis is called statistical test.

## Types of hypotheses (2)

In general, only *statistically noticeable* results of an investigation can indicate a possible incompatibility of sample result and null hypothesis. Therefore, the null hypothesis should (if possible at all) be chosen in such a way that if such an incompatibility occurs and the null hypothesis must consequently be dismissed, the actual *belief* of the investigator (working hypothesis) is substantiated. This can be achieved by reversing the working hypothesis.

If no such incompatibility is found, it does not mean that the null hypothesis is correct. Using the observations of the sample an obvious (significant) contradiction between the sample and the null hypothesis (just) cannot be detected.

# Parameter and distribution hypotheses

## Parameter hypothesis

Determination of parameters of a hypothesis (see Example 10.1):

- ▶  $H_0 : \mu = \mu_0 = 210$
- ▶  $H_0 : \sigma^2 = \sigma_0^2 = 30^2$
- ▶  $H_0 : \mu = \mu_0 = 210 \wedge \sigma^2 = \sigma_0^2 = 30^2$

## Distribution hypothesis

Determination of the distribution type of a hypothesis (see Example 10.1):

- ▶ Fully specified:  
 $H_0$  : The realised variable is  $N(210; 30^2)$ -distributed.
- ▶ Partially specified:  
 $H_0$  : The realised variable is normally distributed.  
→ Goodness of fit tests

# Point and interval hypotheses

## Point hypothesis

Determination of an exact parameter value:

$$H_0 : \mu = \mu_0 = 210$$

## Interval hypothesis

Determination of a value interval:

- ▶ One-sided hypothesis (greater than or equal to):

$$H_0 : \mu \geq \mu_0 = 210$$

- ▶ One-sided hypothesis (less than or equal to):

$$H_0 : \mu \leq \mu_0 = 210$$

## Alternative hypothesis

### Alternative hypothesis

In applications (and because of theoretical considerations) a null hypothesis is always *joined* by an alternative hypothesis. It is specifically called complementary if it is defined as  $H_0$  *does not apply*.

If the null hypothesis is  $H_0 : \mu = \mu_0 = 210$  then  $H_1 : \mu \geq \mu_0 = 210$  is one of the possible alternative hypotheses and  $H_1 : \mu \neq \mu_0 = 210$  is the complementary alternative hypothesis.

The complementary alternative hypothesis to  $H_0 : \mu \geq \mu_0 = 210$  is  $H_1 : \mu < \mu_0 = 210$ .

## General approach to hypothesis testing

Basis: Simple random sample of size  $n$  (WR or WOR)

- ▶  $S$  is set of all possible sample realisations of the sample vector  $(X_1, \dots, X_n)$
- ▶  $H_0$  assumed to be true  
(at first parameter point hypothesis)
- ▶ Decomposition of  $S = C \cup \bar{C}$  with  $P(C)$  small, so that
  - ▶  $(x_1, \dots, x_n) \in C \Rightarrow H_0$  is rejected
  - ▶  $(x_1, \dots, x_n) \in \bar{C} \Rightarrow H_0$  is not rejected

is used as the basis for the test decision.

$P(C) = \alpha$  is called *probability of error* or *level of significance*.  
 $C$  is called *critical region*.

The decomposition of  $S$  is accomplished using a suitable sample function which is called *test statistic*. Its distribution under  $H_0$  is called *null distribution*.

## Example 9.6: see Ex. 9.1 (1)

Let  $X$  be normally distributed with known variance  $\sigma^2 = 30^2$ . We want to test

$$H_0 : \mu = \mu_0 = 210$$

using a level of significance of  $\alpha = 0.05$ .

Under  $H_0$  we have  $\bar{X} \sim N\left(210; \left(\frac{30}{9}\right)^2\right)$ .

Data input in R:

```
sigma <- 30; mu0 <- 210; alpha <- 0.05; n <- 81
```

Very large or very low realisations of  $\bar{X}$  are considered to be unlikely.

## Example 9.6: see Ex. 9.1 (2)

Then

$$P\left(\bar{X} < \mu_0 - z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.025$$

and

$$P\left(\bar{X} > \mu_0 + z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.025$$

characterise the  $\alpha \cdot 100\%$  (here: 5%) of possible values which are *the least plausible* given  $H_0$ . Precisely, we have:

$$C = \{(x_1, \dots, x_n) | \bar{x} < 203.5 \vee \bar{x} > 216.5\} \quad .$$

Calculation of the *borders* of  $C$  in R:

```
C <- vector()
C[1] <- mu0 - qnorm(1-alpha/2)*sigma/sqrt(n)
C[2] <- mu0 + qnorm(1-alpha/2)*sigma/sqrt(n)
C
```

[1] 203.4668 216.5332

# Types of error in testing

		Null hypothesis $H_0$ is	
		not rejected	rejected
$H_0$ is	true	Right decision	Type 1 error
	false	Type 2 error	Right decision

Type 1 error ( $\alpha$  error): A correct hypothesis is wrongly rejected.

Type 2 error ( $\beta$  error): A false hypothesis is wrongly not rejected.

Both errors cannot be simultaneously omitted (their probability kept low).

The level of significance  $\alpha$  is the supremal probability that the null hypothesis is wrongly rejected – regarding all possible parameter values under  $H_0$ .

## Example 9.7: see Ex. 9.6 (1)

Test of  $H_0 : \mu = \mu_0 = 210$ , with  $X \sim N(\mu; 30^2)$ ,  $n = 81$  and  $\alpha = 0.05$

- a) Probability of a type 1 error:  
 $H_0$  is true, but it is rejected.

$$P((x_1, \dots, x_n) \in C | \mu = 210, \alpha = 0.05) = 0.05$$

Determination of significance level  $\alpha$ !

Calculation of the probability for a type 1 error in R:

```
Prob_a <- pnorm(q = C[1], mean = mu0, sd = sigma/sqrt(n)) +  
        1 - pnorm(q = C[2], mean = mu0, sd = sigma/sqrt(n))
```

```
Prob_a
```

```
[1] 0.05
```

## Example 9.7: see Ex. 9.6 (2)

b) Probability of a type 2 error:

$H_0$  is false, but it is not rejected.

The answer depends on  $\mu$ , e.g.  $\mu = 216$ .

$$\begin{aligned} P((x_1, \dots, x_n) \in \bar{C} | \mu = 216, \alpha = 0.05) &= \\ \Phi\left(216.5 | 216; \frac{900}{81}\right) - \Phi\left(203.5 | 216; \frac{900}{81}\right) &= 0.56 \end{aligned}$$

Calculation of the probability for a type 2 error in R:

```
Prob_b <- pnorm(q = C[2], mean = 216, sd = sigma/sqrt(n)) -  
        pnorm(q = C[1], mean = 216, sd = sigma/sqrt(n))
```

```
round(Prob_b, digits=2)
```

```
[1] 0.56
```

# Power function and operating characteristic

## Power function

The power function  $\alpha(\pi)$  specifies the probability of rejecting the null, given  $n$ ,  $\alpha_0$  and  $H_0$ , for all possible parameter values  $\pi$ . We have:

$$\alpha(\pi) = \alpha(\pi | n, \alpha_0, H_0) = P((X_1, \dots, X_n) \in C | \pi)$$

## Operating characteristic

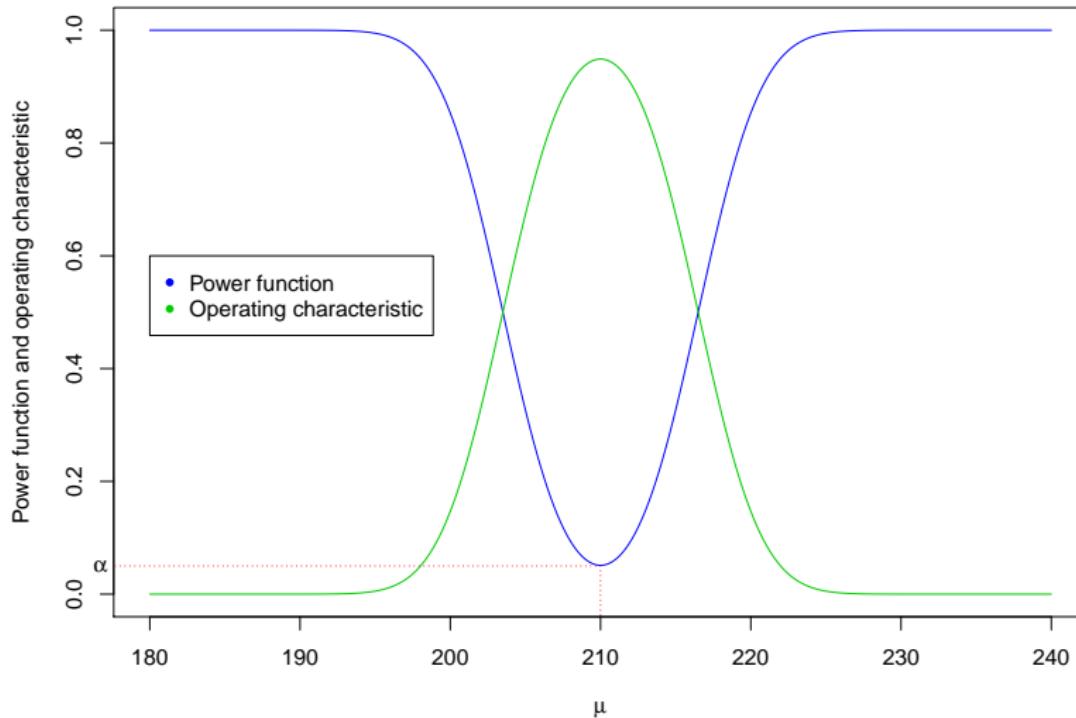
The operating characteristic  $\beta(\pi)$  specifies the probability of not rejecting the null, given  $n$ ,  $\alpha_0$  and  $H_0$ , for all possible parameter values  $\pi$ . We have:

$$\beta(\pi) = \beta(\pi | n, \alpha_0, H_0) = P((X_1, \dots, X_n) \in \bar{C} | \pi)$$

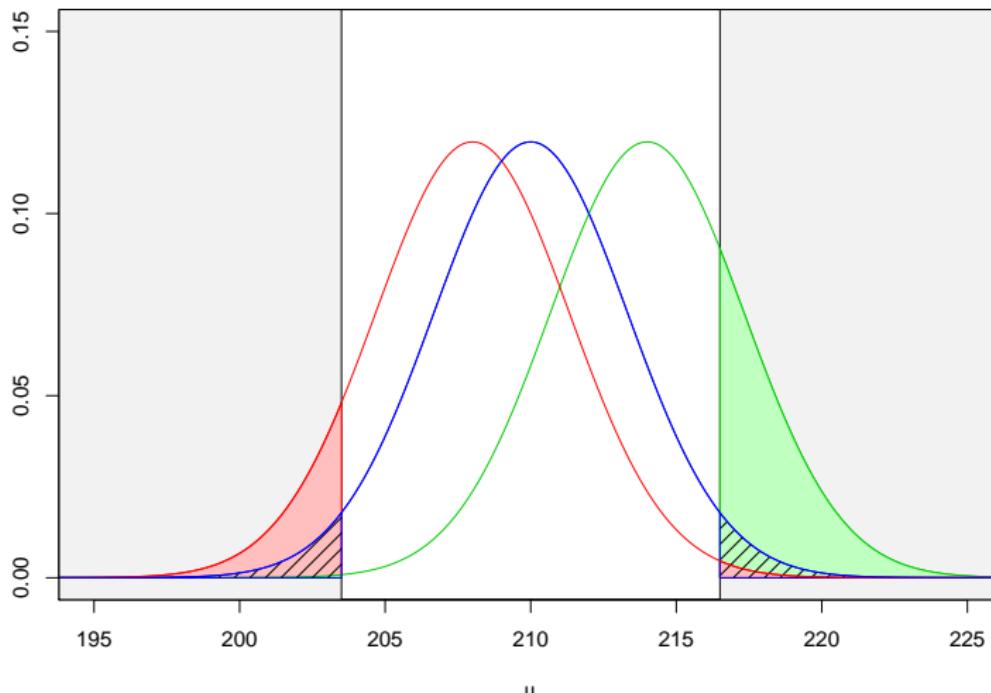
For all possible  $\pi$  the following holds:  $\alpha(\pi) + \beta(\pi) = 1$ .

# Power function and operating characteristic

$$H_0 : \mu = \mu_0 = 210$$



# Determination of the power function for $H_0 : \mu = \mu_0 = 210$



## Example 9.8: One-sided hypothesis

Contrary to Example 10.7, now the null is  $H_0 : \mu \geq \mu_0 = 210$  (one-sided hypothesis; greater than or equal to). The critical region of the sample distribution  $\bar{X} \sim N(\mu; \frac{100}{9})$  for  $\mu \geq 210$  is to be determined. The significance level is at its maximum exactly at  $\mu = 210$  and should not surpass the given value of  $\alpha = 0.05$ . Using

$$\begin{aligned} P\left(\bar{X} < \mu_0 - z(1 - \alpha) \cdot \frac{\sigma}{\sqrt{n}}\right) \\ = P\left(\bar{X} < 210 - 1.64 \cdot \frac{10}{3}\right) = P(\bar{X} < 204.5) = 0.05 \end{aligned}$$

we get the critical region

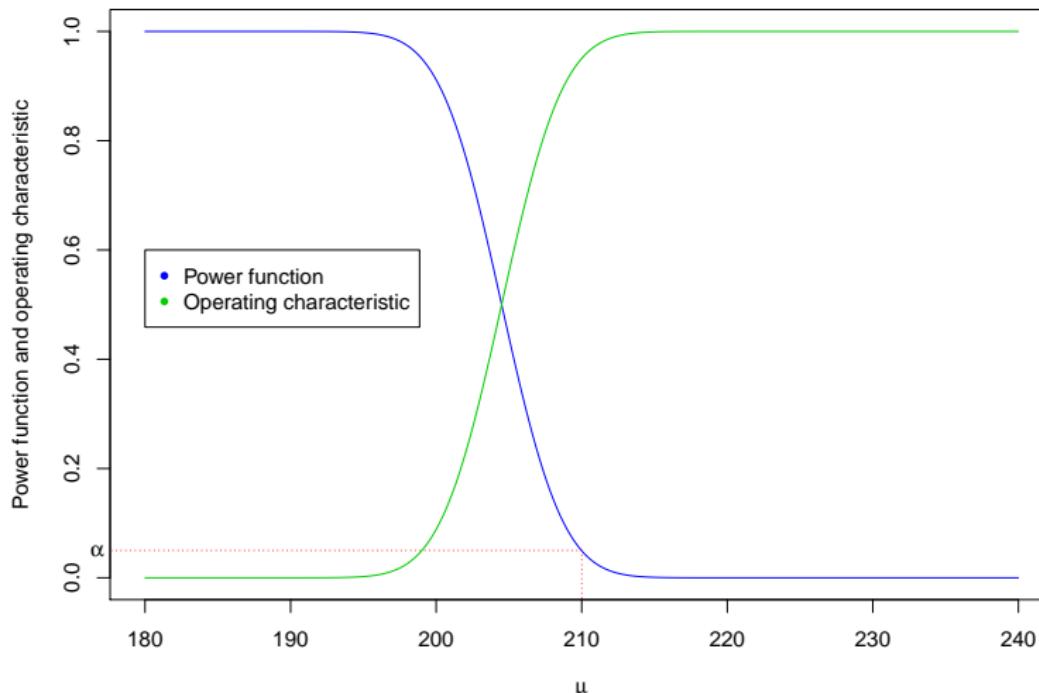
$$C = \left\{ (x_1, \dots, x_n) \mid \bar{x} < 204.5 \right\}$$

The probability for a type 1 error is *at most 5%*.

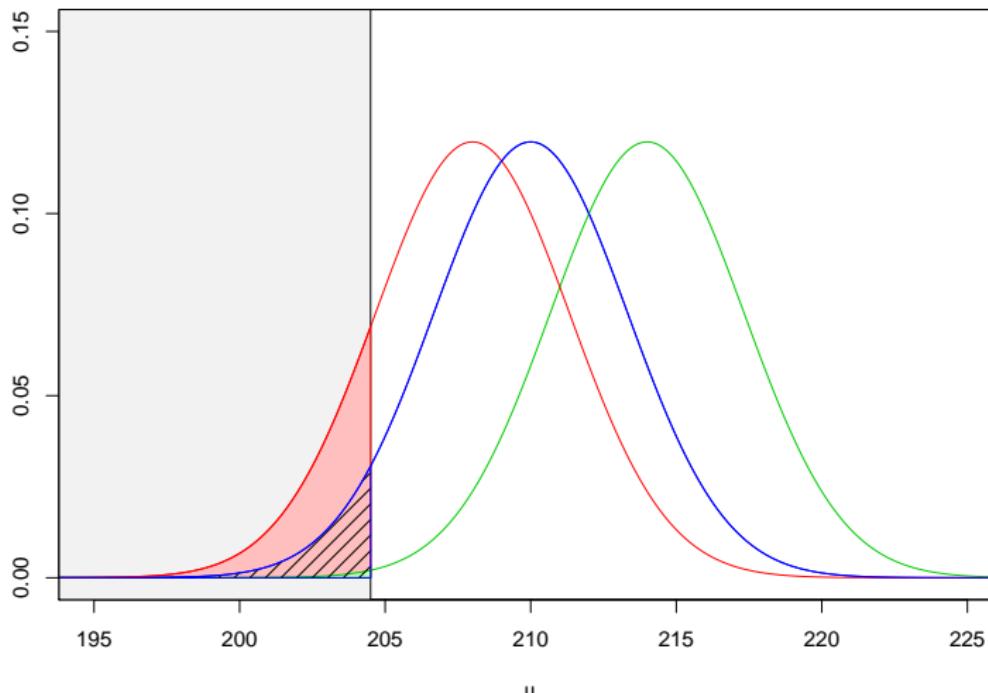
Analogously, for the complementary null hypothesis  $H_0 : \mu \leq \mu_0 = 210$  we have:  $C = \left\{ (x_1, \dots, x_n) \mid \bar{x} > 215.5 \right\}$ .

# Power function and operating characteristic

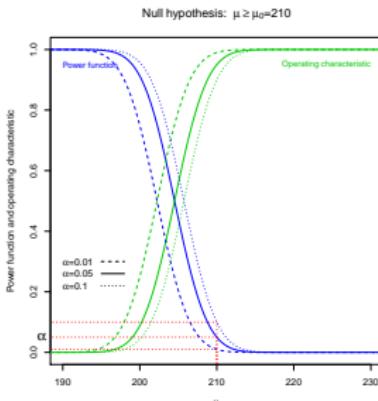
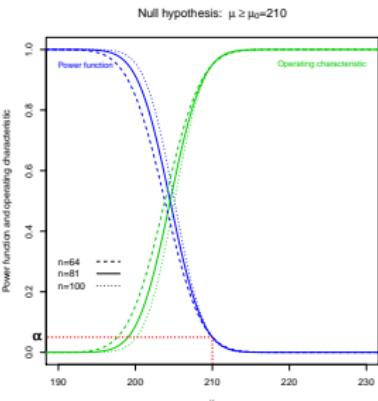
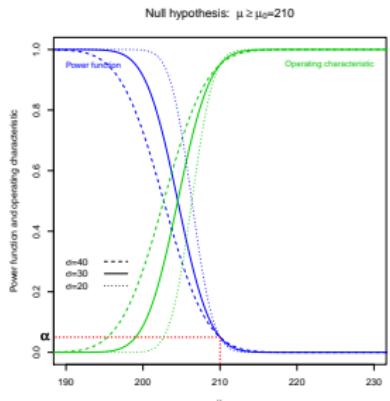
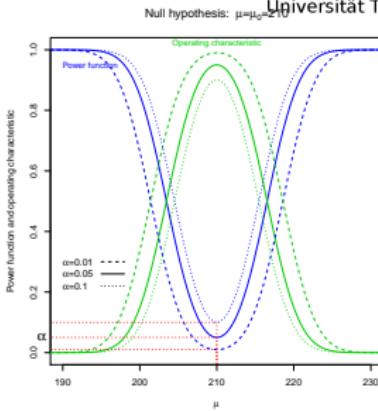
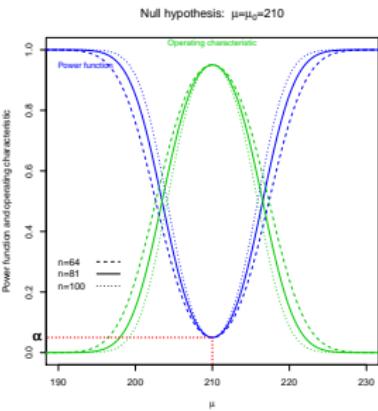
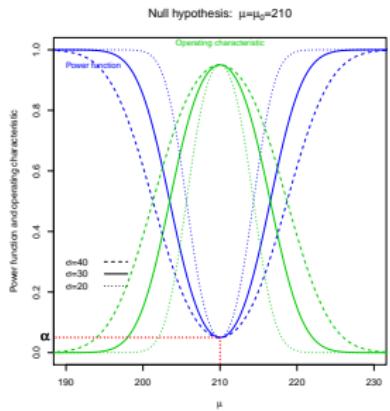
$$H_0 : \mu \geq \mu_0 = 210$$



# Determination of the power function for $H_0 : \mu \geq \mu_0 = 210$



# Influence of $n$ , $\sigma^2$ and $\alpha$ on PF and OC



# General approach of hypothesis testing

1. Postulation of working hypothesis and, consequently, postulation of
  - ▶ null hypothesis  $H_0$  and of
  - ▶ alternative hypothesis  $H_1$
2. Setting of significance level  $\alpha$
3. Statement/Calculation of test statistic
4. Setting of critical region  $C$
5. Implementation of test and interpretation of results

In practice, the null hypothesis is derived by negation of the working hypothesis. Therefore, a rejection of the null hypothesis lends support to the working hypothesis. Such a negation is only possible for one-sided hypotheses.

## Example 9.9: Clinical trial (1)

In a clinical trial after oral intake of an antibiotic a concentration of approx.  $8 \mu\text{g/l}$  of the agent was measured at the centre of inflammation. A doctor assumes that under *normal conditions* at most  $5 \mu\text{g/l}$  are plausible. A control study with  $n = 100$  patients yielded  $\bar{x} = 3 \mu\text{g/l}$  and  $s^2 = 25 (\mu\text{g/l})^2$ .

Test the doctor's assumption using  $\alpha = 0.05$ .

- ▶  $H_0 : \mu \geq \mu_0 = 5$  and  $H_1 : \mu < \mu_0 = 5$  (negation of working h.)
- ▶  $\alpha = 0.05$

Data input in R:

```
SpMean <- 3; SpVar <- 25
mu0 <- 5      ; alpha <- 0.05; n <- 100
```

## Example 9.9: Clinical trial (2)

- ▶ Test statistic:  $\frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \sim N(0; 1)$  ( $n = 100 > 30!$ )
- ▶ Critical region:  $C = \{(x_1, \dots, x_n) | \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} < z(\alpha)\}$
- ▶  $\frac{3 - 5}{5} \cdot \sqrt{100} = -4 < -1.645$ , so that  $-4 \in C$   
 $H_0$  is rejected and the working hypothesis is statistically supported (at  $\alpha = 0.05$ ).

Test decision in R:

```
Teststat <- (SpMean - mu0)/sqrt(SpVar) * sqrt(n)
c_stat <- qnorm(p = alpha)
```

Teststat	c_stat
[1] -4	[1] -1.644854
<b>Teststat &lt; c_stat</b>	
[1] TRUE	

## *p value or exceeding probability*

Software packages typically report the so called *p value (exceeding probability)* of a test.

The *p* value indicates the probability that, given  $H_0$ , the observed value of the test statistic or a value which is *even more unfavourable* (for the null hypothesis) results.

$H_0$  is rejected if the *p* value is smaller than the significance level chosen **in advance**.

## Example 9.9: Clinical trial (3)

In Example 9.9 a value of the test statistic of  $z = -4$  was calculated. In terms of the null hypothesis all values of the test statistic  $z'$  with  $z' < -4$  are *worse*. Therefore, the respective  $p$  value (not tabulated) is

$$P(Z < -4) = \Phi(-4) = 3.167 \cdot 10^{-5} .$$

As  $3.167 \cdot 10^{-5} \ll 0.05$  the null hypothesis has to be rejected.

Determination of the  $p$  value in R:

```
p_value <- pnorm(q = Teststat)
```

```
p_value
```

```
[1] 3.167124e-05
```

### Attention:

Two-sided null hypotheses have a second interval of *more unfavourable* values. Accordingly, the calculated value might have to be doubled.

# Test for $\mu$

We are looking for the null distribution of the standardised random variable  $\bar{X}$ . Analogously to estimation procedures we get:

- ▶ POP normally distributed,  $\sigma^2$  known

$$\frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} \sim N(0; 1)$$

- ▶ POP normally distributed,  $\sigma^2$  unknown

$$\frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \sim t(n - 1)$$

- ▶ POP distribution unknown,  $\text{Var}(X)$  known,  $n > 30$

$$\frac{\bar{X} - \mu_0}{\sqrt{\text{Var}(X)}} \cdot \sqrt{n} \sim N(0; 1)$$

- ▶ POP distribution unknown,  $\text{Var}(X)$  unknown,  $n > 30$ , WR or WOR

$$\frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \sim N(0; 1) \quad \text{or} \quad \frac{\bar{X} - \mu_0}{S \cdot \sqrt{\frac{N-n}{N-1}}} \cdot \sqrt{n} \sim N(0; 1)$$

## Different cases of hypothesis testing for $\mu$

1. Two-sided problem ( $H_0 : \mu = \mu_0$ ):

Comparison of test statistic with  $\alpha/2$ - or  $1 - \alpha/2$ -quantile of null distribution

$$C = \left\{ (x_1, \dots, x_n) \left| \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} < z(\alpha/2) \vee \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} > z(1 - \alpha/2) \right. \right\}$$

2. One-sided problem:

- a) Greater than or equal to ( $H_0 : \mu \geq \mu_0$ ):

Comparison of test statistic with  $\alpha$ -quantile of null distribution

$$C = \left\{ (x_1, \dots, x_n) \left| \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} < z(\alpha) \right. \right\}$$

- b) Less than or equal to ( $H_0 : \mu \leq \mu_0$ ):

Comparison of test statistic with  $1 - \alpha$ -quantile of null distribution

$$C = \left\{ (x_1, \dots, x_n) \left| \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} > z(1 - \alpha) \right. \right\}$$

## Example 9.10: Two-sided test (1)

The following  $n = 20$  values of a normally distributed random variable are observed: 4.30; 4.30; 3.88; 6.81; 5.17; -0.53; 4.05; 3.34; 6.07; 6.58; 6.94; 0.96; 6.14; 4.20; 8.35; 7.01; 3.39; 3.96; 3.19; 4.21.

The null hypothesis  $H_0 : \mu = \mu_0 = 3$  is to be tested using  $\alpha = 0.05$ . We know that  $\sigma^2 = 4$ .

- ▶  $H_0 : \mu = \mu_0 = 3, H_1 : \mu \neq \mu_0 = 3, \alpha = 0.05, n = 20, \sigma^2 = 4$

Data input in R:

```
x9_10 <- c(4.30, 4.30, 3.88, 6.81, 5.17, -0.53, 4.05, 3.34,  
          6.07, 6.58, 6.94, 0.96, 6.14, 4.20, 8.35, 7.01,  
          3.39, 3.96, 3.19, 4.21)  
sigma <- sqrt(4)  
mu0 <- 3  
alpha <- 0.05  
n <- 20  
SpMean <- mean(x9_10)
```

## Example 9.10: Two-sided test (2)

- ▶  $\frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} \sim N(0; 1)$ ,  $z(0.025) = -1.96$ ,  $z(0.975) = 1.96$
- ▶  $C = \left\{ (x_1, \dots, x_n) \mid \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} < z(\alpha/2) \vee \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} > z(1 - \alpha/2) \right\}$
- ▶ As  $\bar{x} = 4.616$  we have  $z = \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \frac{4.616 - 3}{2} \cdot \sqrt{20} = 3.613$ .  
As  $3.613 \in C$  ( $3.613 > 1.96$ ),  $H_0$  is rejected.

Test decision in R:

```
Teststat <- (SpMean - mu0)/sigma * sqrt(n)
c_stat <- vector()
c_stat[1] <- qnorm(p = alpha/2)
c_stat[2] <- qnorm(p = 1 - alpha/2)

round(Teststat, digits = 3)
[1] 3.613

Teststat < c_stat[1] | Teststat > c_stat[2]
[1] TRUE
```

## Example 9.11: see Ex. 9.10 (1)

Now let the distribution of the population and the variance  $\sigma^2$  be unknown. Once again, we want to test the null hypothesis  $H_0 : \mu = \mu_0 = 3$  using a significance level of  $\alpha = 0.05$ .

- ▶  $H_0 : \mu = \mu_0 = 3, H_1 : \mu \neq \mu_0 = 3, \alpha = 0.05, n = 20;$
- ▶  $\frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \sim N(0; 1), z(0.025) = -1.96, z(0.975) = 1.96$
- ▶  $C = \left\{ (x_1, \dots, x_n) \middle| \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} < z(\alpha/2) \vee \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} > z(1-\alpha/2) \right\}$
- ▶ As  $\bar{x} = 4.616$  and  $s^2 = 4.4912$  we have  
$$z = \frac{\bar{x} - \mu_0}{S} \cdot \sqrt{n} = \frac{4.616 - 3}{\sqrt{4.4912}} \cdot \sqrt{20} = 3.410.$$
As  $3.410 \in C$  ( $3.410 > 1.96$ ),  $H_0$  is rejected.

Notice that  $t(0.975; 19) = 2.093 > 1.96 = z(0.975).$

( $t$  test is used in software packages.)

**Attention:**  $n \not> 30!$

Approximation is still inadmissible according to text book.

## Example 9.11 (2): Test of point hypothesis

Application of hypothesis test in R

```
t.test(x9_10, alternative="two.sided", mu=3, conf.level=0.95)
```

One Sample t-test

```
data: x9_10
t = 3.4102, df = 19, p-value = 0.002936
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.624168 5.607832
sample estimates:
mean of x
4.616
```

The output contains the test statistic  $t = 3.4102$  ( $t$ -distribution with  $df = 19$  degrees of freedom), the  $p$  value 0.0029, the point estimate  $\bar{x} = 4.616$  and the corresponding confidence interval [3.624; 5.608].

## Example 9.11 (3): Test of one-sided hypothesis (greater)

Application of hypothesis test in R

```
t.test(x9_10, alternative="less", mu=3, conf.level=0.95)
```

One Sample t-test

```
data: x9_10
t = 3.4102, df = 19, p-value = 0.9985
alternative hypothesis: true mean is less than 3
95 percent confidence interval:
-Inf 5.435393
sample estimates:
mean of x
4.616
```

The test statistic  $t = 3.4102$  has to be compared to  $t(0.05; 19) = -1.729$ .  
 $H_0$  is not rejected here, which is illustrated by the  $p$  value as well:  
 $0.9985 \gg 0.05$ . **Notice: Output states alternative hypothesis!**

## Example 9.11 (4): Test of one-sided hypothesis (less)

Application of hypothesis test in R:

```
t.test(x9_10, alternative="greater", mu=3, conf.level=0.95)
```

One Sample t-test

```
data: x9_10
t = 3.4102, df = 19, p-value = 0.001468
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 3.796607     Inf
sample estimates:
mean of x
 4.616
```

The test statistic  $t = 3.4102$  has to be compared to  $t(0.95; 19) = 1.729$ .  
 $H_0$  is rejected here, which is illustrated by the  $p$  value as well:  
 $0.001468 \ll 0.05$ . **Notice: Output states alternative hypothesis!**

## Example 9.12: One-sided test (1)

A sample of size  $n = 25$  is drawn from a normally distributed population. The following values have been calculated  $\bar{x} = 197.7$  and  $s^2 = 42.25$ .

Test the null hypothesis  $H_0 : \mu \geq \mu_0 = 200$  using  $\alpha = 0.05$ .

- ▶  $H_0 : \mu \geq \mu_0 = 200$ ,  $H_1 : \mu < \mu_0 = 200$ ,  $\alpha = 0.05$ ,  $n = 25$ ;
- ▶  $\frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n} \sim t(n - 1)$ ,  $t(0.05; 24) = -1.711$
- ▶  $C = \left\{ (x_1, \dots, x_n) \mid \frac{\bar{X} - \mu_0}{s} \cdot \sqrt{n} < t(\alpha; n - 1) \right\}$
- ▶ As  $\bar{x} = 197.7$  and  $s^2 = 42.25$  we have  
$$t = \frac{\bar{x} - \mu_0}{s} \cdot \sqrt{n} = \frac{197.7 - 200}{\sqrt{42.25}} \cdot \sqrt{25} = -1.769.$$
As  $-1.769 \in C$  ( $-1.769 < -1.711$ ),  $H_0$  is rejected.

Data input in R:

```
load("Example9-12.RData")
```

## Example 9.12: One-sided test (2)

Test decision in R:

```
SpMean <- mean(x9_12)
SpVar <- var(x9_12)
mu0 <- 200
alpha <- 0.05
n <- length(x9_12)

t.test(x = x9_12, alternative = "less", mu = mu0,
       conf.level = 1 - alpha)

One Sample t-test
```

```
data: x9_12
t = -1.7725, df = 24, p-value = 0.0445
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
-Inf 199.9199
sample estimates:
mean of x
197.6956
```

## Example 9.13: see Ex. 9.12 (1)

Test the null hypothesis  $H_0 : \sigma^2 \leq \sigma_0^2 = 30$  using  $\alpha = 0.05$ .

- $H_0 : \sigma^2 \leq \sigma_0^2 = 30$ ,  $H_1 : \sigma^2 > \sigma_0^2 = 30$ ,  $\alpha = 0.05$ ,  $n = 25$ ;

Data input in R:

```
sigmaq0 <- 30
```

## Example 9.13: see Ex. 9.12 (2)

- ▶  $\frac{n-1}{\sigma_0^2} \cdot s^2 \sim \chi^2(n-1), \chi^2(0.95; 24) = 36.415$
- ▶  $C = \left\{ (x_1, \dots, x_n) \mid \frac{n-1}{\sigma_0^2} \cdot s^2 > \chi^2(1 - \alpha; n-1) \right\}$
- ▶ As  $s^2 = 42.25$  we have  $\chi^2 = \frac{n-1}{\sigma_0^2} \cdot s^2 = \frac{24}{30} \cdot 42.25 = 33.8$ .  
 As  $33.8 \notin C$  ( $33.8 \not> 36.415$ ),  $H_0$  is not rejected.

Test decision in R:

```
c_stat <- qchisq(p = 1 - alpha, df = n - 1)
Teststat <- (n-1) / sigmaq0 * SpVar
```

`c_stat`

[1] 36.41503

`Teststat`

[1] 33.80213

`Teststat > c_stat`

[1] FALSE

## Example 9.14: Another urn example

There are red and black balls in an urn. It is postulated that the share of red balls in the urn is  $\theta = 0.4$ . The following test is performed:

$n = 10$  balls are drawn with replacement.  $H_0$  will be rejected if 0, 1, 9, or 10 red balls are drawn. Then

$$\begin{aligned}\alpha &= P(X = 0, 1, 9, 10 | \theta = \theta_0 = 0.4) = \binom{10}{0} \cdot 0.4^0 \cdot 0.6^{10} \\ &\quad + \binom{10}{1} \cdot 0.4^1 \cdot 0.6^9 + \binom{10}{9} \cdot 0.4^9 \cdot 0.6^1 + \binom{10}{10} \cdot 0.4^{10} \cdot 0.6^0 = 0.0481\end{aligned}$$

is the significance level and

$$\beta(0.5) = 1 - 2 \cdot \left( \binom{10}{0} \cdot 0.5^{10} \cdot 0.5^0 + \binom{10}{1} \cdot 0.5^9 \cdot 0.5^1 \right) = 0.9785$$

is the probability of a type 2 error if  $\theta = 0.5$ .

## Example 9.15: Percentage test (1)

In a study on poverty in Germany  $n = 100$  individuals' income is recorded. A person is considered poor if her income is below the poverty threshold. It is postulated that exactly 25% of the population have to be considered poor. This null hypothesis is tested using  $\alpha = 0.05$ .

$$H_0 : \theta = \theta_0 = 0.25, H_1 : \theta \neq \theta_0, \alpha = 0.05, n = 100$$

Data input in R:

```
theta0 <- 0.25; alpha <- 0.05; n <- 100
```

According to de Moivre and Laplace's theorem we have:

$$\frac{X - n \cdot \theta_0}{\sqrt{n \cdot \theta_0 \cdot (1 - \theta_0)}} \sim N(0; 1).$$

$X$  is the number of poor people ( $n \cdot p$ ) in the sample.

We should consider a continuity correction as well.

The approximation conditions hold, as

$$100 \cdot 0.25 \cdot 0.75 = 18.75 > 9 \text{ and } 0.1 \leq 0.25 \leq 0.9.$$

## Example 9.15: Percentage test (2)

Test of approximation conditions in R:

```
n * theta0 * (1 - theta0) > 9
```

```
[1] TRUE
```

```
0.1 <= theta0 & theta0 <= 0.9
```

```
[1] TRUE
```

- ▶  $C = \left\{ (x_1, \dots, x_n) \middle| \frac{n \cdot p + 0.5 - n \cdot \theta_0}{\sqrt{n \cdot \theta_0 \cdot (1 - \theta_0)}} < z(\alpha/2)$   
 $\quad \vee \quad \frac{n \cdot p - 0.5 - n \cdot \theta_0}{\sqrt{n \cdot \theta_0 \cdot (1 - \theta_0)}} > z(1 - \alpha/2) \right\}$

- ▶ The sample yields 30 poor people:

$$z = \frac{30 + 0.5 - 25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = 1.270 < 1.96$$

$H_0$  is not rejected and the continuity correction is redundant.

## Example 9.15: Percentage test (3)

Hypothesis test in R:

```
p <- 0.3

c_stat <- vector()
c_stat[1] <- qnorm(p = alpha/2)
c_stat[2] <- qnorm(p = 1 - alpha/2)

Teststat <- vector()
Teststat[1] <- (n*p+0.5-n*theta0) / sqrt(n*theta0*(1-theta0))
Teststat[2] <- (n*p-0.5-n*theta0) / sqrt(n*theta0*(1-theta0))

c_stat
[1] -1.959964  1.959964

Teststat
[1] 1.270171 1.039230

Teststat[1] < c_stat[1] | Teststat[2] > c_stat[2]
[1] FALSE
```

## Example 9.15: Percentage test (4)

- ▶ The sample yields 40 poor people:

$$z = \frac{40 - 0.5 - 25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = 3.349 > 1.96$$

$H_0$  is rejected and the continuity correction is redundant.

- ▶ The sample yields 33 poor people:

$$z = \frac{33 + 0.5 - 25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = 1.963 > 1.96$$

$$z = \frac{33 - 0.5 - 25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = 1.732 < 1.96$$

Here we would make different decisions. In case of doubt we do not reject the null (conservative testing).

## Two connected samples

Observations are pairs of values:  $(X_1, X_2) \rightarrow D = X_1 - X_2$ .

We finally get to the one-sample case. We have:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

Often  $\delta_0 = 0$  is of interest (homogeneity hypothesis).

Notice that:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S_D} \cdot \sqrt{n} \sim N(0; 1)$$

We use the central limit theorem (CLT) here  
( $n > 30$ , if  $X_1$  and  $X_2$  are not already normally distributed).

Furthermore, we have  $d = \bar{x}_1 - \bar{x}_2$  and  $s_d^2 = s_1^2 + s_2^2 - 2 \cdot s_{12}$ .

## Example 9.16: Weddings (1)

For  $n = 50$  weddings, the age of the groom and the age of the bride are recorded  $(x_{1i}, x_{2i})$ . The descriptive statistics for the sample are:  $\bar{x}_1 = 31.5$ ;  $\bar{x}_2 = 28.5$ ;  $s_1^2 = 27.3$ ;  $s_2^2 = 23.8$  and  $s_{12} = 15.6$ . We want to test the null that the *typical* bride is at least 5 years younger than the *typical* groom ( $\alpha = 0.05$ ).

- ▶  $H_0 : \mu_1 - \mu_2 \geq \delta_0 = 5$  vs.  $H_1 : \mu_1 - \mu_2 < 5$ ,  $\alpha = 0.05$  and  $n = 50 > 30!$

Data input in R:

```
SpMean_X1 <- 31.5; SpMean_X2 <- 28.5
SpVar_X1 <- 27.3 ; SpVar_X2 <- 23.8
Cov_X1_X2 <- 15.6

delta0 <- 5; alpha <- 0.05; n <- 50
```

Checking approximation conditions in R:

```
n > 30
[1] TRUE
```

## Example 9.16: Weddings (2)

- ▶  $\frac{D - \delta_0}{S_d} \cdot \sqrt{n} \sim N(0; 1)$ ,  $z(0.05) = -1.645$ ,  $s_d^2 = s_1^2 + s_2^2 - 2s_{12} = 19.9$
- ▶  $C = \left\{ (x_1, \dots, x_n) \mid \frac{d - \delta_0}{s_d} \cdot \sqrt{n} < z(\alpha) \right\}$
- ▶ We have  $z = \frac{d - \delta_0}{s_d} \cdot \sqrt{n} = \frac{31.5 - 28.5 - 5}{\sqrt{19.9}} \cdot \sqrt{50} = -3.17$ .

As  $-3.17 \in C$  ( $-3.17 < -1.645$ ),  $H_0$  is rejected.

Test decision in R:

```
S_d2 <- SpVar_X1 + SpVar_X2 - 2 * Cov_X1_X2
Teststat <- (SpMean_X1 - SpMean_X2 - delta0) /
            sqrt(S_d2) * sqrt(n)
c_stat <- qnorm(p = alpha)
```

S_d2	c_stat	Teststat
[1] 19.9	[1] -1.644854	[1] -3.170213
Teststat < c_stat		
[1] TRUE		

## Comparison of expected values of two independent samples

$$H_0 : \mu_1 - \mu_2 = \delta_0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq \delta_0$$

$\bar{X}_1 \sim N(\mu_1; \frac{\sigma_1^2}{n_1})$  and  $\bar{X}_2 \sim N(\mu_2; \frac{\sigma_2^2}{n_2})$  are stochastically independent.

ND, variances known

Test statistic:  $Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Test distribution: Standard normal distribution

Arbitrary distribution, variances known

Test statistic:  $Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{\text{Var}(\bar{X}_1)}{n_1} + \frac{\text{Var}(\bar{X}_2)}{n_2}}}$

Test distribution: Standard normal distribution (CLT,  $n_1, n_2 > 30$ )

## ND, variances unknown but identical

We use the weighted average of the sample variances

$$S^2 := \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}.$$

Test statistic:  $T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Test distribution:  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom

## Arbitrary distribution, variances unknown and potentially not identical

Test statistic:  $Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

Test distribution: Standard normal distribution (CLT,  $n_1, n_2 > 30$ )

ND, variances unknown and not identical

Fisher-Behrens problem (approximative method)

$$\text{Test statistic: } T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Test distribution:  $t$ -distribution with  $[k]$  degrees of freedom, using

$$k = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}} \quad \text{with} \quad c = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

## Example 9.17: Independent samples (1)

Two independent samples from normally distributed populations yielded in R:

```
x9_17 <- c(3,5,8,7,9,12,17,10,5,9,14,22)  
y9_17 <- c(17,25,9,21,12,14,10,12,14,10,16,12,19,14,28)
```

A test for homogeneity of expected values ( $\alpha = 0.01$ ) yielded in R:

```
t.test(x = x9_17, y = y9_17, alternative="two.sided",  
       var.equal=FALSE, paired=FALSE, mu=0, conf.level=0.99)
```

Welch Two Sample t-test

data: x9\_17 and y9\_17

t = -2.5519, df = 23.968, p-value = 0.01751

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

-11.4240159 0.5240159

sample estimates:

mean of x mean of y

10.08333 15.53333

With  $y_{15} = 8$  instead of 28, the p value would have been 0.05066.

Calculate the test statistic and other relevant values *by hand*.

## Example 9.17: Independent samples (2)

An *expert* wants to be sure that the mean of the first population is smaller than the mean of the second population. Accordingly, he postulates  $H_0 : \mu_1 - \mu_2 \geq 0$  (the negation of his working hypothesis).

Hypothesis test in R:

```
t.test(x=x9_17, y=y9_17, alternative="less", var.equal=FALSE,  
       paired=FALSE, mu=0, conf.level=0.99)
```

Welch Two Sample t-test

```
data: x9_17 and y9_17  
t = -2.5519, df = 23.968, p-value = 0.008756  
alternative hypothesis: true difference in means is less than 0  
99 percent confidence interval:  
 -Inf -0.1270684  
sample estimates:  
mean of x mean of y  
 10.08333 15.53333
```

The null has to be rejected this time!

## Comparison of proportions of two independent samples

To be tested:  $H_0 : \theta_1 - \theta_2 = \delta_0$  ( $\geq$  or  $\leq$  possible as well)

$$\delta \neq \delta_0 = 0$$

Test statistic:  $Z = \frac{P_1 - P_2 - \delta_0}{\sqrt{\frac{P_1 \cdot (1 - P_1)}{n_1} + \frac{P_2 \cdot (1 - P_2)}{n_2}}}$

Test distribution: Standard normal distribution (CLT)

$$\delta = \delta_0 = 0 \text{ (Homogeneity hypothesis)}$$

We have  $P = \frac{n_1 \cdot P_1 + n_2 \cdot P_2}{n_1 + n_2}$ .

Test statistic:  $Z = \frac{P_1 - P_2}{\sqrt{P \cdot (1 - P) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Test distribution: Standard normal distribution (CLT)

## Example 9.18: see Ex. 9.4 (1)

In case of non-working mothers there were  $n_1 \cdot p_1 = 60$  children ready for school,  $n_2 \cdot p_2 = 24$  for working mothers. The null hypothesis that the share of children ready for school is 5% higher for non-working mothers is to be tested ( $\alpha = 0.05$ ).

Data input in R:

```
n1 <- 240
n2 <- 160
p_X1 <- 60 / n1
p_X2 <- 24 / n2
delta0 <- 0.05
alpha <- 0.05
```

## Example 9.18: see Ex. 9.4 (2)

Checking approximation conditions in R:

```
(n1 * p_X1 * (1 - p_X1)) > 9 & (n2 * p_X2 * (1 - p_X2)) > 9
[1] TRUE
0.1 <= p_X1 & p_X1 <= 0.9
[1] TRUE
[1] 1 <= p_X2 & p_X2 <= 0.9
```

Test decision in R:

```
c_stat <- vector()
c_stat[1] <- qnorm(p = alpha/2)
c_stat[2] <- qnorm(p = 1 - alpha/2)
Teststat <- (p_X1-p_X2-delta0) / (sqrt((p_X1 * (1-p_X1)) / n1
+ (p_X2 * (1-p_X2)) / n2))
c_stat
[1] -1.959964  1.959964
Teststat
[1] 1.258634
Teststat < c_stat[1] | Teststat > c_stat[2]
[1] FALSE
```

## Example 9.18: see Ex. 9.4 (3)

Now we would like to test the homogeneity hypothesis. The conditions remain the same.

We get  $p = 84/400 = 0.21$  and therefore:

$$z = \frac{0.25 - 0.15 - 0}{\sqrt{0.21 \cdot 0.79 \cdot \left(\frac{1}{240} + \frac{1}{160}\right)}} = 2.4055.$$

As the approximation conditions are met,  $H_0$  is rejected.

Hypothesis test in R:

```
p <- (n1 * p_X1 + n2 * p_X2) / (n1 + n2)
Teststat_new <- (p_X1-p_X2) / sqrt(p*(1-p)*(1/n1+1/n2))
```

p

[1] 0.21

Teststat\_new

[1] 2.405539

```
Teststat_new < c_stat[1] | Teststat_new > c_stat[2]
```

[1] TRUE

## Comparison of variances of two independent samples

We want to test  $H_0 : \sigma_1^2 = \sigma_2^2$ . Then,  $\frac{n_1 - 1}{\sigma_1^2} \cdot S_1^2 \sim \chi_{n_1-1}^2$  and  $\frac{n_2 - 1}{\sigma_2^2} \cdot S_2^2 \sim \chi_{n_2-1}^2$  are stochastically independent. Given  $H_0$  we have

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1; n_2 - 1) .$$

Notice that  $f(p; k_1; k_2) = \frac{1}{f(1-p; k_2; k_1)}$ .

## Example 9.19: Fill quantities (1)

A producer of lemonades wants to buy a new bottling plant and has to choose one of two alternatives. A test revealed no difference in the average fill quantities but an employee thinks that he observed less variation in fill quantities for the first machine. Another test of 101 fillings per plant yielded the following results:  $\bar{x} = 0.99971$ ,  $s_x = 0.01039$ ,  $\bar{y} = 0.99853$  and  $s_y = 0.02078$ . Let the fill quantities be normally distributed for both machines.

We use the null  $H_0 : \sigma_1^2 \geq \sigma_2^2$ . Then,  $H_0$  has to be rejected if the test statistic is too small. We have:

$$\frac{s_x^2}{s_y^2} = 0.24966 < f(0.05; 100; 100) = \frac{1}{f(0.95; 100; 100)} = 0.7185 \quad .$$

$H_0$  is rejected.

## Example 9.19: Fill quantities (2)

Hypothesis test in R:

```
load("Beispiel11-19.RData")  
  
var.test(x = x9_19,y = y9_19, ratio = 1, alternative="less",  
         conf.level = 0.95)
```

F test to compare two variances

```
data: x9_19 and y9_19  
F = 0.24966, num df = 100, denom df = 100, p-value = 1.275e-11  
alternative hypothesis: true ratio of variances is less than 1  
95 percent confidence interval:  
 0.0000000 0.3474606  
sample estimates:  
ratio of variances  
 0.2496628
```

# $\chi^2$ tests

The  $\chi^2$  tests build a class of tests comprising

- ▶ hypotheses regarding the distribution,
- ▶ hypotheses regarding independence,
- ▶ homogeneity of two or more distributions

and many other variations from which we abstract here.

$X$  may be nominal, ordinal or metric. We are looking at a countable number of categories, values or classes  $1, \dots, m$ .

We postulate a hypothesis regarding the distribution over the  $m$  categories, values or classes.  $\theta_j^0$  ( $j = 1, \dots, m$ ) are the corresponding shares. A null hypothesis regarding the distribution of  $X$  is transformed into a postulation regarding the  $\theta_j^0$ .

## Pearson's $\chi^2$ test

$$H_0 : \theta_1 = \theta_1^0 \wedge \cdots \wedge \theta_m = \theta_m^0 \text{ vs. } H_1 : \theta_1 \neq \theta_1^0 \wedge \cdots \wedge \theta_m \neq \theta_m^0$$

Test statistic:  $\chi^2 = \sum_{j=1}^m \frac{(n \cdot p_j - n \cdot \theta_j^0)^2}{n \cdot \theta_j^0} = \sum_{j=1}^m \frac{(n_j - n \cdot \theta_j^0)^2}{n \cdot \theta_j^0}$

If  $H_0$  holds,  $\chi^2$  is asymptotically  $\chi^2_{m-1}$ -distributed for  $n \rightarrow \infty$ ; therefore,  $n$  should be *large*.

Approximation conditions:

$m = 2$ :  $n \geq 30 \wedge n \cdot \theta_j^0 \geq 5$  for all  $j$

$m > 2$ :  $n \geq 30 \wedge n \cdot \theta_j^0 \geq 1$  for all  $j$  and  $n \cdot \theta_j^0 < 5$  for not more than 20% of categories, values or classes

The critical region is always one-sided:

$$C = \{(x_1, \dots, x_n) \mid \chi^2 > \chi^2_{m-1}(1 - \alpha)\}$$

- ▶ The quality of the test is determined by the number and localisation of the classes.
- ▶ A coarsening of the classification using a combination of classes might be needed.

## Example 9.20: Testing a dice (1)

A dice is suspicious and therefore suspected to be *crooked*. To test this hypothesis, the dice is rolled  $n = 300$  times. The recorded outcomes are  $n_1 = 40; n_2 = 45; n_3 = 80; n_4 = 55; n_5 = 45; n_6 = 35$ . The null that the dice *is not crooked* is to be tested using  $\alpha = 0.05$ .

- ▶  $H_0 : \theta_j = \theta_j^0 = \frac{1}{6}$  vs.  $H_1 : \theta_j \neq \theta_j^0 = \frac{1}{6}, \alpha = 0.05, n = 300$

Data input in R:

```
Sp9_20 <- c(40, 45, 80, 55, 45, 35)
n <- 300
m <- 6
theta0 <- rep(x = 1/6, times = m)
alpha <- 0.05
```

Checking approximation conditions in R:

$m > 2$

```
[1] TRUE
all(n * theta0 >= 1)
```

```
[1] TRUE
```

$n \geq 30$

```
[1] TRUE
sum(n * theta0 < 5) / m <= 0.2
```

```
[1] TRUE
```

## Example 9.20: Testing a dice (2)

- ▶  $\chi^2 = \sum_{j=1}^m \frac{(n_j - n \cdot \theta_j^0)^2}{n \cdot \theta_j^0} ; \quad \chi_5^2(0.95) = 11.070$
  - ▶  $C = \{(x_1, \dots, x_n) \mid \chi^2 > 11.0705\}$
  - ▶  $\chi^2 = \frac{(40 - 50)^2}{50} + \frac{(45 - 50)^2}{50} + \frac{(80 - 50)^2}{50} + \frac{(55 - 50)^2}{50} + \frac{(45 - 50)^2}{50} + \frac{(35 - 50)^2}{50} = 26$
- As  $26 \in C$  ( $26 > 11.0705$ ),  $H_0$  is rejected.

## Example 9.20: Testing a dice (3)

Test decision in R:

```
c_stat <- qchisq(p = 1-alpha, df = m-1)
```

```
[1] 11.0705
```

```
chisq.test(x = Sp9_20, p = theta0 * n, rescale.p = TRUE,  
           correct = FALSE) # ATTENTION: argument rescale.p
```

Chi-squared test for given probabilities

```
data: Sp9_20  
X-squared = 26, df = 5, p-value = 8.924e-05
```

## Example 9.21: Test for standard normal distribution (1)

The frequency distribution of a variable  $y$  in the sample looks as follows ( $n = 1000$ ):

Class	$y \leq -1$	$-1 < y \leq 0$	$0 < y \leq 1$	$y > 1$
Frequency	165	360	314	161

Test the null that *variable y follows a standard normal distribution* using  $\alpha = 0.05$ .

- ▶ First, the theoretical frequencies  $n \cdot \theta_j^0$  have to be determined.

$1000 \cdot \theta_1^0 = 1000 \cdot \Phi(-1) = 1000 \cdot (1 - 0.8413) = 159$ . As the SND is symmetric, we have  $\theta_4^0 = 159$  and from  $\theta_2^0 = \theta_3^0$  finally follows that  $n \cdot \theta_2^0 = n \cdot \theta_3^0 = 341$ .

$H_0 : \theta_1 = \theta_1^0 = 0.159 \wedge \theta_2 = \theta_2^0 = 0.341 \wedge \theta_3 = \theta_3^0 = 0.341 \wedge \theta_4 = \theta_4^0 = 0.159$ ,  $H_1$  complementary,  $\alpha = 0.05$  and  $n = 1000$

## Example 9.21: Test for standard normal distribution (2)

Data input in R:

```
load("Example9-21.RData")  
  
y9_21_kl <- table(y9_21)  
n <- 1000  
m <- 4  
theta0 <- c(159, 341, 341, 159) / n  
alpha <- 0.05
```

Checking approximation conditions in R:

```
m > 2           n >= 30
```

```
[1] TRUE          [1] TRUE
```

```
all(n * theta0 >= 1)      sum(n * theta0 < 5) / m <= 0.2
```

```
[1] TRUE          [1] TRUE
```

## Example 9.21: Test for standard normal distribution (3)

- ▶  $\chi^2 = \sum_{j=1}^m \frac{(n_j - n \cdot \theta_j^0)^2}{n \cdot \theta_j^0} ; \quad \chi_3^2(0.95) = 7.815$
- ▶  $C = \{(x_1, \dots, x_n) \mid \chi^2 > 7.815\}$
- ▶  $\chi^2 = \frac{(165 - 159)^2}{159} + \frac{(360 - 341)^2}{341} + \frac{(314 - 341)^2}{341} + \frac{(161 - 159)^2}{159} = 3.448$

As  $3.448 \notin C$  ( $3.448 \geq 7.815$ ),  $H_0$  is not rejected.

Hypothesis test in R:

```
chisq.test(x = y9_21_kl, p = theta0 * n, rescale.p = TRUE,
            correct = FALSE)
```

Chi-squared test for given probabilities

```
data: y9_21_kl
X-squared = 3.4481, df = 3, p-value = 0.3276
```

## Example 9.21: Test for standard normal distribution (4)

In this case the hypothesis has been fully specified as the distribution with all its parameters has been stated (SND). If we still need to estimate  $\omega$  parameters from the sample, we need to use the following test distribution:  $\chi^2_{m-\omega-1}$ .

In the present case we might also have tested for a normal distribution.  $\mu$  and  $\sigma^2$  would have to be estimated using the known estimators  $\bar{X}$  and  $S^2$  using sample data. The theoretical frequencies would then have to be calculated again. The test distribution would then be  $\chi^2_1$ .

Please test again, this time using  $\bar{x} = -0.1$  and  $s^2 = 1.047$ .  
(Hint:  $H_0$  is rejected)

# $\chi^2$ test for independence

The variables of interest have  $m$  and  $r$  values, categories or classes, respectively.  $n_{jk}$  is the joint frequency and  $\theta_{jk}$  is the joint probability with the corresponding marginal probabilities  $\theta_{j\cdot}$  and  $\theta_{\cdot k}$ .  
 We use

$$H_0 : \theta_{jk} = \theta_{j\cdot} \cdot \theta_{\cdot k} \quad \text{for all } j = 1, \dots, m \text{ and } k = 1, \dots, r$$

as the null. The theoretical frequencies on the RHS of  $H_0$  are built using  $\theta_{j\cdot} = n_{j\cdot}/n$  and  $\theta_{\cdot k} = n_{\cdot k}/n$ , respectively.

Then

$$\chi^2 = \sum_{j=1}^m \sum_{k=1}^r \frac{(n_{jk} - n \cdot \theta_{j\cdot} \cdot \theta_{\cdot k})^2}{n \cdot \theta_{j\cdot} \cdot \theta_{\cdot k}} = \sum_{j=1}^m \sum_{k=1}^r \frac{\left(n_{jk} - \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot} \cdot n_{\cdot k}}{n}}$$

is the test statistic. As the test distribution we use  $\chi^2_{(m-1) \cdot (r-1)}$ .

# Two-dimensional frequencies

Cat. of 1st variable \ Cat. of 2nd variable	1	$\dots$	$k$	$\dots$	$r$	$\sum$
1	$n_{11}$	$\dots$	$n_{1k}$	$\dots$	$n_{1r}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
$j$	$n_{j1}$	$\dots$	$n_{jk}$	$\dots$	$n_{jr}$	$n_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
$m$	$n_{m1}$	$\dots$	$n_{mk}$	$\dots$	$n_{mr}$	$n_{m\cdot}$
$\sum$	$n_{\cdot 1}$	$\dots$	$n_{\cdot k}$	$\dots$	$n_{\cdot r}$	$n$

## Example 9.22: exam (see Schaich, 1998) (1)

For  $n = 627$  graduated students at a faculty of a university, the number of semesters and the final grade have been recorded (see table below). We want to test for independence of number of semesters and final grade using  $\alpha = 0.05$ .

Load data in R:

```
load("Example9-22.RData")  
  
alpha <- 0.05  
n <- sum(Sp9_22)  
m <- dim(Sp9_22)[1]  
r <- dim(Sp9_22)[2]  
  
theta0 <- margin.table(x = Sp9_22, margin = 1) %*%  
          t(margin.table(x = Sp9_22, margin = 2)) /  
          margin.table(x = Sp9_22)
```

## Example 9.22: exam (see Schaich, 1998) (2)

Semesters	8	9	10	11	12-13	14-15	over 15	$\Sigma$
Grade								
1	9 (3.1)	19 (10.5)	23 (23.0)	5 (10.8)	4 (8.2)	3 (6.0)	4 (5.4)	67
2	10 (7.3)	34 (24.7)	68 (54.2)	11 (25.5)	11 (19.4)	21 (14.1)	3 (12.9)	158
3	8 (8.6)	34 (29.2)	67 (64.1)	32 (30.1)	20 (23.0)	13 (16.7)	13 (15.2)	187
4	0 (4.9)	10 (16.7)	30 (36.7)	28 (17.2)	18 (13.1)	6 (9.6)	15 (8.7)	107
5	2 (5.0)	1 (16.9)	27 (37.0)	25 (17.4)	24 (13.3)	13 (9.6)	16 (8.8)	108
$\Sigma$	29	98	215	101	77	56	51	627

Empirical frequencies (theoretical frequencies under  $H_0$ )

## Example 9.22: exam (see Schaich, 1998) (3)

Checking approximation conditions in R:

`n >= 30`

[1] TRUE

```
sum(theta0 == 0) == 0
```

[1] TRUE

```
sum(theta0 < 5) / (m*r) <= 0.2
```

[1] TRUE

►  $H_0 : \theta_{jk} = \theta_{j\cdot} \cdot \theta_{\cdot k}$  vs.  $H_1 : \theta_{jk} \neq \theta_{j\cdot} \cdot \theta_{\cdot k}$  for all

$j = 1, \dots, m$  and  $k = 1, \dots, r$ ,  $\alpha = 0.05$ ,  $n = 627$

►  $\chi^2 = \sum_{j=1}^m \sum_{k=1}^r \frac{\left( n_{jk} - \frac{n_{j\cdot} \cdot n_{\cdot k}}{n} \right)^2}{n_{j\cdot} \cdot n_{\cdot k}}$ ,  $\chi^2_{24}(0.95) = 36.415$ ,

►  $C = \left\{ (x_1, \dots, x_n) \mid \chi^2 > 36.415 \right\}$

►  $\chi^2 = \frac{(9 - 3.1)^2}{3.1} + \dots + \frac{(16 - 8.8)^2}{8.8} = 120.54$

As  $120.54 \in C$  ( $120.54 > 36.415$ ),  $H_0$  is rejected.

## Example 9.22: exam (see Schaich, 1998) (4)

Test decision in R:

```
c_stat <- qchisq(p = 1 - alpha, df = (m-1) * (r-1))  
c_stat
```

```
[1] 36.41503
```

```
chisq.test(x = Sp9_22, correct = FALSE, rescale.p = FALSE)
```

Pearson's Chi-squared test

```
data: Sp9_22  
X-squared = 120.55, df = 24, p-value = 7.741e-15
```

Warning message:

```
In chisq.test(x = Sp9_22, correct = FALSE) :  
Chi-squared approximation may be incorrect
```

## $\chi^2$ test for homogeneity

We want to test  $m$  distributions for homogeneity. We look at  $m$  distributions with  $r$  categories, values or classes each. The distributions have to be considered homogeneous if the relative frequencies are identical and do not depend on the respective distribution. Therefore, the test can immediately be derived from the  $\chi^2$  test for independence:

$$\text{Test statistic: } \chi^2 = \sum_{j=1}^m \sum_{k=1}^r \frac{\left( h_{jk} - \frac{n_j \cdot h_{\cdot k}}{n} \right)^2}{\frac{n_j \cdot h_{\cdot k}}{n}}.$$

Here,  $h_{jk}$  are the observed frequencies in the  $j$ -th distribution and the  $k$ -th category, value or class.  $n_j$  is the absolute frequency of observations of distribution  $j$ .

Test distribution:  $\chi^2_{(m-1) \cdot (r-1)}$

## Example 9.23: Income distributions (1)

In a study on income distributions (poor, middle, rich) in three regions A, B and C interviews were conducted with  $n = 200$  people in each region:

Region	poor	middle	rich	$\sum$
A	51	112	37	200
B	65	121	14	200
C	55	115	30	200
$\sum$	171	348	81	600

Test if the income distributions are homogeneous in the three regions using  $\alpha = 0.05$ .

Read data in R:

```
load("Example9-23.RData")
alpha <- 0.05
n<-sum(Sp9_23);m<-dim(Sp9_23)[1];r<-dim(Sp9_23)[2]
theta0 <- margin.table(x = Sp9_23, margin = 1) %*%
  t(margin.table(x = Sp9_23, margin = 2)) /
  sum(x = Sp9_23)
```

## Example 9.23: Income distributions (2)

Checking approximation conditions in R:

```
n >= 30  
[1] TRUE  
  
sum(theta0 == 0) == 0  
[1] TRUE  
  
sum(theta0 < 5) / (m*r) <= 0.2  
[1] TRUE
```

- ▶  $H_0 : F_A = F_B = F_C, \alpha = 0.05, n = 600$
  - ▶ Test statistic: see above, approximation conditions fulfilled
  - ▶  $C = \{(x_1, \dots, x_n) \mid \chi^2 > 9.488\}$
  - ▶  $\chi^2 = \frac{(51 - 200 \cdot 171/600)^2}{200 \cdot 171/600} + \dots + \frac{(30 - 200 \cdot 81/600)^2}{200 \cdot 81/600} = 12.483$
- As  $12.483 \in C$  ( $12.483 > 9.488$ ),  $H_0$  is rejected.

## Example 9.23: Income distributions (3)

Hypothesis test in R:

```
chisq.test(x = Sp9_23, correct = FALSE, rescale.p = FALSE)
  Pearson's Chi-squared test

data: Sp9_23
X-squared = 12.483, df = 4, p-value = 0.0141
```

Alternative determination of the theoretical frequencies under  $H_0$  and of the critical value in R:

```
c_stat <- qchisq(p = 1 - alpha, df = (m-1) * (r-1))
theta0_alternative <- chisq.test(x = Sp9_23, correct = FALSE,
                                   rescale.p = FALSE)$expected
```

c_stat	theta0_alternative		
	poor	middle	rich
[1] 9.487729			
A	57	116	27
B	57	116	27
C	57	116	27

## More tests (1)

**Tests for distribution** In praxis, the Kolmogorow-Smirnow test (KS test) is frequently used. Here, one measures the maximum distance between the theoretical distribution function and the relative empirical sum function in order to use it as test statistic.

**Tests for correlation** The correlation coefficient of Bravais-Pearson has in general no good properties regarding inference. An exception is the bivariate normal distribution! In this case, two tests can be used ( $R_{xy} = \hat{\rho}_{XY}$  is the emp. corr.):

►  $H_0 : \rho_{XY} = 0$

$$\frac{R_{xy}}{\sqrt{1 - R_{xy}^2}} \cdot \sqrt{n - 2} \sim t_{n-2}$$

Since stochastic independence and uncorrelatedness are equal in the case of a bivariate normal distribution, this test is sometimes also called test for independence.

## More tests (2)

Tests for correlation ►  $H_0 : \rho_{XY} = \rho_0 \ (\rho_0 \neq 0), n > 25$

$$\frac{1}{2} \left( \ln \frac{1 + R_{xy}}{1 - R_{xy}} - \ln \frac{1 + \rho_0}{1 - \rho_0} \right) \cdot \sqrt{n - 3} \sim N(0; 1)$$

Tests for correlation in R:

```
cor.test(x, y,
          alternative = c("two.sided", "less", "greater"),
          method = c("pearson", "kendall", "spearman"),
          conf.level = 0.95)
```

Variance analysis Test for homogeneity of means of at least two distributions. The test is based on a decomposition of the total variance within and between the single distributions. As a test distribution, one uses the  $F$  distribution.

# Elements of Statistics

## Chapter 10: Regression analysis

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier  
Faculty IV  
Economic and Social Statistics Department

Winter semester 2022/23

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 22/23 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

# Aim of regression analysis

Let  $(x_i, y_i)$  be pairs of observations from a survey. At first, the attributes can be analysed individually. Furthermore the reciprocal relations of the variables can be examined.

**Symmetric relation:** Is there a statistical correlation between the variables?

- ▶ Contingency coefficient
- ▶ Rank correlation coefficient
- ▶ Bravais-Pearson correlation coefficient

**Asymmetric relation:** Does one variable influence the other, e.g. is there a functional correlation?

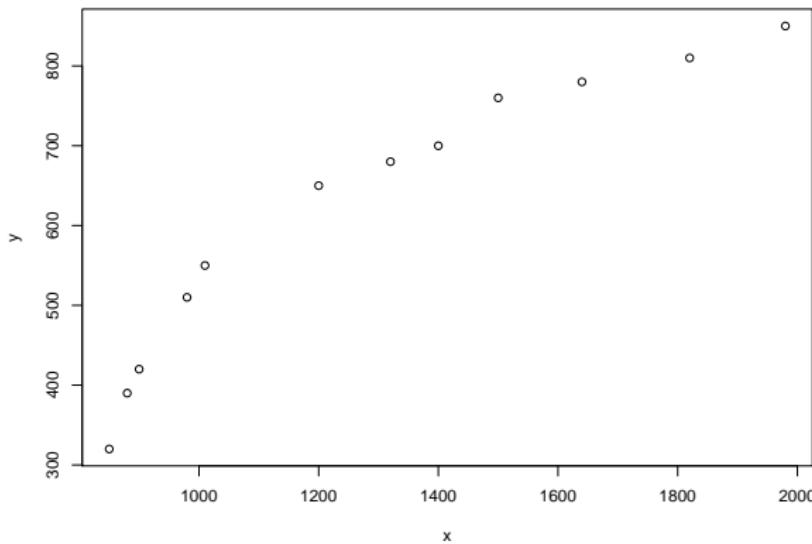
$y$  is called the response variable (regressand / explained variable).

$x$  is called the explanatory variable (regressor).

## Example 10.1: Income and consumption (1)

In a survey for habitudes of income and consumption,  $n = 12$  households were sampled (see HABE in Switzerland).  $x$  is the net income of the household and  $y$  are the expenditures for nutrition.

$x_i$	$y_i$
1010	550
1200	650
1400	700
980	510
850	320
1640	780
1320	680
880	390
1500	760
900	420
1980	850
1820	810



The income level seems to have an influence on consumption, but there seems to be a saturation as well.

## Example 10.1: Income and consumption (2)

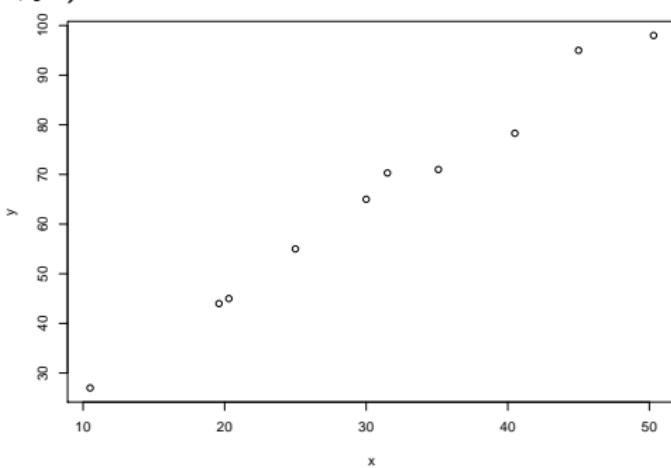
Data input and scatter plot in R:

```
x10_1 <- c(1010, 1200, 1400, 980, 850, 1640,  
          1320, 880, 1500, 900, 1980, 1820)  
  
y10_1 <- c(550, 650, 700, 510, 320, 780,  
          680, 390, 760, 420, 850, 810)  
  
plot(x = x10_1, y = y10_1)
```

## Example 10.2: School readiness test (1)

In order to judge the school readiness of kindergartners, two different tests were used. A SRS of size  $n = 10$  was taken. The values for the points are given through  $(x_i, y_i)$ .

$x_i$	$y_i$
20.3	45.0
31.5	70.3
40.5	78.3
19.6	44.0
25.0	55.0
10.5	27.0
30.0	65.0
35.1	71.0
50.3	98.0
45.0	95.0



We have a strong linear correlation. A functional relation in the sense of independent and dependent variable is however not necessarily plausible.

## Example 10.2: School readiness test (2)

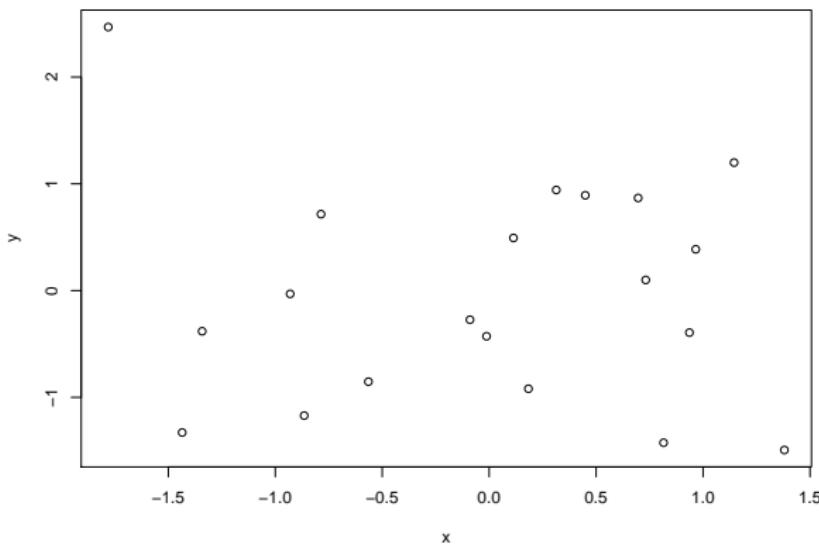
Data input and scatter plot in R:

```
x10_2 <- c(20.3, 31.5, 40.5, 19.6, 25.0,  
          10.5, 30.0, 35.1, 50.3, 45.0)  
  
y10_2 <- c(45.0, 70.3, 78.3, 44.0, 55.0,  
          27.0, 65.0, 71.0, 98.0, 95.0)  
  
plot(x = x10_2, y = y10_2)
```

## Example 10.3: Two random variables (1)

Let the random variables  $X$  and  $Y$  be standard normally distributed and stochastically independent. A two-dimensional random sample of size  $n = 20$  resulted in:

$x_i$	$y_i$
0.815	-1.425
-1.342	-0.381
1.380	-1.493
-1.435	-1.329
0.936	-0.393
0.450	0.893
-0.786	0.716
-0.090	-0.273
0.314	0.943
-0.930	-0.031
-0.012	-0.428
1.145	1.199
0.184	-0.919
0.732	0.100
-1.781	2.469
0.114	0.493
0.966	0.387
-0.865	-1.171
-0.564	-0.853
0.696	0.868



There is no correlation and thus no relation between the two variables.

## Example 10.3: Two random variables (2)

Data input and scatter plot in R:

```
x10_3 <- c(0.815, -1.342, 1.38, -1.435, 0.936,  
          0.45, -0.786, -0.09, 0.314, -0.93,  
         -0.012, 1.145, 0.184, 0.732, -1.781,  
         0.114, 0.966, -0.865, -0.564, 0.696)  
  
y10_3 <- c(-1.425, -0.381, -1.493, -1.329, -0.393,  
          0.893, 0.716, -0.273, 0.943, -0.031,  
         -0.428, 1.199, -0.919, 0.1, 2.469,  
         0.493, 0.387, -1.171, -0.853, 0.868)  
  
plot(x = x10_3, y = y10_3)
```

# Problems for regression ideas

- ▶ Quality and quantity of the involved variables
  - MZ: salary  $\sim$  age, gender, education, experience
  - Rent index: rent  $\sim$  living area, neighbourhood, ...
- ▶ Type of correlation
  - ▶ Linear
  - ▶ Polynomial
  - ▶ Exponential
- ▶ Sample size (and sampling design)
- ▶ Solution in the sense of inferential statistics
  - ▶ **Estimation** of the parameters of interest
  - ▶ Checking of parameters
    - (Example: growth or stagnation)

# Simple linear regression model

We assume the following linear model:

$$Y = \alpha + \beta \cdot X + \varepsilon$$

with error term  $\varepsilon$ . Hence, there might be more than one value of  $y$  corresponding to any given value of  $x$  (**random error**).

Using the data at hand, we would like to determine two parameters: the **intercept**  $a$  and the **slope**  $b$  of

$$\hat{y}_i = a + b \cdot x_i ,$$

where  $\hat{y}_i$  is the vertical projection of observation  $y_i$  onto the **regression line**. Let  $e_i = y_i - \hat{y}_i$  be the **residual** corresponding to observation  $x_i$ .

Using the method of **ordinary least squares (OLS)**, we can reach estimates for the parameters  $a$  and  $b$ :

$$Z(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 \rightarrow \min .$$

## Solution to the minimisation problem

Using the two first order conditions, we get

$$n \cdot a + b \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (\text{I})$$

$$a \cdot \sum_{i=1}^n x_i + b \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i . \quad (\text{II})$$

Solving for  $a$  in (I) and then substituting into (II), we get

$$b = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}^*}{s_x^{*2}}$$

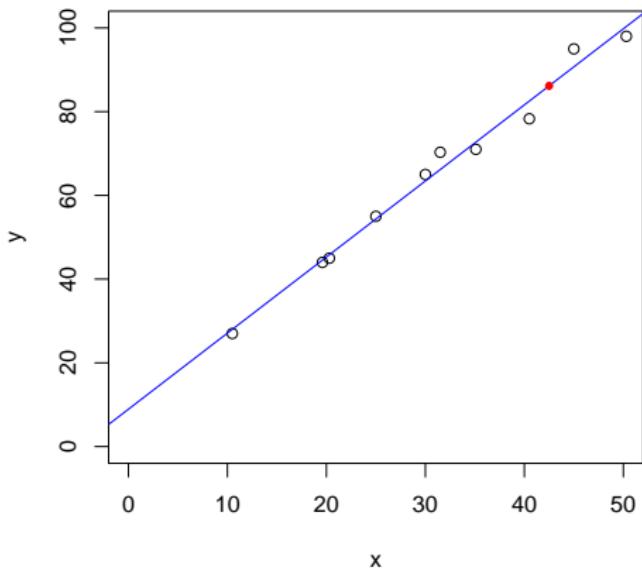
and  $a = \bar{y} - b \cdot \bar{x}$ . Finally, for the sample regression line we have:

$$\hat{y} = \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot (x - \bar{x}) .$$

## Example 10.4: see Ex. 10.2 (1)

We want to obtain estimates for the parameters  $\alpha$  and  $\beta$  using the OLS method. Furthermore, we want to find a suitable estimation for the second school readiness criterion, if  $x_0 = 42,5$  was observed for the first school readiness criterion.

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
20.3	45.0	412.09	2025.00	913.50
31.5	70.3	992.25	4942.09	2214.45
40.5	78.3	1640.25	6130.89	3171.15
19.6	44.0	384.16	1936.00	862.40
25.0	55.0	625.00	3025.00	1375.00
10.5	27.0	110.25	729.00	283.50
30.0	65.0	900.00	4225.00	1950.00
35.1	71.0	1232.01	5041.00	2492.10
50.3	98.0	2530.09	9604.00	4929.40
45.0	95.0	2025.00	9025.00	4275.00
307.8	648.6	10851.10	46682.98	22466.50



## Example 10.4: see Ex. 10.2 (2)

From the table above we get:

$$\bar{x} = \frac{1}{10} \cdot 307.8 = 30.78 \quad \text{sowie} \quad \bar{y} = 64.86 \quad .$$

$\bar{x}$  and  $\bar{y}$  in R:

```
SpMean_x <- mean(x10_2); SpMean_y <- mean(y10_2)
```

```
SpMean_x
```

```
[1] 30.78
```

```
SpMean_y
```

```
[1] 64.86
```

With this, we get:

$$b = \frac{22466.50 - 10 \cdot 30.78 \cdot 64.86}{10851.10 - 10 \cdot 30.78^2} = 1.817$$
$$a = 64.86 - 1.817 \cdot 30.78 = 8.933 \quad .$$

## Example 10.4: see Ex. 10.2 (3)

Regression analysis in R:

```
reg_analysis <- lm(formula = y10_2 ~ x10_2)
summary(reg_analysis)
```

Call:

```
lm(formula = y10_2 ~ x10_2)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.2252	-1.5342	-0.6775	1.3293	4.2965

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.92036	2.56067	3.484	0.00828 **
x10_2	1.81740	0.07774	23.379	1.19e-08 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.885 on 8 degrees of freedom

Multiple R-squared: 0.9856, Adjusted R-squared: 0.9838

F-statistic: 546.6 on 1 and 8 DF, p-value: 1.191e-08

## Example 10.4: see Ex. 10.2 (4)

Obtaining  $b$  and  $a$  in R:

```
b <- summary(reg_analysis)$coeff[2,1]
a <- summary(reg_analysis)$coeff[1,1]
```

b

```
[1] 1.817402
```

a

```
[1] 8.920358
```

Furthermore we have

$$r = \frac{22466.50 - 10 \cdot 30.78 \cdot 64.86}{\sqrt{(10851.10 - 10 \cdot 30.78^2) \cdot (46682.98 - 10 \cdot 64.86^2)}} = 0.9927 \quad .$$

Determination of  $r$  in R:

```
r <- sqrt(summary(reg_analysis)$r.squared)
```

r

```
[1] 0.9927614
```

## Example 10.4: see Ex. 10.2 (5)

As estimation for the second school readiness criterion we get:

$$\hat{y}_0 = 8.933 + 1.817 \cdot 42.5 = 86.159 \quad .$$

Construction of the estimation value  $\hat{y}_0$  in R:

```
x_0 <- 42.5
y_hat_0 <- predict(object = reg_analysis,
                      newdata = data.frame(x10_2 = x_0))
```

```
y_hat_0
```

```
1
86.15995
```

Graphic in R:

```
plot(x = x10_2, y = y10_2)
abline(reg = reg_analysis, col = "blue") # Point not shown
```

## Further issues

- ▶ From (I) follows:

$$\sum_{i=1}^n (y_i - a - b \cdot x_i) = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n e_i = 0$$

and thus  $\bar{y} = \hat{\bar{y}}$ .

- ▶ We have:

$$s_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Variation of } y$$

$$s_{\hat{y}}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{Variation of } \hat{y}$$

$$s_e'^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n e_i^2 \quad \text{Variation of residuals}$$

It is:  $s_y^2 = s_{\hat{y}}^2 + s_e'^2$  and  $1 = \frac{s_{\hat{y}}^2}{s_y^2} + \frac{s_e'^2}{s_y^2}$

# Coefficient of determination

The ratio  $r_{xy}^2 = \frac{s_y^2}{s_x^2} = 1 - \frac{s_e'^2}{s_y^2}$  is called coefficient of determination. It is equal to the squared Bravais-Pearson correlation coefficient. Of special interest are the exceptional cases  $r_{xy}^2 = 0$  and  $r_{xy}^2 = 1$ . It is

$$\begin{aligned}\frac{s_y^2}{s_x^2} &= \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{s_y^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (a + bx_i - (a + b\bar{x}))^2}{s_y^2} \\ &= b^2 \cdot \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{s_y^2} = \left(\frac{s_{xy}}{s_x}\right)^2 \cdot \frac{s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = r_{xy}^2 \quad \square\end{aligned}$$

## Example 10.4: see Ex. 10.2 (6)

As a value for the coefficient of determination, we get

$r_{xy}^2 = 0,9927^2 = 0.9855$ , e.g. 98.55% of the variance of the target variable are explained through the variance of the exogeneous variable.

The coefficient of determination  $r_{xy}^2$  in R:

```
r_q <- summary(reg_analysis)$r.squared
```

```
r_q
```

```
[1] 0.9855751
```

# Inferential statistical properties of OLS

Regression line of the universe:  $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$

Regression line of the sample:  $y_i = a + b \cdot x_i + e_i$

Since we are drawing a random sample,  $A$  and  $B$  are random variables as estimators for  $\alpha$  and  $\beta$ .

System of assumptions:

1. The error terms have the expected value 0:

$$E(\varepsilon_i) = 0.$$

2. The error terms have a constant variance (homoskedasticity)

$$\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2.$$

3. The error terms are uncorrelated

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{for all } i \neq j.$$

4. Normal distribution assumption

$$\varepsilon_i \sim N(0; \sigma_\varepsilon^2)$$

## Statements to the OLS regression line

Following from the assumptions 1. – 3. we have:

- ▶  $A$  and  $B$  are best linear unbiased estimators for  $\alpha$  and  $\beta$ .
- ▶ The estimator  $S_e^2 = \frac{1}{n-2} \sum_{i=1}^n E_i^2$  is unbiased for  $\sigma_\varepsilon^2$ .
- ▶  $\hat{Y} = A + B \cdot x_0$  is the best linear unbiased estimator for the value of the target variable corresponding to  $x_0$ .

If additionally assumption 4 also holds, we have:

- ▶  $A$  and  $B$  are ML-estimators for  $\alpha$  and  $\beta$ .
- ▶ The estimator  $S_e^{*2} = \frac{1}{n} \sum_{i=1}^n E_i^2$  is an ML-estimator for  $\sigma_\varepsilon^2$ .
- ▶  $\hat{Y} = A + B \cdot x_0$  is an ML-estimator for the value of the target variable corresponding to  $x_0$ .

## Point and interval estimation

Let  $E(A) = \alpha$  and  $E(B) = \beta$  as well as  $E(S_e^2) = \sigma_\varepsilon^2$ . Furthermore, we have:

$$A \sim N\left(\alpha; \sigma_\varepsilon^2 \cdot \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$B \sim N\left(\beta; \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\frac{(n-2) \cdot S_e^2}{\sigma_\varepsilon^2} = \frac{\sum_{i=1}^n E_i^2}{\sigma_\varepsilon^2} \sim \chi_{n-2}^2$$

$A$  and  $\sum_{i=1}^n E_i^2 / \sigma_\varepsilon^2$  resp.  $B$  and  $\sum_{i=1}^n E_i^2 / \sigma_\varepsilon^2$  are stochastically independent.

Hence, we obtain the following confidence intervals:

# Confidence intervals

$$\text{CI for } \alpha \left[ a \pm t(1 - \frac{\alpha}{2}; n - 2) \cdot \sqrt{s_e^2 \cdot \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}} \right]$$

$$\text{CI for } \beta \left[ b \pm t(1 - \frac{\alpha}{2}; n - 2) \cdot \sqrt{\frac{s_e^2}{\sum_i (x_i - \bar{x})^2}} \right]$$

$$\text{CI for } \sigma_e^2 \left[ \frac{(n - 2) \cdot s_e^2}{\chi_{n-2}^2(1 - \frac{\alpha}{2})}, \frac{(n - 2) \cdot s_e^2}{\chi_{n-2}^2(\frac{\alpha}{2})} \right]$$

## CI for the mean of an observation

$$\left[ \hat{y}_0 \pm t(1 - \frac{\alpha}{2}; n - 2) \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right]$$

## CI for the single value of an observation

$$\left[ \hat{y}_0 \pm t(1 - \frac{\alpha}{2}; n - 2) \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right]$$

## Example 10.5: see Ex. 10.4 (1)

We want to determine the confidence intervals for  $\alpha$ ,  $\beta$  and  $\sigma_{\epsilon}^2$  as well as for  $x'_0 = 35$  and  $x''_0 = 50$ .

At first, we have:

$$r^2 = 1 - \frac{s_e'^2}{s_y^2} \Leftrightarrow s_e'^2 = s_y^2 \cdot (1 - r^2) \Rightarrow s_e^2 = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r^2) .$$

Using the current results, we get:  $s_e^2 = 8.3210$ .

Standard deviations in R:

```
Std_a <- summary(reg_analysis)$coeff[1,2]
Std_b <- summary(reg_analysis)$coeff[2,2]
Std_e <- summary(reg_analysis)$sigma
```

`round(Std_a,4)`

[1] 2.5607

`round(Std_b,4)`

[1] 0.0777

`round(Std_e,4)`

[1] 2.8846

## Example 10.5: see Ex. 10.4 (2)

With the needed quantiles  $t(0.975; 8) = 2.306$ ,  $\chi^2(0.975; 8) = 17.535$  and  $\chi^2(0.025; 8) = 2.180$  we get:

$$\text{CI}_\alpha : [3,0154; 14,8253] ; \text{CI}_\beta : [1,6381; 1,9967] .$$

Determination of  $\text{CI}_\alpha$  und  $\text{CI}_\beta$  in R:

```
alpha <- 0.05

CI_a_and_b <- confint(object = reg_analysis,
                         level = 1 - alpha
                       )
CI_a_and_b
```

2.5 % 97.5 %

(Intercept) 3.015438 14.82528  
x10\_2 1.638145 1.99666

## Example 10.5: see Ex. 10.4 (3)

Additionally, we get:

$$\text{CI}_{\sigma_{\varepsilon}^2} : [3.7964, 30.5394] \quad .$$

Determination of  $\text{CI}_{\sigma_{\varepsilon}^2}$  in R:

```
n <- length(x10_2)
df <- summary(reg_analysis)$df[2]

CI_sigma_epsilon <- vector()
CI_sigma_epsilon[1] <- ((n - 2) * Std_e^2) /
                        qchisq(p = 1-(alpha/2), df = df)
CI_sigma_epsilon[2] <- ((n - 2) * Std_e^2) /
                        qchisq(p = (alpha/2), df = df)
CI_sigma_epsilon

[1] 3.79637    30.53938
```

## Example 10.5: see Ex. 10.4 (4)

Finally, we get:

CI	Mean	Single value
35	[70.2940,74.7648]	[65.5120,79.5469]
50	[95.7538,103.827]	[92.0095,107.571]

Determination of the confidence intervals above in R:

```
x_0 <- c(35, 50)

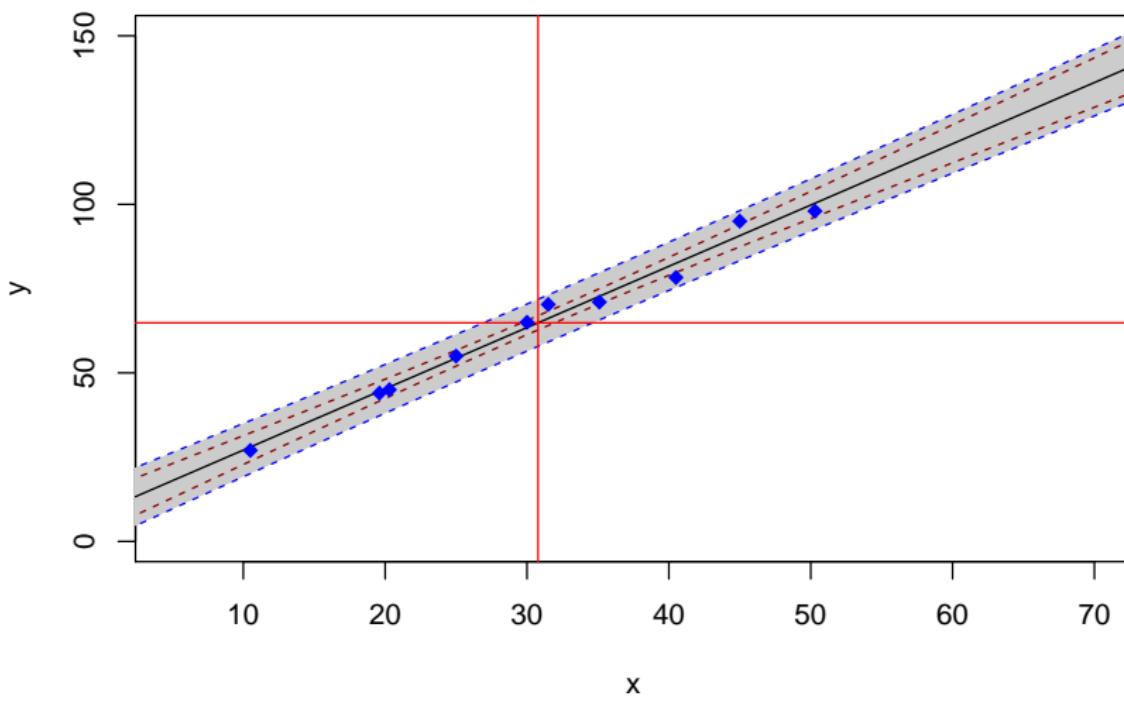
CI_Mean_Obs <- predict(reg_analysis,
                         newdata = data.frame(x10_2 = x_0),
                         interval = "confidence")

CI_Obs <- predict(reg_analysis,
                     newdata = data.frame(x10_2 = x_0),
                     interval = "prediction")
```

	CI_Mean_Obs			CI_Obs			
	fit	lwr	upr	fit	lwr	upr	
1	72.52944	70.29403	74.76484	1	72.52944	65.51196	79.54692
2	99.79047	95.75376	103.82719	2	99.79047	92.00953	107.57141

## Example 10.5: see Ex. 10.4 (5)

Confidence bands for means (red) and single values (blue)



# Hypothesis testing

- ▶  $H_0 : \alpha = \alpha_0$  versus  $H_1 : \alpha \neq \alpha_0$

Test statistic and test distribution:

$$\frac{A - \alpha_0}{S_e} \cdot \sqrt{\frac{n \sum_i (x_i - \bar{x})^2}{\sum_i x_i^2}} \sim t(n - 2)$$

- ▶  $H_0 : \beta = \beta_0$  versus  $H_1 : \beta \neq \beta_0$

Test statistic and test distribution:

$$\frac{B - \beta_0}{S_e} \cdot \sqrt{\sum_i (x_i - \bar{x})^2} \sim t(n - 2)$$

## Example 10.6: see Ex. 10.4 (1)

Repetition of the regression analysis in R:

```
summary(reg_analysis)
```

Call:

```
lm(formula = y10_2 ~ x10_2)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

-4.2252	-1.5342	-0.6775	1.3293	4.2965
---------	---------	---------	--------	--------

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.92036	2.56067	3.484	0.00828 **
x10_2	1.81740	0.07774	23.379	1.19e-08 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.885 on 8 degrees of freedom

Multiple R-squared: 0.9856, Adjusted R-squared: 0.9838

F-statistic: 546.6 on 1 and 8 DF, p-value: 1.191e-08

## Example 10.6: see Ex. 10.4 (2)

For the  $p$  value of the estimations of the coefficients, there are symbols of significance given for different values in the R-Output. We have:

$$\text{t value} = \text{Estimate}/\text{Std. Error} .$$

This is equal to the values of the test statistics for  $\alpha_0 = 0$  resp.  $\beta_0 = 0$ . We are especially interested in the corresponding null hypotheses, thus

- ▶  $H_0 : \alpha = \alpha_0 = 0$  and above all
- ▶  $H_0 : \beta = \beta_0 = 0$ .

We test for *significance* of the single parameters. At a significance level of 5%, we reject both null hypotheses, e.g. that the parameters  $\alpha$  resp.  $\beta$  are zero ( $p$  values are substantially smaller than 0.05). This means that the intercept and the exogenous variable  $X$  have a significant influence in the model.

From the output, we can obtain crucial parts of the beforehand calculated confidence intervals.

## Example 10.6: see Ex. 10.4 (3)

Hypothesis test regarding  $H_0 : \alpha = \alpha_0 = 0$  in R:

```
alpha <- 0.05

Teststat_a <- summary(reg_analysis)$coeff[1,3]
p_value_a <- summary(reg_analysis)$coeff[1,4]

c_stat_a <- vector()
c_stat_a[1] <- qt(p = alpha/2, df = df)
c_stat_a[2] <- qt(p = 1-alpha/2, df = df)
```

Teststat_a	c_stat_a[1]	c_stat_a[2]
[1] 3.483601	[1] -2.306004	[1] 2.306004

```
Teststat_a < c_stat_a[1] | Teststat_a > c_stat_a[2]
```

```
[1] TRUE
```

Alternative test decision in R:

```
p_value_a < alpha 0 < CI_a_and_b[1,1] | 0 > CI_a_and_b[1,2]
[1] TRUE [1] TRUE
```

## Example 10.6: see Ex. 10.4 (4)

Hypothesis test regarding  $H_0 : \beta = \beta_0 = 0$  in R:

```
Teststat_b <- summary(reg_analysis)$coeff[2,3]
p_value_b <- summary(reg_analysis)$coeff[2,4]

c_stat_b <- vector()
c_stat_b[1] <- qt(p = alpha/2, df = df)
c_stat_b[2] <- qt(p = 1-alpha/2, df = df)
```

Teststat_b	c_stat_b[1]	c_stat_b[2]
[1] 23.37944	[1] -2.306004	[1] 2.306004

```
Teststat_b < c_stat_b[1] | Teststat_b > c_stat_b[2]
```

```
[1] TRUE
```

Alternative test decision in R:

```
p_value_b < alpha 0 < CI_a_and_b[2,1] | 0 > CI_a_and_b[2,2]
[1] TRUE [1] TRUE
```

## Example 10.6: see Ex. 10.4 (5)

Now, we're not interested in the *significance* of  $\beta$ , but whether the value of the parameter is at least 1. By negating the working hypothesis, we obtain the null hypothesis  $H_0 : \beta \leq \beta_0 = 1$  and the corresponding alternative hypothesis.

We obtain the test statistic with the R outputs via

$$\frac{B - \beta_0}{S_e} \cdot \sqrt{\sum_i (x_i - \bar{x})^2} = \frac{B - \beta_0}{S_e / \underbrace{\sqrt{\sum_i (x_i - \bar{x})^2}}_{=0.07774}} = \frac{1.81740 - 1}{0.07774} = 10.5 \quad .$$

At a significance level of  $\alpha = 0.05$ , we derive from  $t(0.95|8) = 1.860$  the rejection of the null hypothesis.

Practical example: The consumption rate is at most 0.8.

## Example 10.6: see Ex. 10.4 (6)

Test decision regarding  $H_0 : \beta \leq \beta_0 = 1$  in R:

```
b0 <- 1

Teststat_b0 <- (b - b0) / Std_e *
               sqrt(sum((x10_2 - SpMean_x)^2))

c_stat_b0 <- qt(p = 1-alpha, df = df)
```

```
Teststat_b0      c_stat_b0
```

```
[1] 10.51523      [1] 1.859548
```

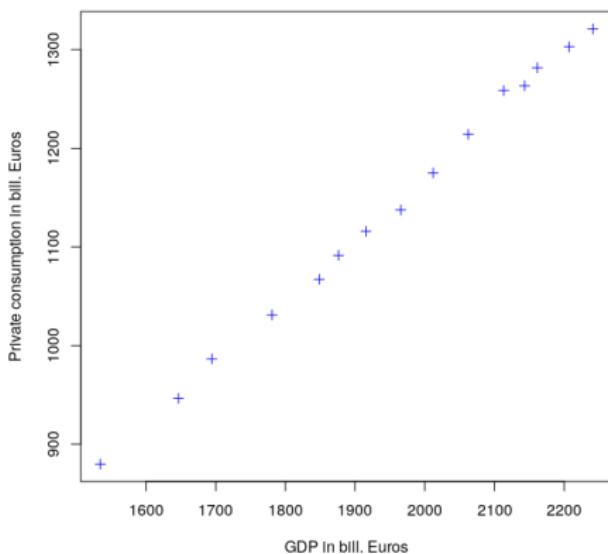
```
Teststat_b0 > c_stat_b0
```

```
[1] TRUE
```

## Example 10.7: GDP and cons. expenditure

The following table contains the gross domestic product (bill. Euro) and the private consumption expenditure (bill. Euro) (**dependent variable**) for the years 1991 until 2005.

Year	$GDP_i$	$C_i$
1991	1534.60	879.86
1992	1646.62	946.60
1993	1694.37	986.54
1994	1780.78	1031.10
1995	1848.45	1067.19
1996	1876.18	1091.50
1997	1915.58	1115.78
1998	1965.38	1137.51
1999	2012.00	1175.01
2000	2062.50	1214.16
2001	2113.16	1258.57
2002	2143.18	1263.46
2003	2161.50	1281.76
2004	2207.20	1302.94
2005	2241.00	1321.06



You can find the data in the file Example10-7.RData.

Estimate  $\alpha$  and  $\beta$  using OLS and calculate  $r_{xy}^2$ .

# Change of measuring units

- ▶ Multiplying the dependent variable  $y$  with a constant  $c$  also multiplies  $a$  and  $b$  with this constant.
- ▶ Multiplying the independent variable  $x$  with a constant  $c$  changes nothing for  $a$ .  $b$  on the other hand is divided by this constant.
- ▶ The value of the coefficient of determination is independent of changing the measuring units and stays the same in both cases.

## Why multiple regressors?

In reality, a variable  $y$  depends seldom on only one regressor  $x$ .

For example, income doesn't only depend on the level of education, but also on other determinates like gender, age and job tenure.

The *multiple regression* extends the model of simple linear regression by letting more than one independent variable determine the dependent variable.

The assumptions of the simple linear regression must be kept also here. Additionally, there are some further assumptions to be made.

# The multiple regression model (1)

We have the following regression model:

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_{p-1} \cdot X_{p-1} + \varepsilon$$

- ▶  $y$  is metric in the linear regression model.
- ▶ The independent variables however don't need to be. For example, income (metric) depends on age (metric) as well as on job tenure (metric) and gender (categorial).
- ▶ In order to include gender for example in the multiple regression model, a *dummy variable*  $D_i$  is created.
- ▶ It will take the values 0 for the *reference category* and 1 for the category of interest.

## The multiple regression model (2)

Let now the reference category for the variable gender be *male* ( $D_i = 0$ ). Then, we set  $D_i = 1$  for *female* observations.

From this, we implicitly obtain **two** regression equations:

Model for men (reference group):

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon$$

Model for women (group of interest):

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \delta + \varepsilon$$

If a categorial variable has more than two domains, we have to generate multiple dummy variables.

In this case, we take one domain as reference category. For  $m$  domains, we get  $m - 1$  dummy variables  $D_1, \dots, D_{(m-1)}$ .

If the  $i$ -th observation is part of the reference category, we have

$D_{ij} = 0, \forall j = 1, \dots, m - 1$  and the following model:

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \delta_1 \cdot D_1 + \dots + \delta_{(m-1)} \cdot D_{(m-1)} + \varepsilon$$

## Example 10.8: see Ex. 10.6

Multiple regression model in R:

```
g10_8 <- factor(x = c("w", "w", "m", "w", "m", "w", "w",
                      "m", "m", "w"), levels = c("m", "w"))
reg_analysis <- lm(formula = y10_2 ~ x10_2 + g10_8)
summary(reg_analysis)
```

Call:

```
lm(formula = y10_2 ~ x10_2 + g10_8)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.5504	-1.4267	-0.5005	1.1902	3.6159

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.85787	2.82465	1.366	0.214
x10_2	1.90105	0.06877	27.643	2.08e-08 ***
g10_8w	4.14633	1.64727	2.517	0.040 *

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.234 on 7 degrees of freedom

Multiple R-squared: 0.9924, Adjusted R-squared: 0.9903

F-statistic: 458.7 on 2 and 7 DF, p-value: 3.777e-08

## F-test

Testing the whole model requires the simultaneous test of all  $p - 1$  parameters, excluding the intercept.

$$H_0 : \beta_1 = \dots = \beta_{p-1} = 0 \text{ versus}$$

$$H_1 : \text{there is at least one } j \in \{1, \dots, p - 1\} \text{ with } \beta_j \neq 0.$$

For linear regression, we use the  $F$ -test:

$$F = \frac{\frac{1}{p-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\frac{1}{n-p} \sum_{i=1}^n e_i^2} = \frac{\frac{1}{p-1} r_{xy}^2}{\frac{1}{n-p} (1 - r_{xy}^2)}$$

The test statistic is  $F$ -distributed with  $(p - 1, n - p)$  degrees of freedom.

# Common violations of assumptions

Linear regression models are based on certain assumptions (see above). If those assumptions are not met, the quality of the estimates decreases. In order to *cure* these violations, a transformation of the model can be useful.

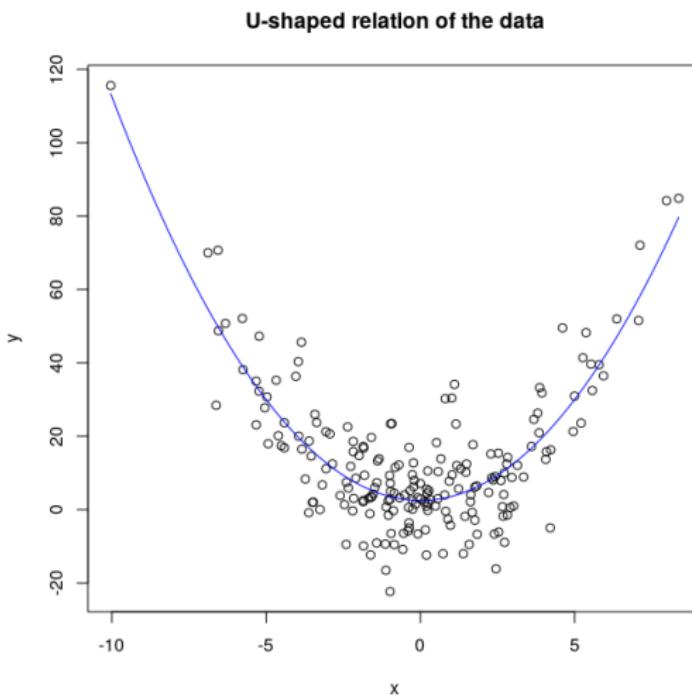
Some empirical problems are:

- ▶ Autocorrelation of independent variables
- ▶ Non-linearity of independent variables
- ▶ Heteroskedasticity of error terms
- ▶ Non-normality of the dependent variable
- ▶ Multicollinearity

# Quadratic correlation

In order to model a u-shaped correlation, one variable can be squared

Model:  $y = \beta_0 + \beta_1 x_1^2 + \varepsilon$



# Transformations

Depending on the relation of dependent and independent variable, there exist a number of different transformations.

Transformation	Formula	Linearisation
Linear	$Y = \alpha + \beta \cdot x$	
Logarithmic	$Y = \alpha + \beta \cdot \ln(x)$	
Inverse	$Y = \alpha + \beta/x$	$Y = \alpha + \beta \cdot 1/x$
Quadratic	$Y = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2$	
Cubic	$Y = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$	
Power	$Y = \alpha \cdot x^\beta$	$\ln(Y) = \ln(\alpha) + \beta \cdot \ln(x)$
Composite	$Y = \alpha \cdot \beta^x$	$\ln(Y) = \ln(\alpha) + \ln(\beta) \cdot x$
S-curve	$Y = e^{\alpha+\beta/x}$	$\ln(Y) = \alpha + \beta \cdot 1/x$
Logistic	$Y = \frac{1}{1/M+\alpha \cdot \beta^x}$	$\ln\left(\frac{1}{Y} - \frac{1}{M}\right) = \ln(\alpha) + \ln(\beta) \cdot x$
Buildup	$Y = e^{\alpha+\beta \cdot x}$	$\ln(Y) = \alpha + \beta \cdot x$
Exponential	$Y = \alpha \cdot e^{\beta \cdot x}$	$\ln(Y) = \ln(\alpha) + \beta \cdot x$

Overview from Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2011): Multivariate Analysemethoden, Springer-Verlag Berlin Heidelberg, Aufl. ?, p. 141.

# Linear regression with heteroskedasticity

If we have heteroskedasticity, OLS is not efficient. Furthermore, the standard errors of the coefficients are biased and thus inconsistent. Particularly, the parameters of the model can't be tested like before.

Two transformations to be applied in this context are

- ▶ to logarithmise the dependent variable and
- ▶ to square the dependent variable.

With this, we want to reach homoskedastic error terms. Then, we can apply our familiar tests. We have to proof, if the transformation was successful in the particular case.

Alternative approaches are using *robust* standard errors and a weighted estimation.

## Linear regression to calculate (constant) elasticities

An elasticity gives the relative change of a dependent variable due to a change in the independent variable.

Elasticities are of interest for example in economics (price elasticity) and are defined by:

$$E = \frac{\Delta y}{\Delta x} \cdot \frac{x}{y}$$

As

$$\beta = \frac{\Delta \ln(y)}{\Delta \ln(x)} \approx \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = E$$

the  $\beta$  of a *log-log-model* can be directly interpreted as (constant) elasticity.

## Example 10.9: Elasticity estimation

In order to estimate the elasticities, a log-log-model is stated as follows:

$$\ln(y) = \alpha + \beta \cdot \ln(x)$$

We obtain an estimation of  $\hat{\alpha} = 1$  and  $\hat{\beta} = 3$ .

x	$\Delta x/x$	% $\Delta x$	$y = e^{1+3 \cdot \ln(x)}$	$\Delta y/y$	% $\Delta y$
2			21.74625		
2.02	$\frac{2.02-2}{2} = 0.01$	1%	22.40519	0.030301	$\approx 3\%$

This means that a relative change of  $x$  by 1% results approximately, according to the model, in a relative change of  $y$  by 3%.