Introduction to Statistics and Data Analysis

A Case-Based Approach

Conrad Ziller

"Introduction to Statistics and Data Analysis – A Case-Based Approach"

Conrad Ziller, University of Duisburg-Essen

Contents

Preface **3**
- Motivation for this Book . . . 3
- How to Use the Book . . . 3
- Organization of the Book . . . 3
- Acknowledgments . . . 4
- About the Author . . . 4
- Outlook . . . 4

Univariate Statistics – Case Study Socio-Demographic Reporting **7**
- Introduction . . . 7
- Preparing the data . . . 8
- Representation of spatial patterns using maps . . . 8
- Distributions: First insights into the data structure . . . 10
- Mean . . . 11
- Measures of dispersion: Variance and standard deviation . . . 13
- Position measures: Boxplots and interquartile range . . . 16
- Conclusion . . . 18
- Outlook: Does unemployment cause more crime? . . . 18

Bivariate Statistics – Case Study United States Presidential Election **23**
- Introduction . . . 23
- Data overview and descriptive analyses . . . 23
- "Who supports Trump?" - Delving into bivariate statistics . . . 26
- Scatter plot and correlation . . . 33
- Chi-square and Cramér's V . . . 38

Statistical Inference - Case Study Satisfaction with Government **40**
- Introduction . . . 40
- Probability . . . 40
- Basics of statistical inference . . . 43
- Confidence intervals . . . 45
- Null Hypothesis Testing . . . 54
- Outlook . . . 62

Regression Analysis - Case Study Attitudes Toward Justice **65**
- Data preparation and bivariate regression . . . 65
- Multiple Regression . . . 74
- OLS assumptions and model diagnostics, non-linear relationships between variables, and interactions 78

Preface

Suggested citation:

Ziller, Conrad (2024). *Introduction to Statistics and Data Analysis – A Case-Based Approach.* Available online at https://bookdown.org/conradziller/introstatistics

To download the R-Scripts and data used in this book, go HERE.

Motivation for this Book

This short book is a complete introduction to statistics and data analysis using R and RStudio. It contains hands-on exercises with real data—mostly from social sciences. In addition, this book presents four key ingredients of statistical data analysis (univariate statistics, bivariate statistics, statistical inference, and regression analysis) as brief case studies. The motivation for this was to provide students with practical cases that help them navigate new concepts and serve as an anchor for recalling the acquired knowledge in exams or while conducting their own data analysis.

The case study logic is expected to increase motivation for engaging with the materials. As we all know, academic teaching is not the same as before the pandemic. Students are (rightfully) increasingly reluctant to chalk-and-talk techniques of teaching, and we have all developed dopamine-related addictions to social media content which have considerably shortened our ability to concentrate. This poses challenges to academic teaching in general and complex content such as statistics and data science in particular.

How to Use the Book

This book consists of four case studies that provide a short, yet comprehensive, introduction to statistics and data analysis. The examples used in the book are based on real data from official statistics and publicly available surveys. While each case study follows its own logic, I advise reading them consecutively. The goal is to provide readers with an opportunity to learn independently and to gather a solid foundation of hands-on knowledge of statistics and data analysis. Each case study contains questions that can be answered in the boxes below. The solutions to the questions can be viewed below the boxes (by clicking on the arrow next to the word "solution"). It is advised to save answers to a separate document because this content is not saved and cannot be accessed after reloading the book page.

A working sheet with questions, answer boxes, and solutions can be downloaded together with the R-Scripts HERE. You can read this book online for free. Copies in printable format may be ordered from the author.

This book can be used for teaching by university instructors, who may use data examples and analyses provided in this book as illustrations in lectures (and by acknowledging the source). This book can be used for self-study by everyone who wants to acquire foundational knowledge in basic statistics and practical skills in data analysis. The materials can also be used as a refresher on statistical foundations.

Beginners in R and RStudio are advised to install the programs via the following link https://posit.co/download/rstudio-desktop/ and to download the materials from HERE. The scripts from this material can then be executed while reading the book. This helps to get familiar with statistical analysis, and it is just an awesome feeling to get your own script running! (On the downside, it is completely normal and part of the process that code for statistical analysis does not work. This is what helpboards across the web and, more recently, ChatGPT are for. Just google your problem and keep on trying, it is, as always, 20% inspiration and 80% consistency.)

Organization of the Book

The book contains four case studies, each showcasing unique statistical and data-analysis-related techniques.

- Section 2: Univariate Statistics – Case Study Socio-Demographic Reporting

Section 2 contains material on the analysis of one variable. It presents measures of typical values (e.g., the mean) and the distribution of data.

- Section 3: Bivariate Statistics - Case Study 2020 United States Presidential Election

Section 3 contains material on the analysis of the relationship between two variables, including cross tabs and correlations.

- Section 4: Statistical Inference - Case Study Satisfaction with Government

Section 4 introduces the concept of statistical inference, which refers to inferring population characteristics from a random sample. It also covers the concepts of hypothesis testing, confidence intervals, and statistical significance.

- Section 5: Regression Analysis - Case Study Attitudes Toward Justice

Section 5 covers how to conduct multiple regression analysis and interpret the corresponding results. Multiple regression investigates the relationship between an outcome variable (e.g., beliefs about justice) and multiple variables that represent different competing explanations for the outcome.

Acknowledgments

Thank you to Paul Gies, Phillip Kemper, Jonas Verlande, Teresa Hummler, Paul Vierus, and Felix Diehl for helpful feedback on previous versions of this book. I want to thank Achim Goerres for his feedback early on and for granting me maximal freedom in revising and updating the materials of his introductory lectures on Methods and Statistics, which led to the writing of this book. Earlier versions of this book have been used in teaching courses on statistics in the Political Science undergraduate program at the University of Duisburg-Essen.

About the Author

Conrad Ziller is a Senior Researcher in the Department of Political Science at the University of Duisburg-Essen. His research interests focus on the role of immigration in politics and society, immigrant integration, policy effects on citizens, and quantitative methods. He is the principal investigator of research projects funded by the German Research Foundation and the Fritz Thyssen Foundation. More information about his research can be found here: https://conradziller.com.

Outlook

The final part of the book is about linear regression analysis, which is the natural endpoint for a course on introductory statistics. However, the "ordinary" regression is where many further useful techniques come into play—most of which can subsumed under the label "Advanced Regression Models". You will need them when analyzing, for example, panel data where the same respondents were interviewed multiple times or spatially clustered data from cross-national surveys.

I will extend this introduction with case studies on advanced regression techniques soon. If you want to get notified when this material is online, please sign up with your email address here: https://forms.gle/T8Hvhq3EmcywkTdFA.

In the meantime, I have a chapter on *"Multiple Regression with Non-Independent Observations: Random-Effects and Fixed-Effects"* that can be downloaded via https://ssrn.com/abstract=4747607.

For feedback on the usefulness of the introduction and/or reports on errors and misspellings, I would be utmost thankful if you would send me a short notification at conrad.ziller@uni-due.de.

Thanks much for engaging with this introduction!

This syntax loads all packages necessary to conduct the subsequent analyses.

```
# Required packages are installed and loaded here. If more packages are needed, they can
# be added to the list

if (!require(dplyr)){install.packages("dplyr"); library(dplyr)}
if (!require(ggplot2)){install.packages("ggplot2"); library(ggplot2)}
if (!require(tidyverse)){install.packages("tidyverse"); library(tidyverse)}
if (!require(viridis)){install.packages("viridis"); library(viridis)}
if (!require(knitr)){install.packages("knitr"); library(knitr)}
if (!require(sf)){install.packages("sf"); library(sf)}
if (!require(readxl)){install.packages("readxl"); library(readxl)}
if (!require(Hmisc)){install.packages("Hmisc"); library(Hmisc)}
if (!require(foreign)){install.packages("foreign"); library(foreign)}
if (!require(sjPlot)){install.packages("sjPlot"); library(sjPlot)}
if (!require(lattice)){install.packages("lattice"); library(lattice)}
if (!require(lmtest)){install.packages("lmtest"); library(lmtest)}
if (!require(rcompanion)){install.packages("rcompanion"); library(rcompanion)}
if (!require(flextable)){install.packages("flextable"); library(flextable)}
if (!require(haven)){install.packages("haven"); library(haven)}
if (!require(descr)){install.packages("descr"); library(descr)}
if (!require(stargazer)){install.packages("stargazer"); library(stargazer)}
if (!require(interactions)){install.packages("interactions"); library(interactions)}
if (!require(car)){install.packages("car"); library(car)}
```

Univariate Statistics – Case Study Socio-Demographic Reporting

Introduction

The 2020 socio-demographic report on the German state North-Rhine Westphalia continues the state's social reporting since its beginning in 1992. These reports aim to provide the public with information on the social and demographic conditions and dynamics of North Rhine-Westphalia (NRW).

Social indicators refer to demographic conditions and developments (e.g., population figures, the proportion of people with a migration background), the economy (e.g., unemployment rates, GDP), health, education, housing, public finances, and social inequality.

The report and underlying data we are dealing with in this case study can be accessed online via https://www.sozialberichte.nrw.de.

Cover Sozialbericht

Analyses based on data from the socio-demographic report are politically relevant. Insights from data analyses may provide politicians and public officials with information on social problems that need to be addressed by political measures. It is thus important that interpretations derived from data analyses are accurate. In the following case study, we take a closer look at the data and use methods from univariate statistics. Please note that the analyses and interpretations below are meant to illustrate data analysis techniques and do not reflect any policy advice.

We will use the following indicators or *variables* (source https://www.inkar.de):

- Population density
- Proportion of foreigners
- Proportion of unemployed
- Crime rate
- Average age

Preparing the data

Before we can analyze the data, they must be read in the data analysis program. R uses a lot of so-called "packages" (i.e., software add-ons that facilitate or simplify specific parts of data analysis). For example, the package "readxl" allows us to read data that is stored in an Excel format. As another example, the package "ggplot2" allows us to create print-ready figures. Packages often have a specific syntax to which we refer along the way.

```
library(readxl) # This command loads the package "readxl". In the markdown document, all
# required packages are installed in the background using the command
# "install.packages('packagename')". In the downloadable script files, the installation
# and activation of the packages is automated.

data_nrw <- read_excel("data/inkar_nrw.xls") # Reads data from Excel format; the
# directory where the data is to be found can be changed (e.g.,
# "C:/user/documents/folderxyz/inkar_nrw.xls")
```

After reading the file "inkar_nrw.xlsx", it is stored as an object in R. We can now work with the data. As a first step, we print the first six rows of the data frame. For reasons of space, we only show a selection of the variables in the table.

```
ft1 <- flextable(head(select(data_nrw, kkziff, area, population, unemp, crimerate)))
autofit(ft1)
```

kkziff	area	population	unemp	crimerate
05111	Düsseldorf, Stadt	620,523	7.77	9,998.762
05112	Duisburg, Stadt	495,885	12.10	8,640.908
05113	Essen, Stadt	582,415	11.04	7,472.201
05114	Krefeld, Stadt	226,844	11.11	8,866.532
05116	Mönchengladbach, Stadt	259,665	10.02	8,256.396
05117	Mülheim an der Ruhr, Stadt	170,921	8.31	5,285.058

The data contains the following variables for all 53 districts and large cities in North-Rhine Westphalia (reference year is 2020):

`kkziff` = identifier code for the county (*Landkreis*) or city (*kreisfreie Stadt*)

`area` = County / city name

`nonnationall` = Proportion of foreign-born residents in %

`population` = Number of inhabitants

`flaeche` = Land area in square kilometers

`unemp` = Unemployment rate in %

`avage` = Average age of the population

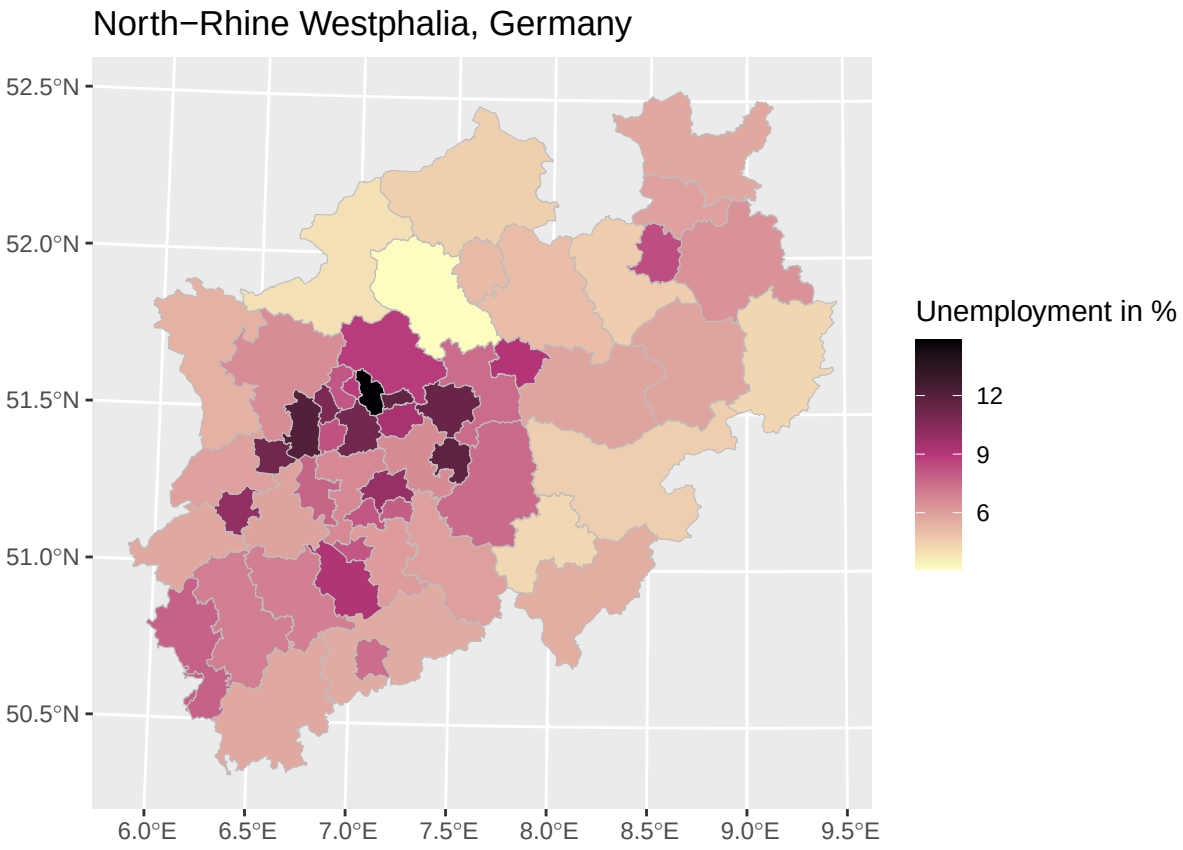
`crimerate` = Crime rate in cases per 100,000 inhabitants

`KN` = Alternative id code that we need for the map below

Representation of spatial patterns using maps

One option to represent spatially structured data is in the form of a map.

```r
# Read a so-called shape file for NRW which contains the polygones (e.g., district
# boundaries) necessary to build maps and merging with structural data
nrw_shp <- st_read("data/dvg2krs_nw.shp")
nrw_shp$KN <- as.numeric(as.character(nrw_shp$KN))
nrw_shp <- nrw_shp %>%
  right_join(data_nrw, by = c("KN"))

# Building the map using the "ggplot2" package
ggplot() +
  geom_sf(data = nrw_shp, aes(fill = unemp), color = 'grey', size = 0.1) +
  ggtitle("North-Rhine Westphalia, Germany") +
  guides(fill = guide_colorbar(title = "Unemployment in %")) +
  scale_fill_gradientn(colours = rev(magma(3)), na.value = "grey100")
```



Question: What information is conveyed by the map? What information remains implicit?

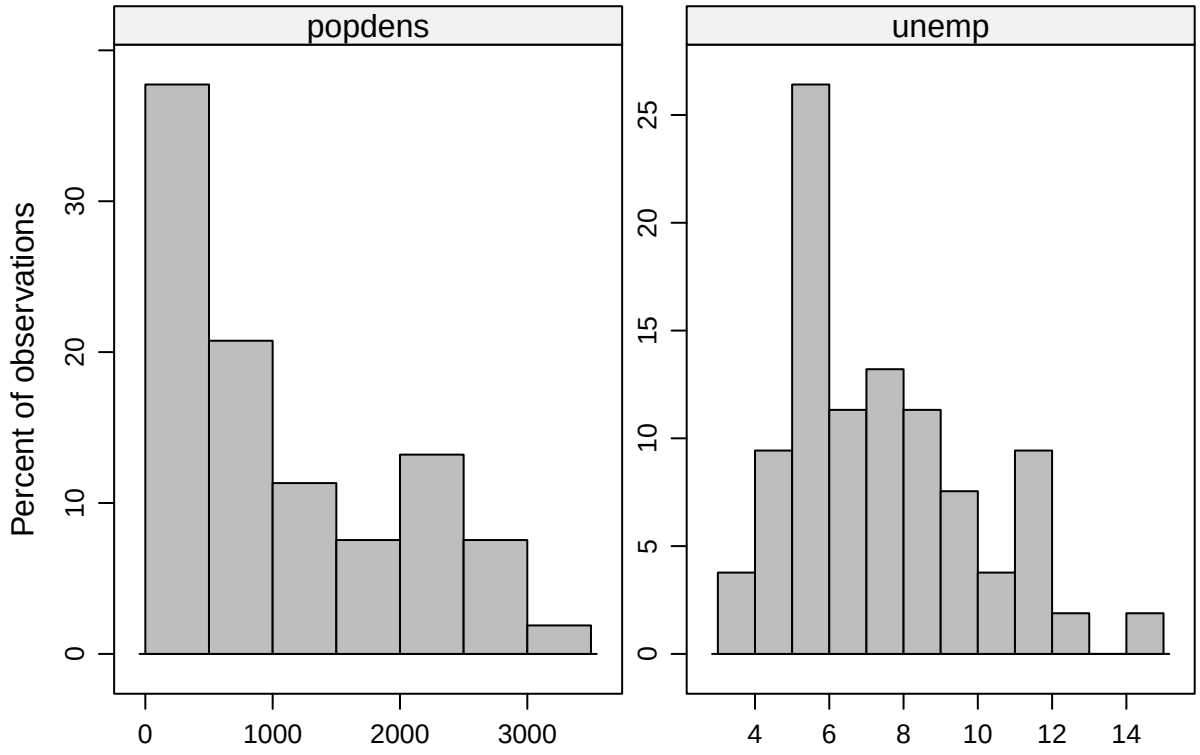
Solution:

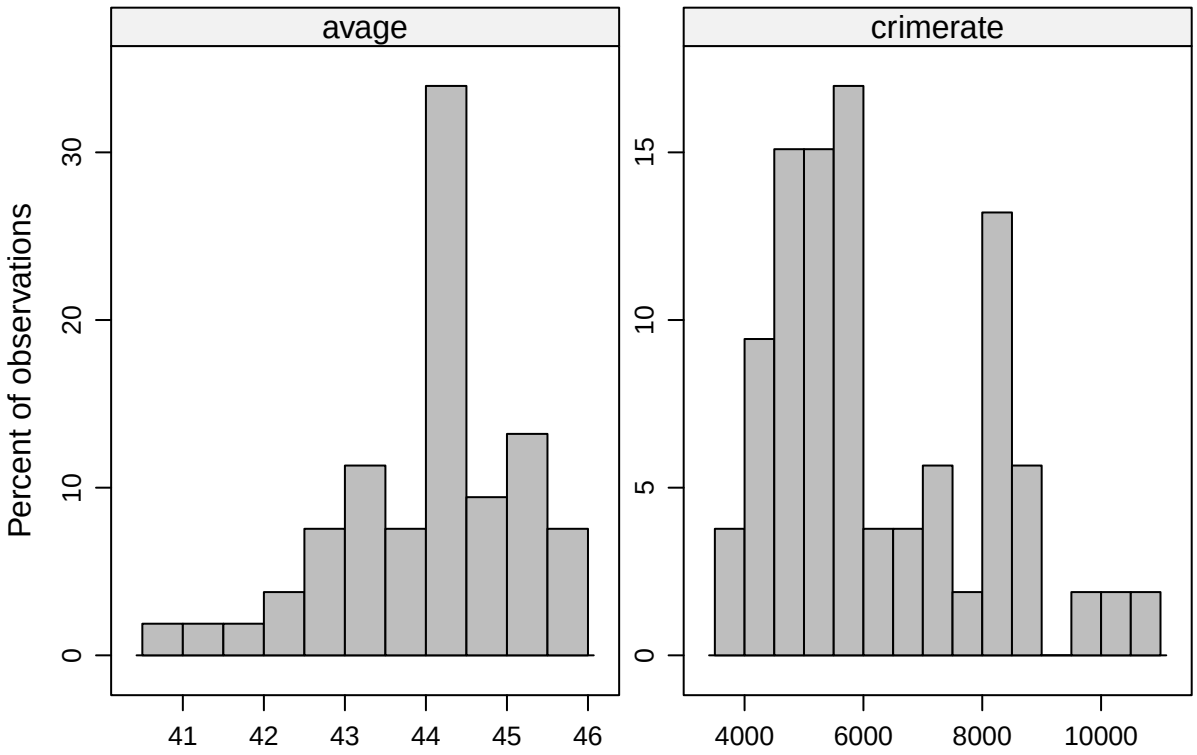
- Maps give a good first impression of the spatial distribution of data.
- However, insights depend on further (implicit) information ("What are the specifics of the high-unemployment cities in the middle of the map? Which areas are rural, which are urban?").
- Difficult to make general statements about maps and to compare maps (especially if the breaks that underly the colorization differ).

Distributions: First insights into the data structure

To get an impression about the distribution of the data, histograms can be used (e.g., What are minimum or maximum values? Do observed values are clusterd around the center of the distrubtion or are they widely spread?). Histograms show the frequency of observations across the range of values of a variable. It does not matter how the single observations (i.e., districts in our case) appear in the data set. The histogram ranks observations depending on the values they have on a specific variable. The height of the displayed bars is proportional to the frequency of observations that have corresponding variable values. The width of the bars (also known as bins) represents an interval of values and can be customized (i.e., a few wide bars comprise more observations than many narrow ones).

```
library(lattice)

data_nrw$popdens <- data_nrw$population / data_nrw$flaeche # This generates the variable
                                                           # for the population density

histogram( ~ popdens + unemp + avage + crimerate,
           breaks = 10,
           type = "percent",
           xlab = "",
           ylab = "Percent of observations",
           layout = c(2, 1),
           scales = list(relation = "free"),
           col = 'grey',
           data = data_nrw)
```






```
## Alternative approach with the "hist"-function
#hist(data_nrw$unemp)
#hist(data_nrw$avage)
#hist(data_nrw$crimerate)
```

Question: Which characteristics of the data can be identified with histograms? Which aspects remain hidden?

Solution:

What can be seen: The distribution of the data, including the center(s) where most of the observations are located, patterns of skewness, and outliers.

What remains hidden: The specific estimates of the summary statistics (e.g., mean, variance).

Mean

The mean of a variable is also referred to as a measure of central tendency. It thus represents a good first impression of what is a typical (or expectable) value of a measured characteristic (i.e., a variable). Comparing the mean and the variable value of a specific observation provides information about whether the observation is close to or far from the mean. Besides the arithmetic mean, other measures exist (e.g., the geometric mean), but we do not consider them here.

For the calculation of the mean, all observed values for a variable are added and the sum is then divided by the total number of observations (n).

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ or $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

$\bar{x}$ and n refer to given data we are working with (e.g., from a *sample*), μ and N refer to the *population* we

want to make statistical statements about. (Note: This distinction becomes relevant in statistical testing, where we use a random sample to draw inferences about an underlying larger population.)

Let's take a look at the means and other information on the variables in the data set:

```
summary(data_nrw)
```

```
##     kkziff              area             aggregat          nonnational
##  Length:53          Length:53          Length:53          Min.   : 6.05
##  Class :character   Class :character   Class :character   1st Qu.: 9.82
##  Mode  :character   Mode  :character   Mode  :character   Median :12.19
##                                                           Mean   :13.45
##                                                           3rd Qu.:16.87
##                                                           Max.   :21.86
##    population         flaeche           unemp            avage
##  Min.   : 111516   Min.   :  51.0   Min.   : 3.110   Min.   :40.97
##  1st Qu.: 226844   1st Qu.: 170.0   1st Qu.: 5.740   1st Qu.:43.33
##  Median : 308335   Median : 543.0   Median : 6.950   Median :44.15
##  Mean   : 338218   Mean   : 643.6   Mean   : 7.432   Mean   :44.01
##  3rd Qu.: 408662   3rd Qu.:1112.0   3rd Qu.: 8.850   3rd Qu.:44.62
##  Max.   :1083498   Max.   :1960.0   Max.   :14.870   Max.   :45.81
##    crimerate           KN             popdens
##  Min.   : 3718   Min.   :5111000   Min.   : 116.3
##  1st Qu.: 4906   1st Qu.:5166000   1st Qu.: 278.5
##  Median : 5636   Median :5512000   Median : 784.7
##  Mean   : 6244   Mean   :5506094   Mean   :1072.0
##  3rd Qu.: 7568   3rd Qu.:5770000   3rd Qu.:1768.8
##  Max.   :10500   Max.   :5978000   Max.   :3077.3
```

Question: Interpret two means of your choice. Why is no mean displayed for the variable `area`?

Solution:

The average unemployment rate across the observed districts is 7.4 percent. The mean of the population variable is 338,218, which means that about 338,218 people live in a region, on average.

Means can only be calculated for metric or quasi-metric (typically an ordinal variable with five or more categories) variables. `area` is a nominal variable.

The displayed tables contain more information than means:

- The **median** is another measure of centrality (besides the mean and the mode, i.e., the variable value that occurs the most often for a given set of observations):
  - The median is the midpoint of a frequency distribution of observed variable values (i.e., it divides ordered observed data into two equal parts).
  - Often preferred over the mean for skewed data because the median is not sensitive to outliers (i.e., extreme values).
- Measures of dispersion, such as the **range** (i.e., maximum value – minimum value) and the **standard deviation**, both providing information about how the data is distributed.
- Measures of position, such as **quartiles** and the interquartile range, which are often graphically depicted using boxplots.

Measures of dispersion: Variance and standard deviation

Imagine for a moment that we not only have the socio-demographic report of NRW available but also a report from another German state. Comparing county characteristics, such as crime rates, across the two states might result in finding that both have the exact same mean. However, the crime rate in State 1 is much more extremely distributed, with particularly high and low values for some counties. In State 2, in contrast, the crime rate is much more evenly distributed around the mean. Determining the degree of dispersion (or scatteredness) of data is an important piece of information that is useful for several further statistical analyses It might also yield practical relevance, as different patterns of dispersion of crime may lead to different policy measures for combating crime. Let us start with calculating the variance.

The variance is the sum of squared deviations.

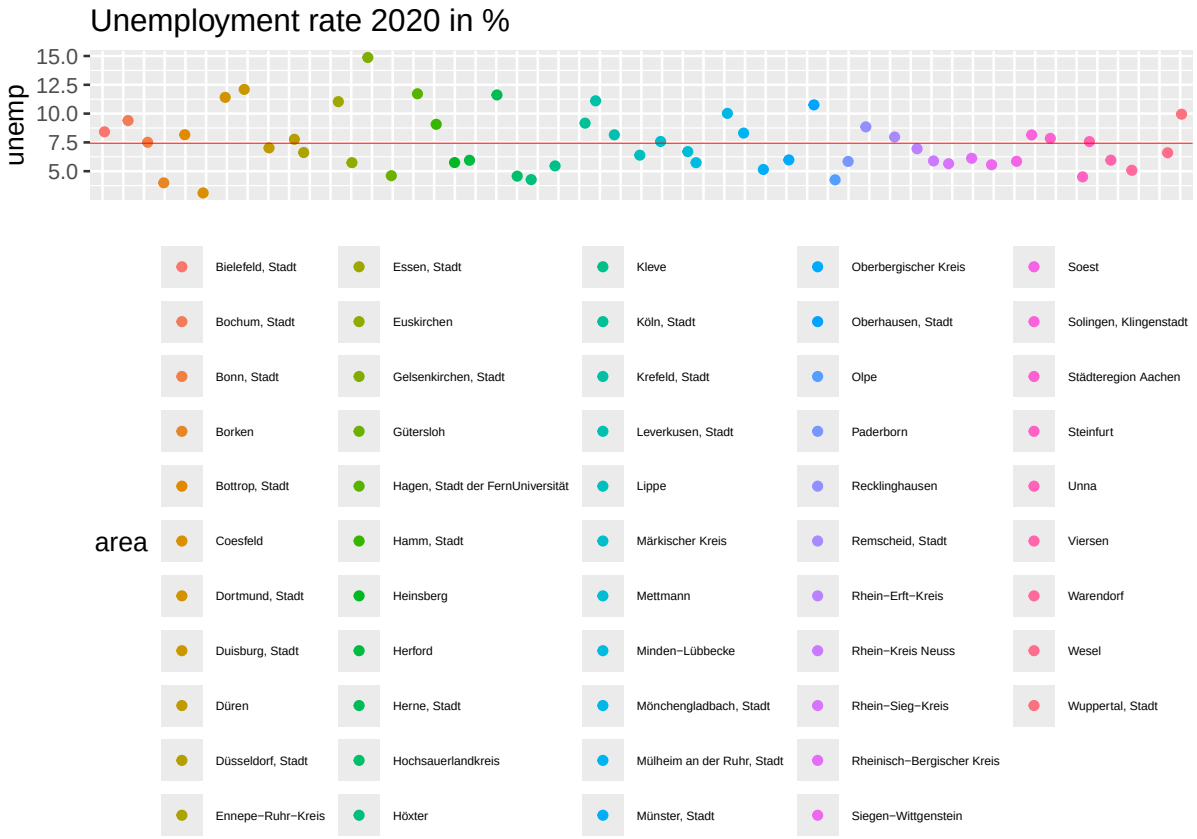
$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{N}$ (Notation for the population) $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ (Notation for the sample; If we apply statistical inference and use a random sample, we need to apply a correction in the denominator "n-1" – referred to as Bessel's correction)

The standard deviation re-transforms the value back to the scale on which the variable is measured, it is therefore easier to interpret and more commonly used.

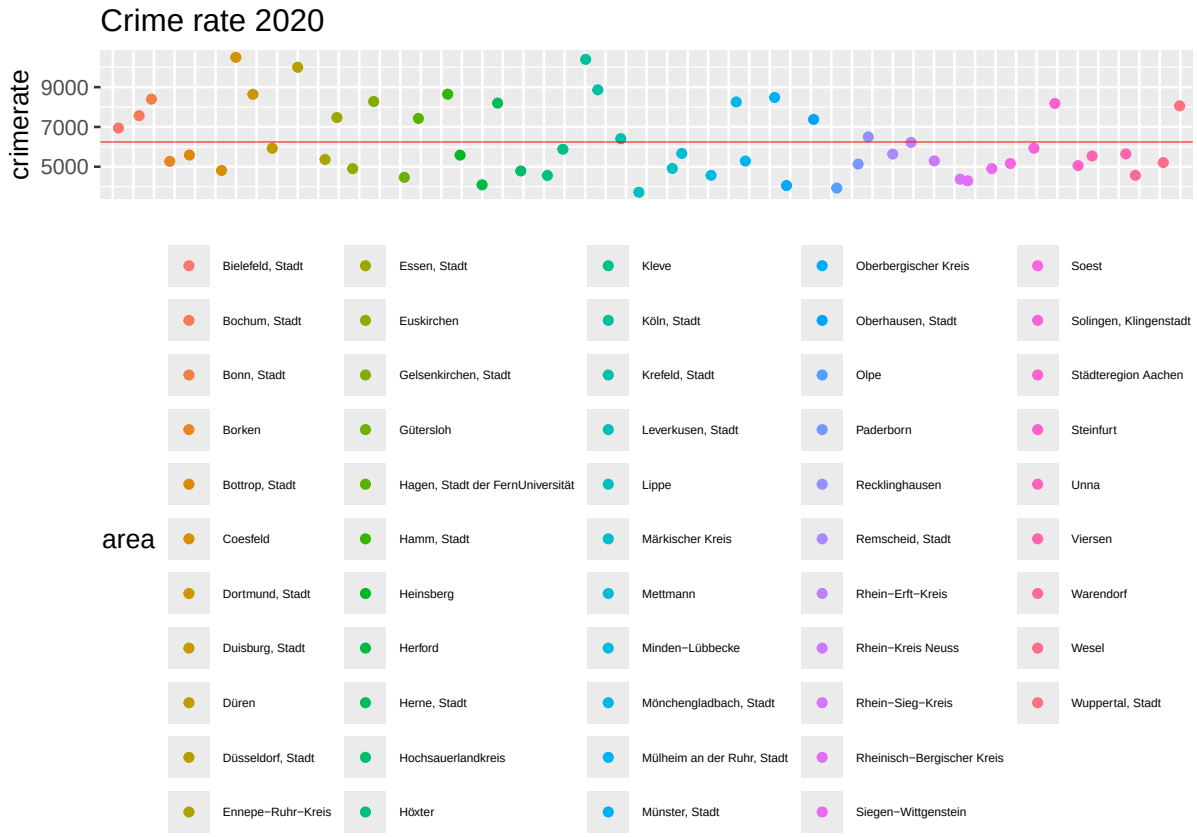
$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

Here you find a graphical representation of the dispersion of the variables `unemp` and `crimerate`. Note that the deviations between observations on the x-axis are purely random to increase the readability of the plot (otherwise, all observations would be lined up like a vertical pearl necklace). The red line is the mean value.

```
s_unemp <- ggplot(data=data_nrw, aes(y=unemp, x=reorder(area, area), color=area)) +
  geom_jitter(height = 0) +
  ggtitle("Unemployment rate 2020 in %")

s_unemp +
  geom_hline(yintercept = 7.432, linetype = "solid", color = "red", size = 0.1) +
  theme(legend.position = "bottom") +
  theme(legend.text = element_text(size = 5)) +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```




```
s_crime <- ggplot(data=data_nrw, aes(y=crimerate, x=reorder(area, area), color=area)) +
  geom_jitter(height = 0) +
  ggtitle("Crime rate 2020")

s_crime +
  geom_hline(yintercept = 6244, linetype = "solid", color = "red", size = 0.1)  +
  theme(legend.position = "bottom") +
  theme(legend.text = element_text(size = 5)) +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```



Question: Can you spot which district has the highest unemployment rate? Which county has the lowest crime rate? Can we infer from the graphs which of the two variables is more dispersed?

Solution:

It is quite hard to match the color of the dots with the ones of the legend. Gelsenkirchen has the highest unemployment rate at 14.9%. The district with the lowest crime rate is Lippe.

Because of the different scaling of the variables, it is not possible to infer from this plot which variable is more dispersed. To do so, we would need to calculate the so-called coefficient of variation (see below).

Standard deviation, range, and coefficient of variation calculated with R (unemployment):

```
sd(data_nrw$unemp)
```

```
## [1] 2.477591
```

```
range <- max(data_nrw$unemp, na.rm = TRUE) - min(data_nrw$unemp, na.rm = TRUE)
range
```

```
## [1] 11.76
```

```
varcoef <- sd(data_nrw$unemp) / mean(data_nrw$unemp) * 100 # how much percent of the mean
                                                           # is the standard deviation?
varcoef
```

```
## [1] 33.33815
```

Standard deviation, min/max, and coefficient of variation calculated with R (crime rate):

```
sd(data_nrw$crimerate)
```

```
## [1] 1775.389
```

```
range <- max(data_nrw$crimerate, na.rm = TRUE) - min(data_nrw$crimerate, na.rm = TRUE)
range
```

```
## [1] 6782.056
```

```
varcoef <- sd(data_nrw$crimerate) / mean(data_nrw$crimerate) * 100
varcoef
```

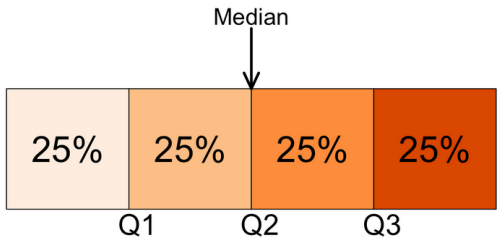
```
## [1] 28.43261
```

To compare the dispersion of variables measured on different scales, we cannot use the variance or standard deviation (as they are measured on the units of the scale) but employ the coefficient of variation (i.e., standard deviation over the mean times 100). Here, higher values imply a higher dispersion of the data.

> ***Interpretation: The variable unemployment is more heterogeneously dispersed (varcoef = 33.3) than the variable crime rate (varcoef = 28.4).***

Position measures: Boxplots and interquartile range

Quartiles divide the ordered data into four parts (Q1 to Q4, where Q4 is the maximum of the observed values), and the median is represented by Q2. Hence, 25% of the observations are below or at Q1, 75% above. Q3-Q1 = interquartile range (IQR), which reflects the range in which the middle 50% of the observations are located.



Quartiles

Quartiles and IQR for unemployment rate:

```
quantile(data_nrw$unemp)
```

```
##    0%   25%   50%   75%  100%
##  3.11  5.74  6.95  8.85 14.87
```

```
unemp.score.quart <- quantile(data_nrw$unemp, names = FALSE)
unemp.score.quart[4] - unemp.score.quart[2]
```

```
## [1] 3.11
```

Quartiles and IQR for crime rate:

```
quantile(data_nrw$crimerate)
```

```
##        0%       25%       50%       75%      100%
##  3718.411  4906.122  5635.991  7568.376 10500.467
```

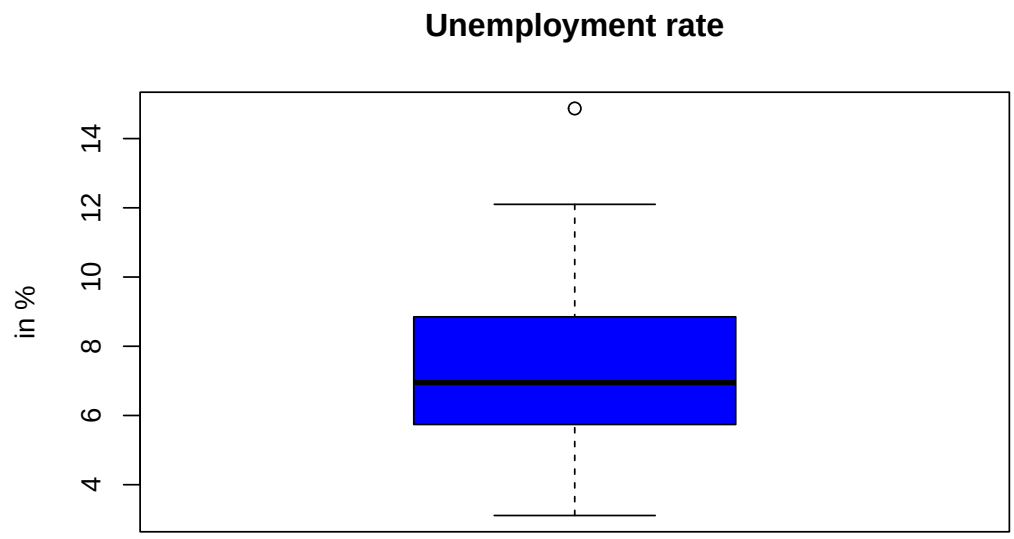
```
crimerate.score.quart <- quantile(data_nrw$crimerate, names = FALSE)
crimerate.score.quart[4] - crimerate.score.quart[2]
```

```
## [1] 2662.254
```

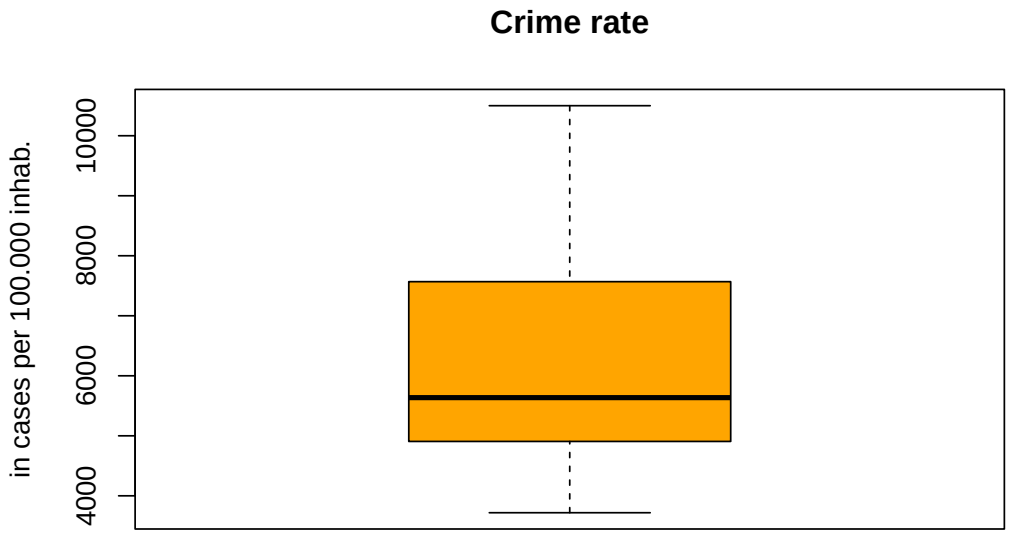
Boxplots graphically represent the key figures Q1, Q2, Q3. The lower and upper ends of the whiskers are usually the minimum and maximum observed values. To mark observations that are far away from the median (so-called outliers), the maximal width of a whisker is restricted to Q1 - 1.5 x IQR (for the lower whisker) and Q3 + 1.5 x IQR (for the higher upper whisker) respectively. Outliers are represented by points outside of the range of the whisker. The lowest and highest oulier would then mark the minimum/maximum value. For crime rate, the end of the whiskers represent the minimum/maximum value. For unemployment, we find an outlier that represents the maximum of observed values (i.e., Gelsenkirchen with 14.9% unemployment).

Apart from that, the larger the single parts in the boxplot, the greater the dispersion of the data in this area (which can also give an impression of the skewness of the data distribution).

```
boxplot(data_nrw$unemp,
        col = 'blue',
        horizontal = FALSE,
        ylab = 'in %',
        main = 'Unemployment rate')
```




```
boxplot(data_nrw$crimerate,
        col = 'orange',
        horizontal = FALSE,
        ylab = 'in cases per 100.000 inhab.',
        main = 'Crime rate')
```



Conclusion

The provided examples illustrate some basic principles of univariate descriptive statistics. Univariate statistics are important to get an overview of the structure of the data, which may also inform more complex statistical methods. In terms of further steps, we could apply the learned methods and ask: Which counties are most plagued by crime? Which counties have the lowest unemployment, and which have the highest? Where are particularly large numbers of people moving to or from? This may inspire additional questions, like: Why are the observed structures the way they are? These questions concern explanatory analyses. In the following, we take an outlook on this and – by doing so – we refer to further topics that are explained in detail in subsequent chapters.

Outlook: Does unemployment cause more crime?

To test causal claims like "more unemployment leads to more crime" with observational data (i.e., from surveys or official records) is heroic and rests on various assumptions that need to be fulfilled. The main reason why causal inference with observational (i.e., non-experimental) data is so difficult is that all possible alternative explanations must be eliminated. Otherwise, we cannot be sure whether unemployment is the cause or some other unobserved phenomenon that correlates with unemployment.

For causal inference, the following three conditions must be met:

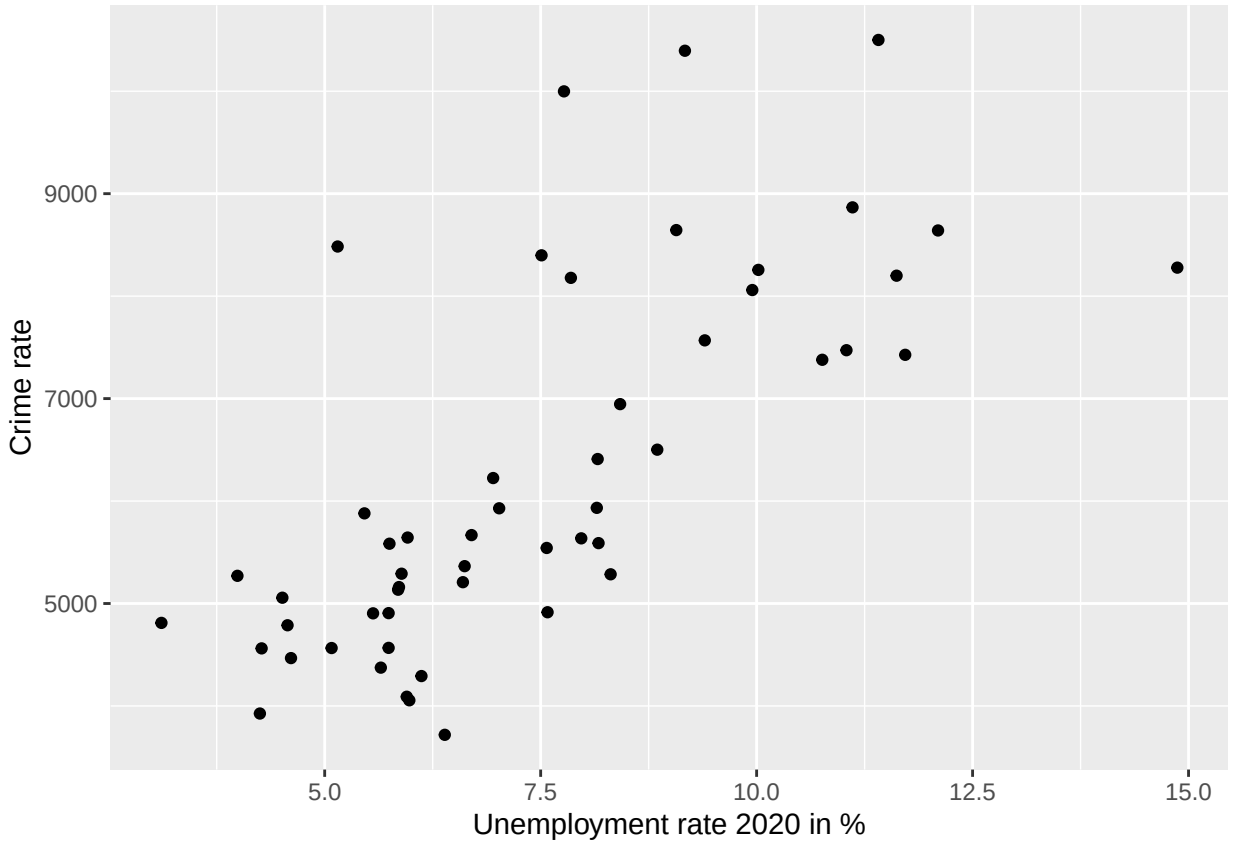
- X (cause) and Y (effect) must be empirically related to each other (e.g., correlate with each other).
- X must precede Y in time (e.g., can be mapped with panel data).
- Most importantly: All possible alternative explanations must be excluded (e.g., via a randomized experiment or by using control variables in multiple regression).

We will revisit these assumptions in the case study on regression analysis.

Bivariate analysis: Correlation

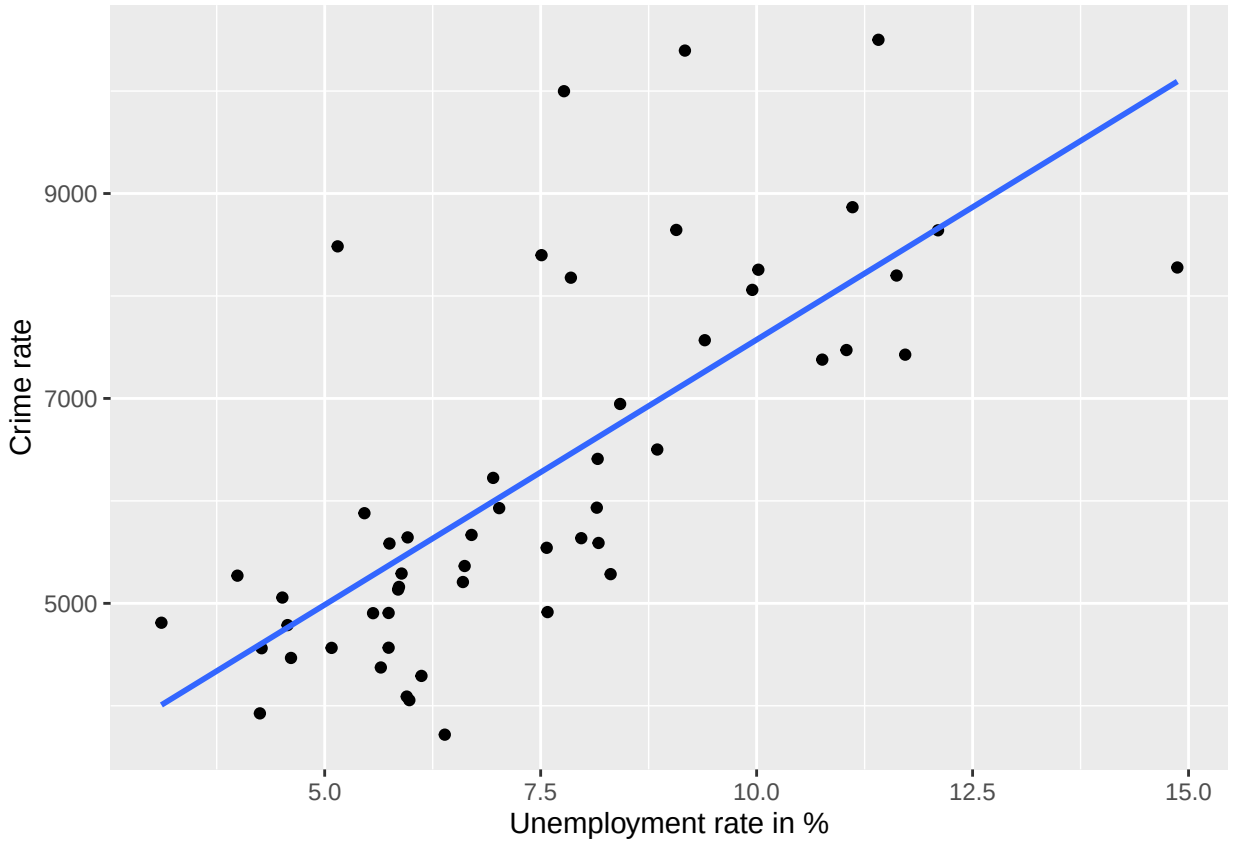
A good way to get an impression of the direction of a correlation is using a scatter plot.

```
sc1 <- ggplot(data = data_nrw, aes(x = unemp, y = crimerate)) +
  geom_point() +
  xlab("Unemployment rate 2020 in %") +
  ylab("Crime rate")

sc1
```



A line that describes the linear relationship between the two variables can be added:

```
sc1 <- ggplot(data = data_nrw, aes(x = unemp, y = crimerate)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  xlab("Unemployment rate in %") +
  ylab("Crime rate")

sc1
```



Interpretation: The depicted observations and the fitted line imply a positive correlation. (The higher the unemployment rate, the higher the crime rate.)

Pearson's correlation coefficient "r" quantifies the correlation: -1 = perfect neg. correlation, 0 = no correlation, +1 = perfect pos. correlation.

```
vars <- c("unemp", "crimerate")
cor.vars <- data_nrw[vars]
rcorr(as.matrix(cor.vars))
```

```
##           unemp crimerate
## unemp      1.00      0.72
## crimerate  0.72      1.00
##
## n= 53
##
##
```

```
## P
##           unemp crimerate
## unemp            0
## crimerate  0
```

Interpretation: In this case, the correlation coefficient is $r = 0.72$ with a p-value of < 0.001. That means that we find a strong positive correlation. The correlation is also statistically significant (we typically speak of a statistically significant result if the p-value is below 0.05), which means that we can be fairly certain that the result is not due to chance but can be interpreted as systematic.

Regression with control variables

The correlation showed that there is a positive (and statistically significant) bivariate relationship between unemployment and crime. But is this relationship causal? One way to assess this is to control for alternative explanations using multiple regression. We use population density as an alternative explanation. After all, it may well be that unemployment and crime occur primarily in an urban context. If that was the case, part of the correlation would not be due to unemployment, but to population density, and unemployment would so to say partly "transport" the effect of population density (if not included in the regression model). Let's take a look at this.

- Model 1 includes unemployment as an explanatory variable
- Model 2 includes unemployment **and** population density

```
model1 <- lm(crimerate ~ 1 + unemp, data = data_nrw)
model2 <- lm(crimerate ~ 1 + unemp + popdens, data = data_nrw)

stargazer(model1, model2, type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                       crimerate
##                                (1)                   (2)
## ---------------------------------------------------------------
## unemp                       517.460***            215.409**
##                              (69.412)              (100.125)
##
## popdens                                            1.054***
##                                                    (0.275)
##
## Constant                   2,398.594***          3,513.062***
##                             (543.248)             (563.493)
##
## ---------------------------------------------------------------
## Observations                   53                    53
## R2                            0.521                 0.630
## Adjusted R2                   0.512                 0.615
## Residual Std. Error  1,240.129 (df = 51)    1,101.398 (df = 50)
## F Statistic         55.575*** (df = 1; 51) 42.557*** (df = 2; 50)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Interpretation: If unemployment increases by one percentage point, crime increases by 517 cases per 100,000 inhabitants (Model 1). The relationship is statistically

significant. If we now control for population density in Model 2 (and thus hold the influence of population density constant), a one-unit increase in unemployment is associated with an increase of only 215 cases. Population density is also positively associated with crime. Both coefficient estimates are statistically significant at p < 0.01. It thus seems to be a good idea to control for the influence of population density. We might be now much closer to the real effect of unemployment on crime now. However, we do not know this for certain because other alternative explanations are still possible to think of (e.g., the role of age composition or local social policies, as well as alternative explanations at the level of individuals).

Disclaimer: If this outlook has been too complex, no worries. We will go through all the steps in detail in the subsequent case studies.

Bivariate Statistics – Case Study United States Presidential Election

Introduction

Studying the results of the 2020 US presidential election holds profound relevance in terms of understanding contemporary American politics, as well as research on electoral behavior. Analyzing the voting patterns unveils the interplay of demographics, ideology, and socio-economic factors shaping political preferences. Such insights not only inform electoral strategies but also deepen our comprehension of societal divisions and cleavages. The 2020 election also stands out through an effort to overturn the election result by Donald Trump and co-conspirators. This has caused turmoil and abet the January 6 United States Capitol attack (e.g., see https://www.britannica.com/event/January-6-U-S-Capitol-attack and https://statesuniteddemocracy.org/resources/doj-charges-trump/).

Source: https://unsplash.com/de/s/fotos/us-election

Before the election, a team of researchers from the American National Election Study (ANES) conducted a large representative survey (based on a random sample of the US population) to study voting intentions. We will work with these data in this case study. We are particularly interested in which individuals are more likely to support Donald Trump compared to Joe Biden. In doing so, we will present data in the form of cross tabulations. We will also take a closer look at the topic of correlation analysis by investigating the relationship between average education and average Trum support across US states.

Data overview and descriptive analyses

The data are read from a CSV file (comma-separated values). By using the "head" command, the first six lines of the data are displayed.

```
data_us <- read.csv("data/data_election.csv")
ft2 <- flextable(head(data_us))
autofit(ft2)
```

vote	trump	education	age
trump	1	3	46
other		2	37
biden	0	1	40
biden	0	3	41
trump	1	4	72
biden	0	2	71

We can see from this that the following variables are included in the dataset:

`vote` = voting intention in the 2020 United States presidential election: "trump", "biden", "other", "refused or don't know".

`trump` = numerical recode of the `vote` variable. 1="trump", 0="biden", NA="other / refused / don't know".

`education` = respondent's highest educational qualification: 1=no degree or high school, 2=some collage, 3=Associate or Bachelor's degree, 4=Master's or postgraduate degree, NA=not specified

`age` = age of respondent in years.

Question: What is the scale of measurement (nominal, ordinal, metric) of each variable?

Solution:

`vote` and `trump` –> nominal (i.e., information on if something is either the case or not, categories cannot be ranked and have no numerical meaning) `education` –> ordinal (i.e., categories (or variable values) can be ranked, information on whether a variable value is "higher"/"more" or "lower"/"less") `age` –> metric (i.e., interval between ranked variable values can be compared; e.g., moving up 2 years from 20 to 22 is equivalent to moving up from 52 to 54)

Frequency tables

The following table shows the observed frequencies of the values of the variable `vote`.

```
dim(data_us) # total number of cases
```

```
## [1] 7272    4
```

```
table(data_us$vote)
```

```
##
##                biden                other refused / don't know
##                 3759                  274                   223
##                trump
##                 3016
```

We can see that 3,016 respondents chose Trump, whereas 3,759 opted for Biden. A difference of a couple of hundred responses. For a variable with only a few categories, looking at absolute numbers of observations might be informative. However, using relative frequencies in terms of proportions is typically more informative. Let's get to it!

```
prop.table(table(data_us$vote))
```

```
##
##                biden                other refused / don't know
##           0.51691419           0.03767877           0.03066557
##                trump
##           0.41474147
```

The relative difference between the group of Trump supporters and Biden supporters seems to be quite marginal in terms of percentage points. To be able to compare the figures from the survey data with the official vote result, we need to exclude the categories "refused (e.g., because respondent stated not going to vote) / don't know". To do so, we assign this category as "missing values".

```r
data_us$vote[data_us$vote == "refused / don't know"] <- NA # Set "refused / don't know" to
# NA (i.e., missing), so that this category is no longer displayed in the table command

prop.table(table(data_us$vote))
```

```
##
##      biden      other      trump
## 0.53326713 0.03887076 0.42786211
```

Question: How well did the results from the survey predict the actual election outcome? Can we conclude from that whether or not the sample is representative?

Solution:

In the 2020 US presidential election, Biden won the popular vote against Trump by 51.3% to 46.9%. The survey results indicate that at the time the survey was conducted, 53.3% would have voted for Biden (compared to 42.8% for Trump). While the survey correctly predicted Biden as the winner, the result deviated from the actual election outcome by about 2 percentage points. Inferring from samples to the underlying population is always subject to some uncertainty. We can quantify margins of uncertainty, and we will do so in the case study on statistical inference. Beyond the role of statistical uncertainty, several further factors (e.g., political campaigning or situational circumstances on voting day) might have contributed to the outcome of the election result and possibly account for the difference between the survey results and the actual vote.

The representativeness of a survey relates to its features, such as random sampling of respondents and no systematic non-response of those being interviewed. For the American Election Study, we have good reasons to assume that these features are given and the results are representative of the US population.

We will below also exclude the category "other candidates" that could have chosen in the survey and the election (together these candidates gathered less than 2% of the votes in the election). We will show later on how to recode variables. In the meantime, we can rely on the variable `trump` in which the categories "refused/don't know" and "other" have already been set to NA (*not available* which is equal to *missing values* in terms of data analysis).

```r
prop.table(table(data_us$trump))
```

```
##
##         0         1
## 0.5548339 0.4451661
```

A note on missing data and missing values

In survey data, missing data can occur through deliberate non-response of participants or other processes (e.g., a person has moved to a new address or is not eligible to vote). If the process that generated missing values is not known, this can potentially lead to biased results. A useful vignette on how to deal with missing data can be found here: https://cran.r-project.org/web/packages/finalfit/vignettes/missing.html

Regarding how to deal with missing values in data analyses, remember that in the variable `trump` the category "won't vote" was recoded to NA, which means declaring to the program to not include these cases in the statistical analysis. Note that for some functions in R, it is necessary to explicitly exclude missing values (e.g., using the option "na.rm = TRUE"). Other functions do this automatically.

For example, the *table* command excludes missing values automatically, and if you want to display the frequency of missing values, you have to specify "`exclude = NULL`". The *mean* command, however, does not

exclude missing values automatically and therefore we have to specify the option "`na.rm = TRUE`" (i.e., "*NA remove* = TRUE").

"Who supports Trump?" - Delving into bivariate statistics

In the following, we are interested in which groups of people were more likely to support Trump in the US presidential election. To do so, we focus on respondents' education and age. Education is measured with a few categories. Hence, setting up a cross tabulation or cross tab (= a table involving two variables) is appropriate. In contrast, age is measured in years and thus comprises many more categories. This would make a cross tab involving age rather cumbersome. We could recode age into a few age categories and use them in a cross tab or we could compare age distributions across voting intentions. Below, we will explore both options.

Education and Trump vote

First, we are interested in whether or not people's vote choice differs across education groups. If this is the case, we could state that education and voting intentions are statistically related or associated with each other. If not, both would be independent. Following conventions, the "variable to be explained" (a.k.a. outcome or dependent variable) is represented by the rows of the cross tab and the "explanatory variable" (a.k.a. causal factor or independent variable) by the columns. We calculate column percentages for the interpretation (i.e., the number of observations for each cell divided by the total number of observations for each column. Hence, adding all proportions per column adds up to 100%). In our case, the causal order is quite clear. We have good reason to believe that (if anything) `education` determines voting intentions (`trump`) and **not** the other way around (i.e., voting intentions would cause higher or lower education). Note that in many other cases, the causal order is not as clear and we can freely choose which variable goes in the rows and which in the columns.

```
data_us_nomissings <- na.omit(data_us) # We store a new version of the data set without
# missings, so that the subsequent "proc_freq" command only displays valid cases
# (otherwise another row with "missings" would be displayed)

proc_freq(data_us_nomissings, "trump", "education",
          include.row_percent = FALSE, include.table_percent = FALSE) # To only display
# column percentages, row and table (i.e., total) percent are set to "FALSE". Otherwise,
# the table would be a bit overloaded.
```

		education				
trump		1	2	3	4	Total
0	Count	555	676	1,438	936	3,605
	Col. pct	45.7%	50.4%	55.6%	71.0%	
1	Count	660	666	1,147	382	2,855
	Col. pct	54.3%	49.6%	44.4%	29.0%	
Total	Count	1,215	1,342	2,585	1,318	6,460

For column percentages, the interpretation begins by picking two categories (adjacent or extreme) of the explanatory variable and then comparing a relevant category of the outcome.

> ***Interpretation: Among those with low formal education (education=1), 54.8% of respondents support Trump, while in the highly educated group (education=4) only 29.5% support Trump. From this comparison, we can already conclude that education and Trump support are strongly related. The higher the education, the lower the preference for Trump, on average.***

Note that in the case of no systematic relationship between variables, the difference between two adjacent or opposite column values would be zero or close to zero (both values would closely correspond to the total proportion of 44.5%).

Important additions to the made interpretation would be (1) to quantify the strength of the empirical relationship (e.g., weak, medium, strong) and (2) to make a statement about how confident we can be that the empirical relationship found in the sample reflects the properties of the population the sample has been drawn from. There are procedures for both purposes, to which we will return to at the end of this case study. Before that, however, let's go back to cross tabs for a moment. Here, we see the same cross tab as above but with row percentages instead.

```
proc_freq(data_us_nomissings, "trump", "education",
          include.column_percent = FALSE,
          include.table_percent = FALSE) # Here, column and total (table) percentages are
                                         # omitted which leaves us with row percentages
```

		education				
trump		1	2	3	4	Total
0	Count	555	676	1,438	936	3,605
	Row pct	15.4%	18.8%	39.9%	26.0%	
1	Count	660	666	1,147	382	2,855
	Row pct	23.1%	23.3%	40.2%	13.4%	
Total	Count	1,215	1,342	2,585	1,318	6,460

Question: Which conclusions can be drawn from a cross tab with row percentages?

Solution:

Row percentages are the number of observations for each cell is divided by the total number of observations for each row (adding all proportions per row adds up to 100%). The interpretation would be similar to before, except that we now focus on comparing different categories of the row variable for one selected value of the column variable. Let's select education = 1: Among those who do not support Trump (= Biden supporters), 15.3% have a low education profile, while among Trump supporters 23.1% have a low education profile. The difference of 7.8 percentage points suggests that Trump support and low education are systematically associated. Hence, the conclusion from the cross tab analysis with row percentages is equivalent to that one from column percentages (given that we adjust the interpretation scheme). Moreover, row percentages in this case also show the descriptive *distribution* of the education variable within each category of the vote variable. However, it is often not easy to compare the distributions at one glance.

As a third possibility, below a cross tab with total percentages:

```
proc_freq(data_us_nomissings, "trump", "education",
          include.column_percent = FALSE,
          include.row_percent = FALSE) # Here, column and row percentages are omitted
                                       # which leaves us with total percentages
```

	education				
trump	1	2	3	4	Total
0	555 (8.6%)	676 (10.5%)	1,438 (22.3%)	936 (14.5%)	3,605 (55.8%)
1	660 (10.2%)	666 (10.3%)	1,147 (17.8%)	382 (5.9%)	2,855 (44.2%)
Total	1,215 (18.8%)	1,342 (20.8%)	2,585 (40.0%)	1,318 (20.4%)	6,460 (100.0%)

Question: Which conclusions can be drawn from a cross tab with cell percentages?

Solution:

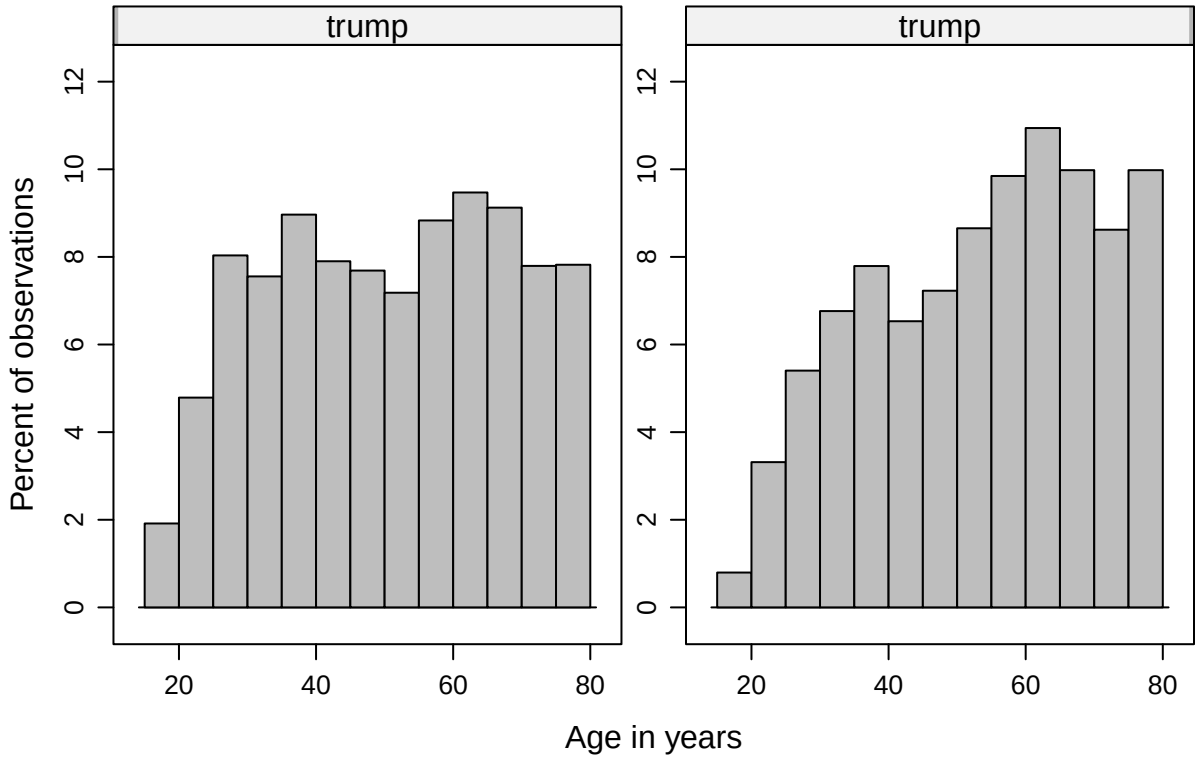
Total percentages are the number of observations for each cell divided by the total number of observations. This shows the relative frequency of each combination of values of the two variables. The goal is a descriptive understanding of the distribution of cases. (Note: This is rarely used in applied research.)

Age and Trump vote

Due to the 'age' variable being recorded in years and thus having a large number of categories, a cross tab would be quite cumbersome and thus less informative. Therefore, we first create histograms displaying the distribution of the age variable for both categories of `trump` and then compare them with each other.

```
histogram( ~ age | trump,
           breaks = 10,
           ylim=c(0, 12),
           type = "percent",
           main = "Left: trump=0 (support of Biden), Right: trump=1 (support of Trump)",
           ylab = "Percent of observations",
           xlab = "Age in years",
           layout = c(2, 1),
           scales = list(relation = "free"),
           col = 'grey',
           data = data_us)
```

Left: trump=0 (support of Biden), Right: trump=1 (support of Trump)



Question: What can be inferred from the graphs?

Solution:

The age distribution of the respondents supporting Biden (left-hand) is a bit more even than the distribution of those supporting Trump (right-hand side). The latter is also more skewed to the left and its center is more on the right, which indicates that respondents who support Trump are older, on average, than those who support Biden.

In addition, here are the means and standard deviations of age by vote group.

Age mean, median, and standard deviation for Biden supporters:

```
mean(data_us$age[data_us$trump == 0], na.rm = TRUE)
```

```
## [1] 51.22253
```

```
median(data_us$age[data_us$trump == 0], na.rm = TRUE)
```

```
## [1] 52
```

```
sd(data_us$age[data_us$trump == 0], na.rm = TRUE)
```

```
## [1] 17.22161
```

Age mean, median, and standard deviation for Trump supporters:

```
mean(data_us$age[data_us$trump == 1], na.rm = TRUE)
```

```
## [1] 54.2871
```

```
median(data_us$age[data_us$trump == 1], na.rm = TRUE)
```

```
## [1] 56
```

```
sd(data_us$age[data_us$trump == 1], na.rm = TRUE)
```

```
## [1] 16.51148
```

By comparing the mean and median values of both groups, we find that the Trump-supporting group is older than the group that supports Biden. The standard deviation (which can be thought of as the average deviation between the age of respondents and the mean) in the Biden group is slightly larger (17.2 years) than the standard deviation of the Trump group (16.5 years). This means that the age of respondents in the Biden group is slightly more dispersed than in the Trump group.

In the next step, we recode the age variable into the categories "young" (1), "middle-aged" (2), and "old" (3). We then use a cross tab with column percentages.

```
data_us <- data_us %>%
  mutate(age_cat =
           case_when(age <= 35 ~ 1,
                     age > 36 & age <= 59 ~ 2,
                     age > 59 ~ 3))


data_us_nomissings <- na.omit(data_us) # Exlude missings again by overwriting with the new
                                       # dataset that contains the recoded age variable

proc_freq(data_us_nomissings, "trump", "age_cat",
          include.row_percent = FALSE, include.table_percent = FALSE)
```

		age_cat			
trump		1	2	3	Total
0	Count	831	1,363	1,349	3,543
	Col. pct	63.2%	55.7%	52.3%	
1	Count	484	1,083	1,230	2,797
	Col. pct	36.8%	44.3%	47.7%	
Total	Count	1,315	2,446	2,579	6,340

Question: Interpret the cross tab. What can be concluded? Why don't we always go straight to categorizing metric variables and using the recoded variable in a cross tab?

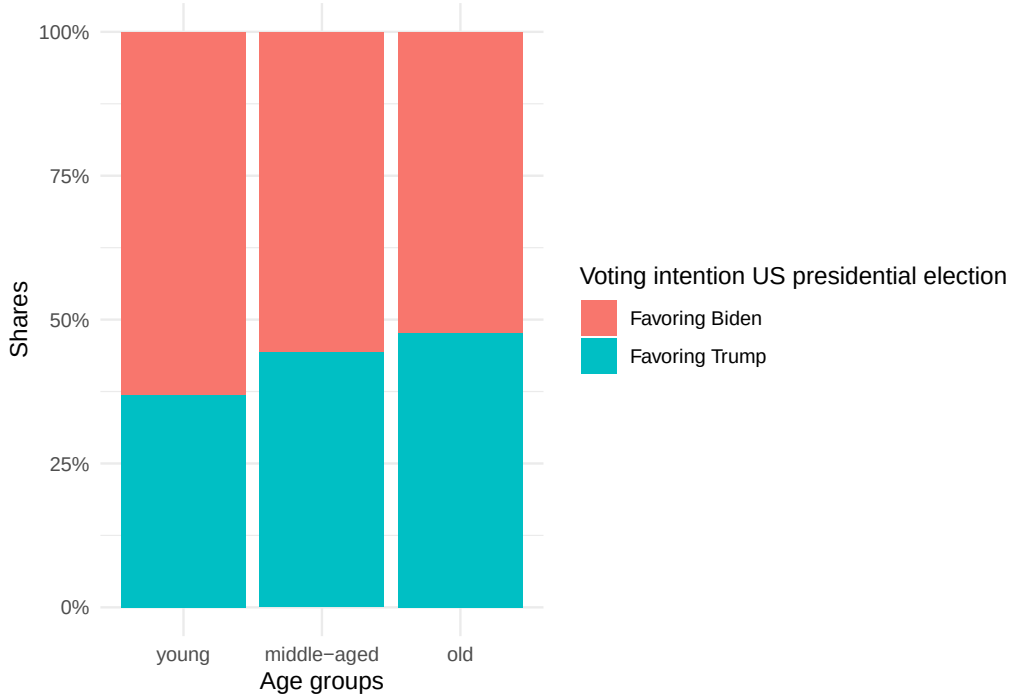
Solution:

Looking at the category of young respondents (age_cat = 1), 36.8% prefer Trump as president, whereas among old respondents (age_cat = 3), 47.7% opt for Trump. This difference of 10.9 percentage points indicates a systematic (positive) empirical relationship between age and Trump support: Older respondents tend to prefer Trump more than younger voters.

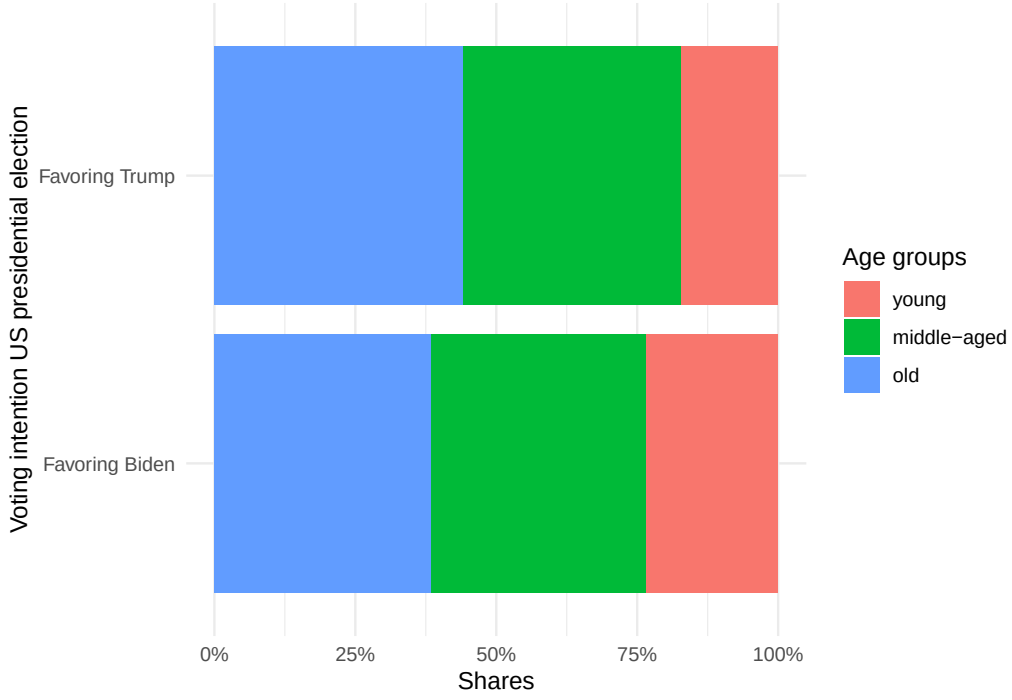
The reason why it is not always recommended to categorize metric variables is that the process of categorization involves a loss of information. Usually, more information is better as it produces more detailed results. Note that there are exceptions to this such as striving for a simplified description of empirical relationships.

Here is a way to visualize a cross tab with stacked bar charts. The first figure corresponds to a cross tab with column percentages and the second figure relates to a cross tab with row percentages.

```
# Age variable will be subdivided into the three dummy variables "young", "middle-aged",
# and "old"
data_us <- data_us %>%
  mutate(age_cat_nom =
           case_when(age <= 35 ~ "young",
                     age > 36 & age <= 59 ~ "middle-aged",
                     age > 59 ~ "old"))

data_us$age_cat_nom <- factor(data_us$age_cat_nom,
                              levels = c('young', 'middle-aged', 'old'))

# Recoding of electoral participation
data_us <- data_us %>%
  mutate(trump_nom =
           case_when(trump == 0 ~ "Favoring Biden",
                     trump == 1 ~ "Favoring Trump"))

data_us_counted <- data_us %>% count(age_cat_nom, trump_nom)
data_us_counted <- subset(data_us_counted,
                          !is.na(data_us_counted$age_cat_nom)) # removal of missing values
data_us_counted <- subset(data_us_counted,
                          !is.na(data_us_counted$trump_nom)) # removal of missing values


ggplot(data_us_counted, aes(fill = trump_nom, y = n, x = age_cat_nom)) +
    geom_bar(position = "fill", stat = "identity") +
    scale_y_continuous(labels = scales::percent) +
    labs(x = "Age groups",
         y = "Shares",
         fill = "Voting intention US presidential election") +
    theme_minimal()
```

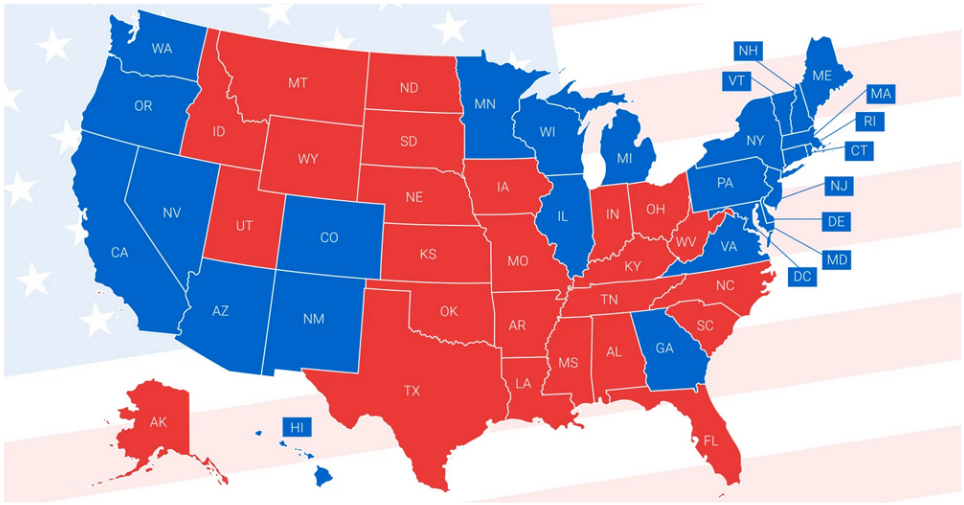



```
ggplot(data_us_counted, aes(fill = age_cat_nom, y = n, x = trump_nom)) +
    geom_bar(position = "fill", stat = "identity") +
    scale_y_continuous(labels = scales::percent) +
    coord_flip() +
    labs(x = "Voting intention US presidential election",
         y = "Shares",
         fill = "Age groups") +
    theme_minimal()
```



Scatter plot and correlation

Scatter plots and correlation are suitable for showing relationships between two variables. In the following, we use the average approval of Trump across US states in % (`perc_trump`) as the variable to be explained (a.k.a. the outcome or dependent variable). State-specific variations in voting patterns are illustrated in the following map. As an explanatory variable, we use the proportion of people in the state who have a high level of education (Master's or postgraduate) in % (`perc_higheducation`). Both variables originate from individual-level data of the ANES study and have been aggregated to the state level (i.e., the means per region were stored in a new dataset).



Regional US presidential election results

Source: https://www.governing.com/assessments/what-painted-us-so-indelibly-red-and-blue

Next, the data are read and the first six rows are displayed:

```
data_states <- read.csv("data/data_states.csv")
ft3 <- flextable(head(data_states))
autofit(ft3)
```

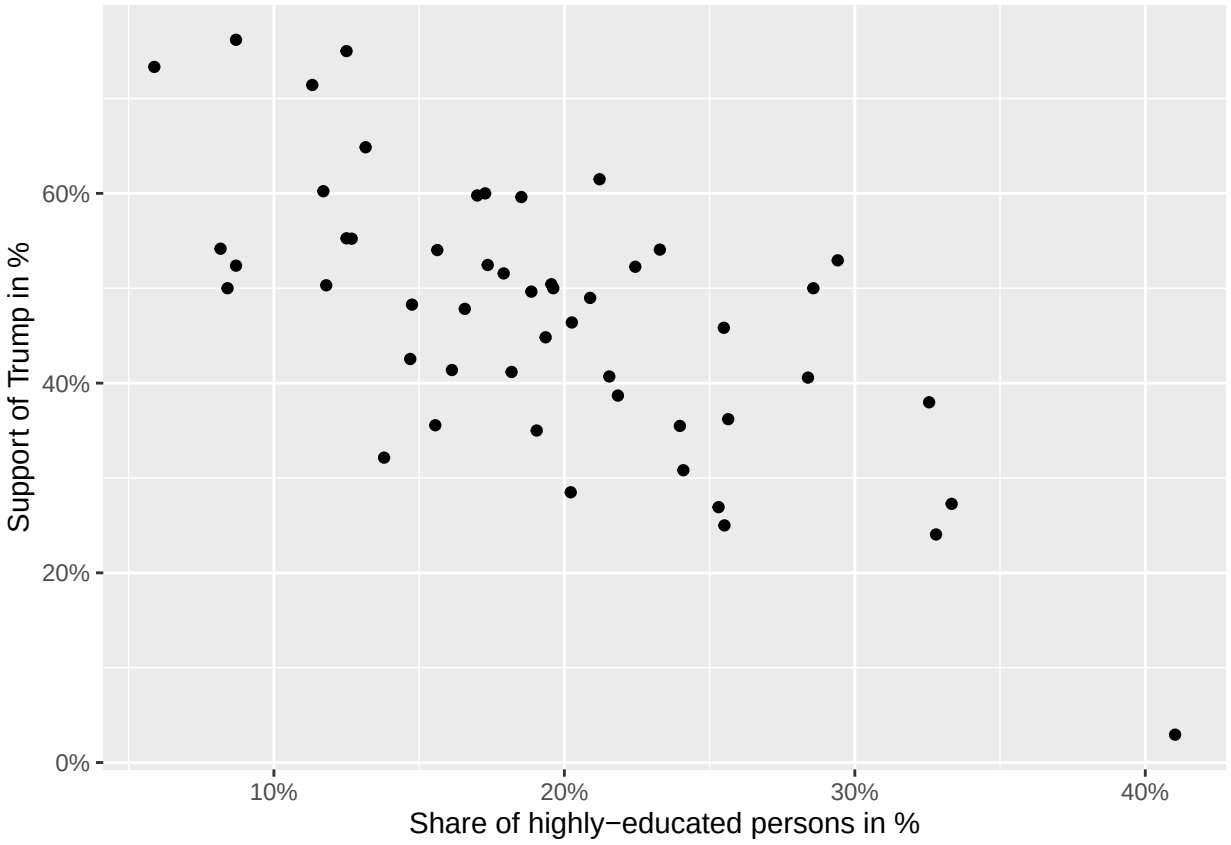
state	perc_trump	perc_higheducation
46. South Dakota	0.7333333	0.05882353
31. Nebraska	0.5416667	0.08163265
45. South Carolina	0.5000000	0.08403361
38. North Dakota	0.7619048	0.08695652
30. Montana	0.5238095	0.08695652
5. Arkansas	0.7142857	0.11320755

The variable `state` identifies the region.

Scatter plot

To get a first impression of the empirical relationship between both variables, we use a scatter plot. Each point represents one of the 50 US states and Washington D.C. Note that it is a convention that the dependent variable is shown on the y-axis, whereas the independent variable is shown on the x-axis.

```r
sc1 <- ggplot(data=data_states, aes(x = perc_higheducation, y = perc_trump)) +
  geom_point() +
  xlab("Share of highly-educated persons in %") +
  ylab("Support of Trump in %") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_continuous(labels = scales::percent)

sc1
```



Question: What can be inferred from the figure?

Solution:

The pattern indicates a *negative* relationship between the two variables. If the proportion of highly educated people in a region increases (moving to the right on the x-axis), we see that this is related to *lower* support for Trump (lower scores on the y-axis).

Correlation

In the next step, we will take a graphical approach to how a correlation is determined. For this, we plot a vertical and a horizontal line that represents the mean of the variables through the cloud of observations represented by the black dots. We see that the dots are mainly in the upper-left quadrant (= low proportion of high education AND high Trump approval) and the lower-right quadrant (= high proportion of high education AND low Trumo approval). This indicates that higher average education tends to be associated with *lower* average Trump approval. In other words: The correlation is negative.

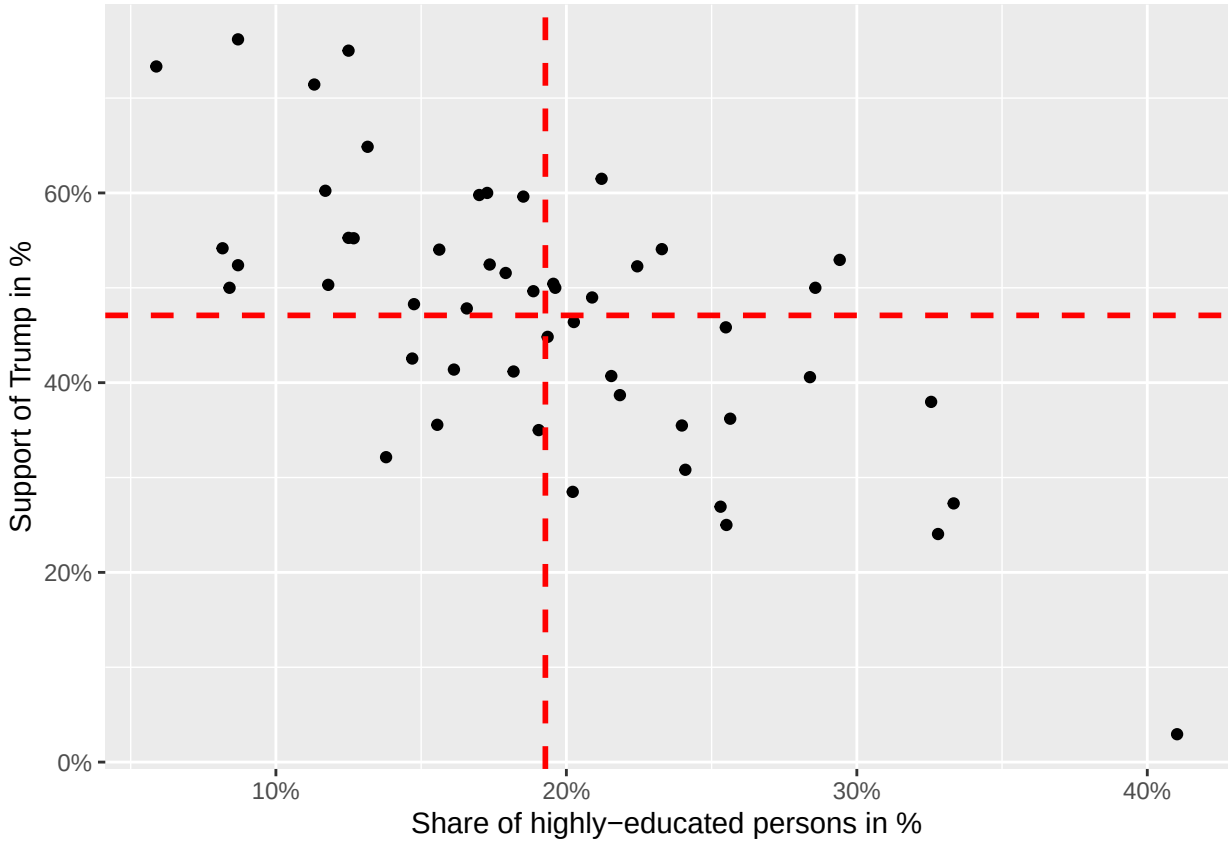
```r
# Obtaining means for each variable
mean(data_states$perc_trump, na.rm = TRUE)
```

```
## [1] 0.470922
```

```r
mean(data_states$perc_higheducation, na.rm = TRUE)
```

```
## [1] 0.1927645
```

```r
# Scatter plot using percentages as scale units
sc2 <- ggplot(data=data_states, aes(x = perc_higheducation, y = perc_trump)) +
  geom_point() +
  xlab("Share of highly-educated persons in %") +
  ylab("Support of Trump in %") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_continuous(labels = scales::percent) +
  geom_hline(yintercept = 0.470922, linetype = "dashed", color = "red", size = 1) +
  geom_vline(xintercept = 0.1927645, linetype = "dashed", color = "red", size = 1)

sc2
```



To quantify the relationship, we can calculate the covariance and subsequently the correlation (= standardized covariance fitted into the value range of -1 to +1). The widely used Pearson's correlation coefficient is given as:

$r = \frac{Covariance(x,y)}{SD(x) * SD(y)}$

$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

The numerator $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is particularly important. It represents the **covariance**, which is an *unstandardized* measure of association between x and y because its value heavily depends on the scales x and y are measured. At the same time, it serves to determine the direction of the empirical association. To do so, the classification of observations into above/below the mean quadrants is useful again. Observations that are above both means (i.e., upper right corner) have a positive sign and contribute as a product to a "positive" covariance or correlation. The same is true for observations in the lower left quadrant because here both differences have a negative sign, which again contributes as a product to a positive covariance (and thus correlation). In contrast, observations in the upper left and lower right quadrants contribute to a negative covariance (and thus correlation). If observations are evenly distributed across all quadrants, positive and negative products cancel each other out, which would lead to a zero covariance (and thus correlation). The step from covariance to correlation is that the covariance is *standardized* by the standard deviation of x and y which removes the scale units covariance is measured on and transforms the measure to range between -1 and +1.

Here is an overview of possible scenarios in the correlation coefficient:

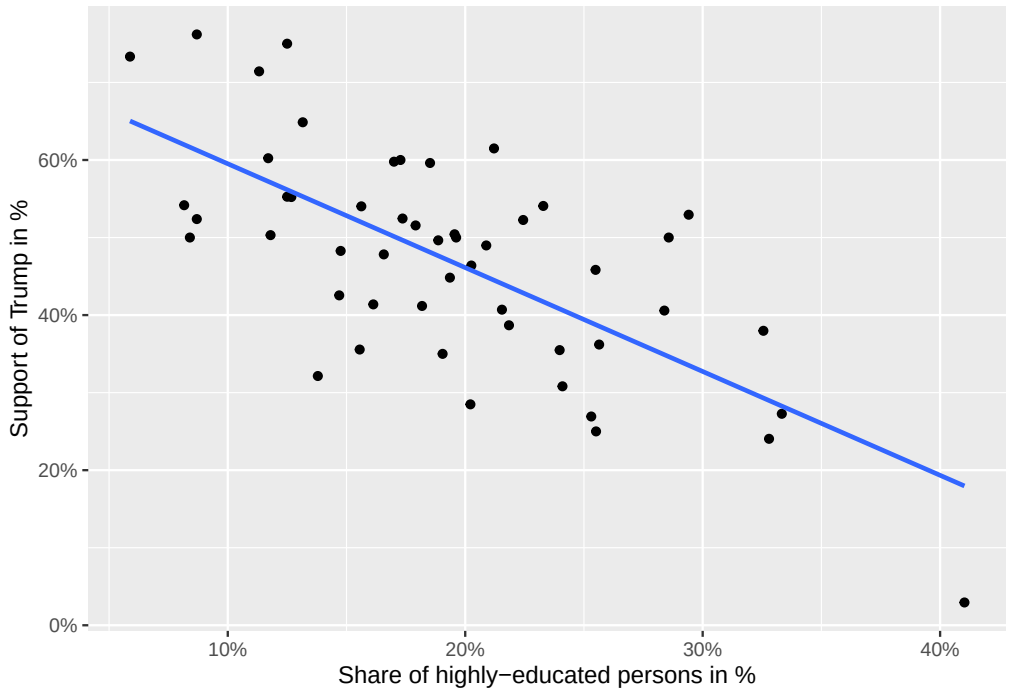
Different correlations

Source: https://en.wikipedia.org/wiki/Correlation#/media/File:Correlation_examples2.svg

Scatter plots showing correlations (from left to right) of -1, -0.8, -0.4, 0, 0.4, 0.8 and 1.

In the next step, the correlation between the share of highly educated people and the share of Trump supporters is calculated and a line representing their relationship is plotted. The line essentially depicts the bivariate association in the form of a regression slope, which always runs in the direction of the correlation. However, its interpretation differs from correlation.

```
sc3 <- ggplot(data=data_states, aes(x = perc_higheducation, y = perc_trump)) +
  geom_point() +
  xlab("Share of highly-educated persons in %") +
  ylab("Support of Trump in %") +
  scale_y_continuous(labels = scales::percent) +
  scale_x_continuous(labels = scales::percent) +
  geom_smooth(method = lm, se = FALSE)

sc3
```




```
vars <- c("perc_higheducation", "perc_trump")
cor.vars <- data_states[vars]
rcorr(as.matrix(cor.vars))
```

```
##                    perc_higheducation perc_trump
## perc_higheducation               1.00      -0.69
## perc_trump                      -0.69       1.00
##
## n= 51
##
##
## P
##                    perc_higheducation perc_trump
## perc_higheducation                     0
## perc_trump          0
```

In the case of correlation and regression coefficients, the interpretation always begins with mentioning x (or the **independent** variable) which exerts an effect on y (or the **dependent** variable).

- **Correlation coefficient:** "If x increases, then y increases / decreases (depending on the sign of the correlation coefficient), on average. The magnitude of the correlation coefficient r indicates a *weak/moderate/strong* empirical association." (Rule of thumb: +/- 0.1 *weak*, +/- 0.3 *moderate*, +/- 0.5 or higher *strong* association)
- **Regression coefficient:** "If x increases by one unit, then y increases / decreases (depending on the sign of the regression coefficient) by *coefficient* units, on average." The specific coefficient estimate that is referred to, is given in the regression output. To what extent the empirical relationship is strong, moderate, or weak can be determined by a standardization of the regression coefficients. Note that the details of regression analysis and its interpretation are covered in the case study on regression analysis.
- Correlation is a standardized form of covariance, which comes with the advantage of being intuitively interpretable (-1 to +1) without any reference to the underlying scale the two variables are measured. However, the price to pay is that it conceals the *magnitude* by which y changes if x changes by one unit (rather it represents how dense observations cluster together in forming a linear relationship). Bivariate

regression quantifies the expected increase in y if x increases by 1. We thus need to standardize (or "correct") the covariance between x and y for the scale on which x is measured. This is achieved by dividing the covariance by the variance of x: $b = \frac{Covariance(x,y)}{Var(x)}$

Interpretation of the correlation coefficient from the example: The correlation between the proportion of people with high education and Trump support is negative. If the share of people with high education increases, then Trump support decreases, on average. The correlation coefficient of -0.69 indicates a strong empirical relationship between both variables.

Chi-square and Cramér's V

Now we return to cross tabs and how to make a statement about whether there is a correlation between the displayed variables and if so, how strong it is. To do so, we once again revisit the individual-level survey data and display the cross tab between `education` and `trump`.

```
proc_freq(data_us, "trump", "education",
          include.row_percent = FALSE, include.table_percent = FALSE, na.rm = FALSE)
# To only display column percentages, row and table (i.e., total) percent has to be set
# to "FALSE". Otherwise, the table would be a bit overloaded
```

		education					
trump		1	2	3	4	Missing	Total
0	Count	567	687	1,483	974	48	3,759
	Col. pct	42.1%	46.3%	51.7%	66.6%	43.2%	
1	Count	687	685	1,195	407	42	3,016
	Col. pct	51.0%	46.2%	41.7%	27.8%	37.8%	
Missing	Count	93	111	191	81	21	497
	Col. pct	6.9%	7.5%	6.7%	5.5%	18.9%	
Total	Count	1,347	1,483	2,869	1,462	111	7,272

It has already become quite clear that low-educated people were more likely to support Trump compared to high-educated people (who were more in favor of Biden). To quantify the empirical association, we can calculate Cramér's V. This measure quantifies empirical relationships if a nominal variable is involved (here: `trump`). It is based on a chi-square test for independence. The chi-square value indicates how strongly the real observations deviate from a theoretical distribution, wherein all observations are equally distributed (relative to marginal distribution) across the cells of the cross tab. The more the measured observations deviate from the theoretical distribution, the higher the chi-square value and subsequently Cramér's V gets, which is normalized for the table size and ranges between 0 (= no association) and 1 (= exceedingly strong association).

The formula for the chi-square test for cross tabs is the following: $Chi^2 = \sum \frac{(O-E)^2}{E}$

O refers to the observed frequencies and E to the expected ones. The expected frequencies are obtained by multiplying the marginal frequencies ("total") for each cell, divided by the total number of cases ($E = \frac{R \times C}{n}$). For the first cell in the table above, this would be (1254*3711)/6685=696.12. Thus, 696 observations would be theoretically expected in the first cell and 567 were observed. Quite different. This is done (usually through statistical software) for each cell and summed up according to the formula.

Cramér's V is a measure based on Chi-square and adjusted for different-sized cross tabs. It is given by the formula: $V = \sqrt{\frac{Chi^2}{n \times (m-1)}}$, where n is the number of observations and M is the number of categories (or

variable values) of the variable with the fewer categories (here `trump` with 2).

Regarding interpretation, the following guidelines apply:

- Cramér's V ranges from 0 to 1, with no negative values possible
- Hence, Cramér's V provides no information on whether the association is positive or negative, we have to figure this out from the table (e.g., by calculating percentage differences between column values –> see above)
- Rules of thumb regarding the size of the relationship:
  - In a small table (e.g., 2x2): between 0.1 and 0.3 –> weak, more than 0.3 and less than 0.5 –> moderate, more than 0.5 –> strong association
  - In a large table (e.g., 5x5): between 0.05 and 0.15 –> weak, more than 0.15 and less than 0.25 –> moderate, more than 0.25 –> strong association

In R, chi-square and Cramér's V can be calculated with the following commands:

```
chisq.test(data_us$trump, data_us$education)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_us$trump and data_us$education
## X-squared = 196.39, df = 3, p-value < 2.2e-16
```

```
cramerV_tabelle <- table(data_us$trump, data_us$education)
cramerV(cramerV_tabelle)
```

```
## Cramer V
##   0.1714
```

Question: How can Cramér's V be interpreted in this case?

Solution:

> ***Interpretation: There is an association between the variables `education` and `trump`. Given the rule of thumb for small tables, the association of 0.17 is weak. In our case, we know from the interpretation above that the association is negative (the higher the education, the lower the approval of Trump, on average).***

Question: Why didn't we calculate the Pearson's correlation coefficient instead?

Solution:

Pearson's correlation coefficient requires (quasi-)metric variables with several categories. For ordinal variables, we have tools such as Kendall's Tau or Spearman's correlation coefficient (see here for more on different correlation types: https://ademos.people.uic.edu/Chapter22.html). If nominal variables are involved, we use coefficients such as Cramér's V.

Statistical Inference - Case Study Satisfaction with Government

Introduction

In Western democracies, citizens' satisfaction with the government is critically related to the electoral success of governing parties in an upcoming election. At the same time, satisfied citizens are more willing to support non-preferred policies and are less likely to support anti-system or populist parties.

If we are interested in the degree of satisfaction with the government in Germany, for example, we first need to define the population we aim to make statements about (e.g., people over 18 years old who live in Germany). Conducting interviews with each person that belongs to the population would be pointless. Instead, we can rely on the "shortcut" of random sampling, in which we randomly select observational units from this population (i.e., randomly selected Germans) and survey them. The idea is that the results (e.g., mean approval of the government using a survey item) from the random sample are representative of the underlying population - that is, a close match between the parameter estimated from the sample (e.g., mean approval) and the (unobserved) parameter of the population.

Federal Chancellery (Bundeskanzleramt)

Source: https://www.bundesregierung.de/breg-de/bundesregierung/bundeskanzleramt/dienstsitze-bundeskanzleramt-466834

In statistical inference, we are primarily interested in how accurately an estimate from a sample reflects the true population parameter. Thus, the goal is to find a suitable measure of precision (or uncertainty, as the other "side of the coin"). We do this by matching the estimate from the sample with a test distribution that reflects likely or unlikely values in terms of probabilities. How this is done will be discussed in this case study.

Before that, we take a short excursion into the realm of probability theory.

Probability

Probabilities correspond to the ratio of the number of observations that meet a certain criterion (i.e., an "event") to the total number of observations. A classic example is the coin toss and the ratio between heads and the total number of tosses. Probabilities are always positive ($P(A) \geq 0$). The probability of all possible events (e.g., heads and tails in the case of a coin toss) is 1 ($P(\Omega) = 1$). If events A and B cannot occur simultaneously, then the probability of A or B occurring is the summed probability for both events ($P(AorB) = P(A) + P(B)$).

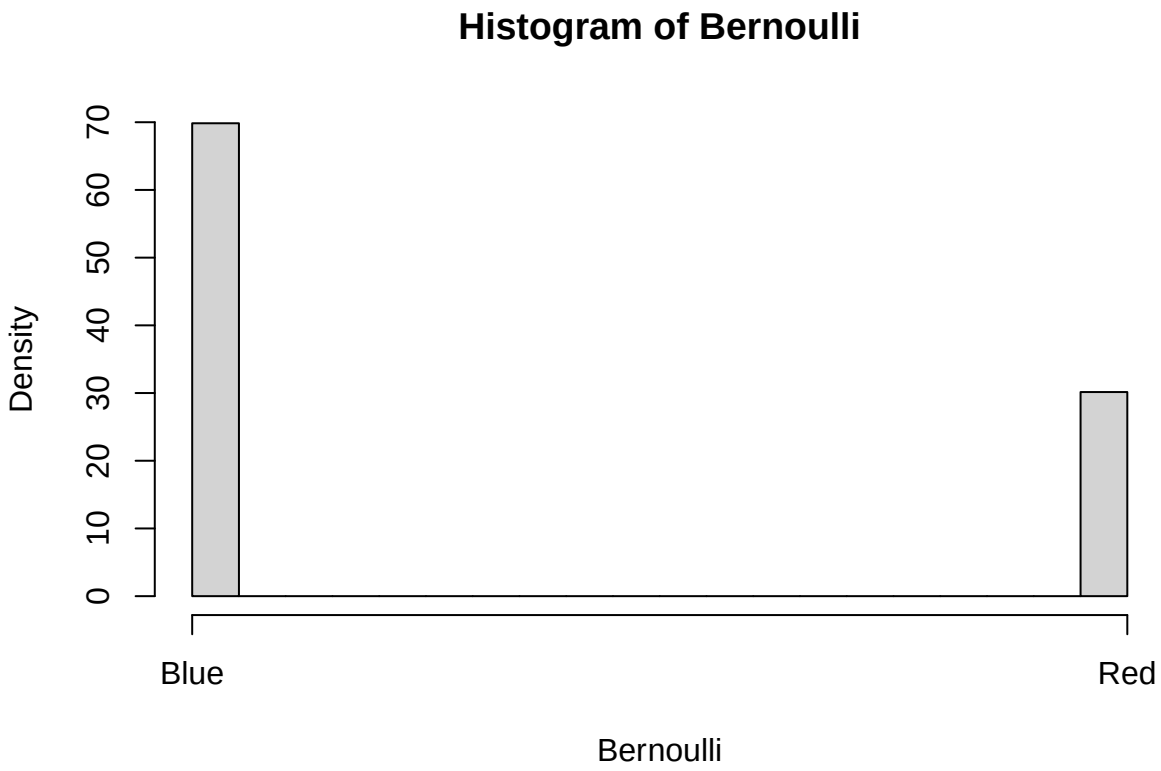
The probabilities of events that occur in a chance experiment (e.g., a coin toss) can be represented as so-called random variables. Random variables have a probability distribution, for binary variables (0/1) this is the Bernoulli distribution, for continuous variables this is the normal distribution. (There are other distributions like the F-distribution or chi-square distribution, which we will use later for further purposes).

Here we see a probability distribution for 10 balls (7 blue, 3 red) randomly drawn with replacement, in total 10,000 draws were made. Accordingly, the probability distribution is approximately 0.7 for blue and 0.3 for red:

```r
possible_values <- c(1, 0)
Bernoulli <- sample(possible_values,
                    size = 10000,
                    replace = TRUE,
                    prob = c(0.3, 0.7))
prop.table(table(Bernoulli))
```

```
## Bernoulli
##      0      1
## 0.6985 0.3015
```

```r
h <- hist(Bernoulli, plot = FALSE)
h$density = h$counts / sum(h$counts) * 100
plot(h, freq = FALSE, axes = FALSE)
axis(1, at = c(0, 1), labels = c("Blue", "Red"))
axis(2, at = c(0, 10, 20, 30, 40, 50, 60, 70))
```

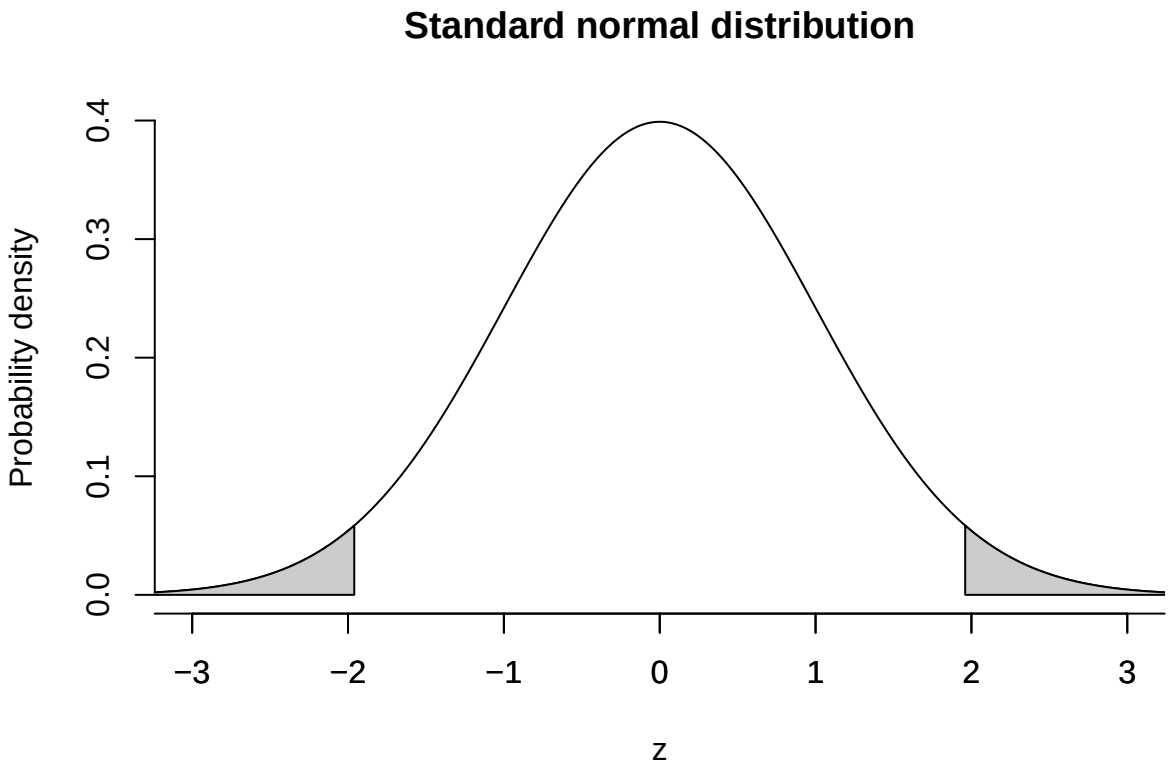


The Bernoulli distribution is characterized by the probability of occurrence of an event p. The probability for the counter-event automatically follows from this: $1 - p$. **We will need this distribution for the calculation of test statistics of proportions.**

The normal distribution for random variables is a bell-shaped probability density function. It gives the probabilities for a **continuous random variable** with infinite intermediate values. Think of time as a continuous concept. Theoretically, we can measure time in infinitely small units (milliseconds, nanoseconds, etc.). However, in applied research, we would assign discrete values to the variable time. Hence, continuous variables are rather a theoretical concept. Regarding the distribution of a continuous random variable: The normal distribution is characterized by a mean μ and a standard deviation σ. The notation for the normal distribution of a random variable is $X \sim N(\mu, \sigma)$. Each combination of μ and σ gives a differently shaped normal distribution. The mean indicates where the center of the distribution is on the x-axis and σ indicates how flat or steep the distribution is (the higher, the more spread out and thus the flatter the curve becomes).

The so-called *standard normal distribution* has a mean of 0 and a standard deviation of 1 ($X \sim N(0, 1)$). As with all other normal distributions, (1) the ends of the curve spread out to the left and right, approaching the x-axis without ever touching it. Thus, the values can range from - to + infinity. (2) The area under the curve is equal to 1 in total. (3) The so-called "Empirical Rule" states that for a standard normal distribution, 68% of the observations lie between -1 and +1 standard deviations. 95% of the observations lie between approx. -2 and +2 standard deviations (to be exact -1.96 and +1.96), and 99.7% of the observations lie between approximately -3 and +3 standard deviations.

```
# Draw a standard normal distribution:
z = seq(-4, 4, length.out = 1001)
x = rnorm(z)
plot(x = z, y = dnorm(z), bty = 'n', type = 'l', main = "Standard normal distribution",
     ylab = "Probability density", xlab = "z", xlim = c(-3, 3))
axis(1, at = seq(-4, 4, by = 1))

# annotate the density function with the 5% probability mass tails
polygon(x = c(z[z <= qnorm(0.025)], qnorm(0.025), min(z)),
        y = c(dnorm(z[z <= qnorm(0.025)]), 0, 0),
        col = grey(0.8))
polygon(x = c(z[z >= qnorm(0.975)], max(z), qnorm(0.975)),
        y = c(dnorm(z[z >= qnorm(0.975)]), 0, 0),
        col = grey(0.8))
```



With probability distributions such as the standard normal distribution, we do not calculate the probability of a specific value but rather a range of values. For example, the probability that a value is greater than 1.96 standard deviations, i.e., $P(Z \geq 1.96)$, is 2.5%. And the probability that a value is less than -1.96 standard deviations, i.e., $P(Z \leq -1.96)$, is 2.5% (see the grey areas in the plot above, both areas sum up to 5%).

We will use the standard normal distribution (also referred to as Z-distribution) for various statistical tests. The rule that 95% of observable values are located between -1.96 and 1.96, and 5% at the outer margins will also become important in Null Hypothesis Testing.

Basics of statistical inference

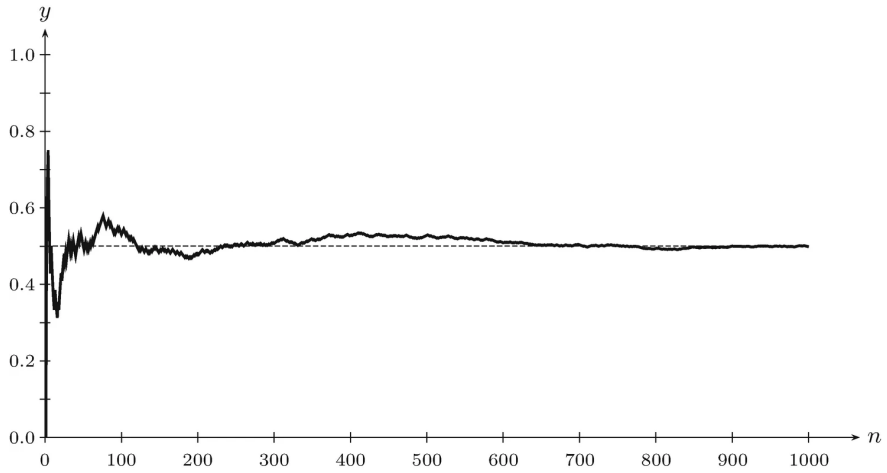
Inference means that we use facts we know to learn about facts we don't know (yet). Statistical inference means that we take information from a (random) sample and infer from it (not yet known) properties of the population (from which the sample was drawn). Basically, only random samples are suitable for this purpose, since they represent the underlying population in all properties in the best possible way (plus minus some random variation).

While random samples are the best way to conduct statistical inference, the process is not flawless. There is still a chance that the units in our sample do not perfectly match all properties of the population due to random deviations. It may be that a sample was drawn in which disproportionately many women accidentally "slipped in". These random deviations are called **sampling variability**. The following sections should illustrate that sampling variability decreases with the number of observations in a sample ("more data is better") and that if we did the process of random sampling over and over again (with a sufficiently large number of observations), we would obtain a distribution of sampled estimates that are shaped like a normal distribution. This will be tremendously useful when we conduct statistical tests, in which we use the normal distribution as if it was the sampling distribution of a quantity of interest we can match and test our estimate with.

Law of Large Numbers

In general, a small number of cases leads to higher sampling variability. You may be familiar with this from the coin toss experiment. A coin is often tossed in succession. After 10 tosses, it is quite possible that heads were tossed 7 times and tails only 3 times. Nevertheless, the underlying probability is always 0.5/0.5 (or 50%/50%) and in the population of all coin tosses, heads and tails should occur equally often. If we now toss the coin even more often, after 30, 300, 3000, etc. the observed coin tosses will increasingly match the 0.5/0.5 probability distribution.

The law of large numbers states that as the sample size increases, the mean of the sample will increasingly approxiamte that of the population. (Eventually, the sample would be identical to the entire population.) This means that more data is better, and the larger our sample, the more confident we can be that it matches or at least approximates the population parameter.



Law of large numbers: Coin toss

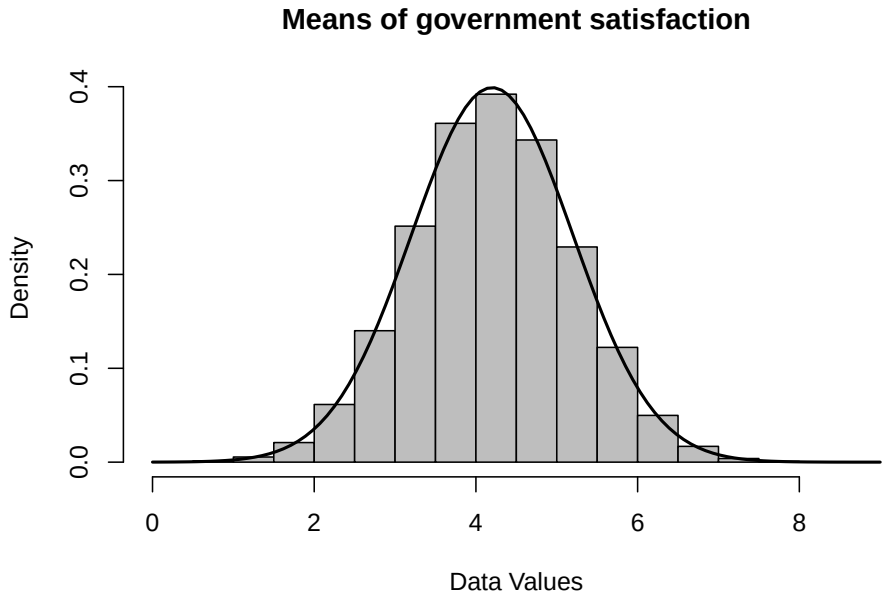
Source: https://link.springer.com/chapter/10.1007/978-3-030-45553-8_5

Central Limit Theorem

If we draw many (or even infinite) samples from a population and are interested in, for example, the mean of satisfaction with government, then plotting the means of each sample as a histogram will lead to a distribution that (with added samples increasingly) follows a normal distribution. The mean of the sample means thereby corresponds to the population parameter we are essentially interested in (i.e., satisfaction with the government of all over-18-year-olds living in Germany).

The implications of the central limit theorem are illustrated in the following figure. The mean of the means from 100,000 simulated samples is 4.2 and corresponds to the "true" population mean of government satisfaction. Note that in reality, we do not have this information and we are not able to draw these many samples as that would be too costly.

```
set.seed(123)
data <- rnorm(100000, 4.2, 1)
hist(data, freq = FALSE, col = "grey",
     xlab = "Data Values", main = "Means of government satisfaction")
curve(dnorm(x, mean = mean(data), sd = sd(data)), col = "black", lwd = 2, add = TRUE)
```



Another property of the central limit theorem is that the larger the number of cases (n) of each of the individual samples, the better the distribution of sample estimates corresponds to a normal distribution. This works very well with $n > 30$. Moreover, it is not important how the distribution of the characteristic to be examined is in the population (e.g., right- or left-skewed). In any case, with increasing samples (and $n > 30$), the distribution of sample estimates will (with added samples increasingly) correspond to a normal distribution.

Why statistical laws are important

Only one sample

- We calculate everything on the basis of **one sample**, everything else would be too costly.
- We can never be completely sure that the ONE specific sample and the statistics estimated using it correspond to those of the population, but:
  - The **larger the sample**, the steeper the distribution of sample estimates and the more certain we are to get an estimated parameter **near the population parameter**, even with only one sample (Law of Large Numbers).

– For a **sample of 30 or more observations**, we know that the distribution of sample estimates is normal no matter how the distribution of the population looks like; this gives us the rationale that most samples mingle around the population mean and that there is quite a high probability that the estimate from our one sample **is near the center of the distribution (i.e., close to the population parameter)**; getting an estimate that is far away from the center of the distribution is rather unlikely (Central Limit Theorem).

A Primer on Null Hypothesis Testing

- The Central Limit Theorem also gives us a rationale for **quantifying uncertainty**, because if repeated sample estimates can be represented as a **normal distribution**, then we can use the normal distribution as a frame of reference for how credible the results from our one sample are.
- To do so, we make specific assumptions about the distribution we test our sample estimate against (e.g., a distribution that assumes no difference in means between two groups).
- This procedure is called Null Hypothesis Testing (or, NHT):

(1) We develop a research hypothesis or a so-called **alternative hypothesis (e.g., women are more satisfied with the government than men)**.

(2) We use a random sample and estimate the difference in means of government satisfaction between men and women.

(3) We test our result against a (theoretical) distribution under which we assume the **null hypothesis (-> there is NO difference in government satisfaction between men and women)** to be true (i.e., the test distribution).

(4) If the estimate from our sample is located near the center of the distribution that assumes the null hypothesis, this would be not much evidence for the alternative hypothesis; however: If the estimate is at the margins of that distribution, we would conclude that there is evidence that favor the alternative hypothesis, which states that **in the population, women are more satisfied with the government than men**.

- The prerequisite for this testing procedure is that we standardize our sample estimate to make it comparable to the test distribution (For example, this can be achieved by: $z = \frac{estimate}{standarderror}$).

Standard error

- To determine the variation around a population parameter, we could use the standard deviation of the sampling distribution.
- However, since in the applied case we do not have many samples, but only one, we need to estimate this variation from the one sample.
- This estimated variation of an estimate is called the standard error, and the formula for calculating it differs depending on the estimator (usually we normalize a **measure of variance or standard deviation** using a measure that involves the **number of observations**; for example, the standard error of mean the formula is given as: $\bar{x} = \frac{s}{\sqrt{n}}$).

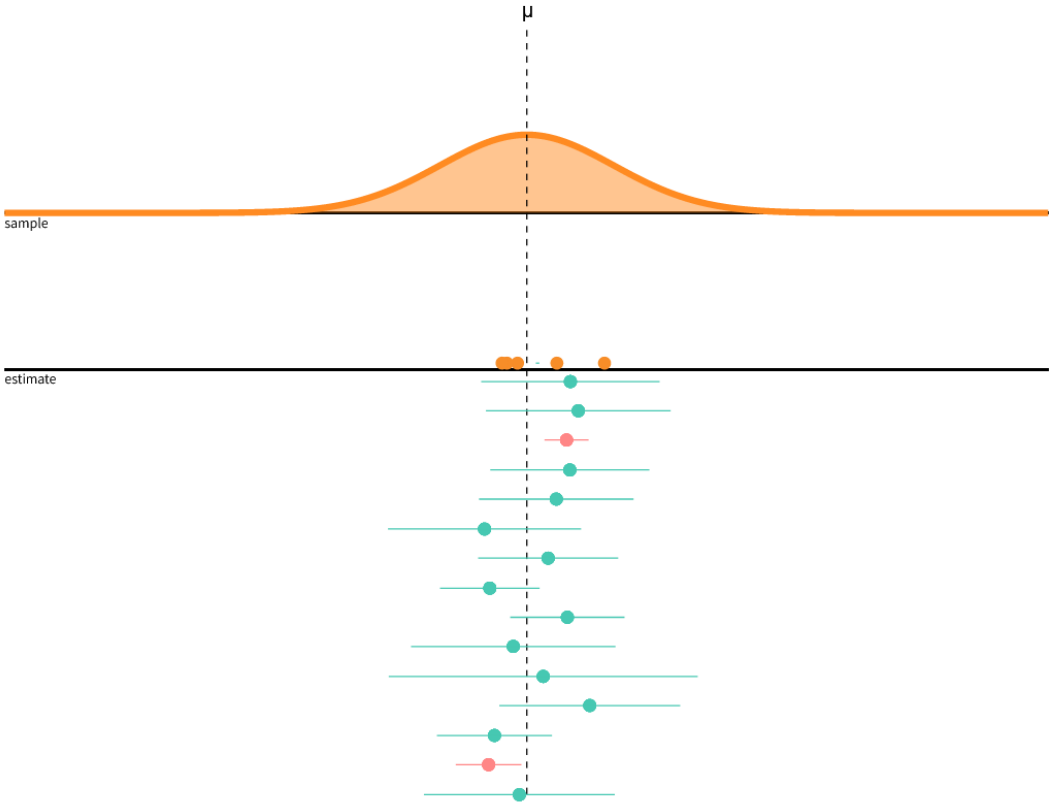
Confidence intervals

Basic idea

In theory, the confidence intervals (CIs) for a point estimate (e.g., a mean or a correlation coefficient) shows a range of values that contain the true population parameter with a certain probability. It is thus a measure of uncertainty (or, conversely, precision) of an estimate. There are conventionally three types of confidence intervals: 90%, 95%, and 99% confidence intervals. These three types reflect the level of probability that the true population value lies within the interval for a very large number of samples. For example, if one

is interested in the population mean and conducts repeated ("infinite") sampling, 95% of the confidence intervals of the respective samples would contain the true population mean, and 5% would not contain it.

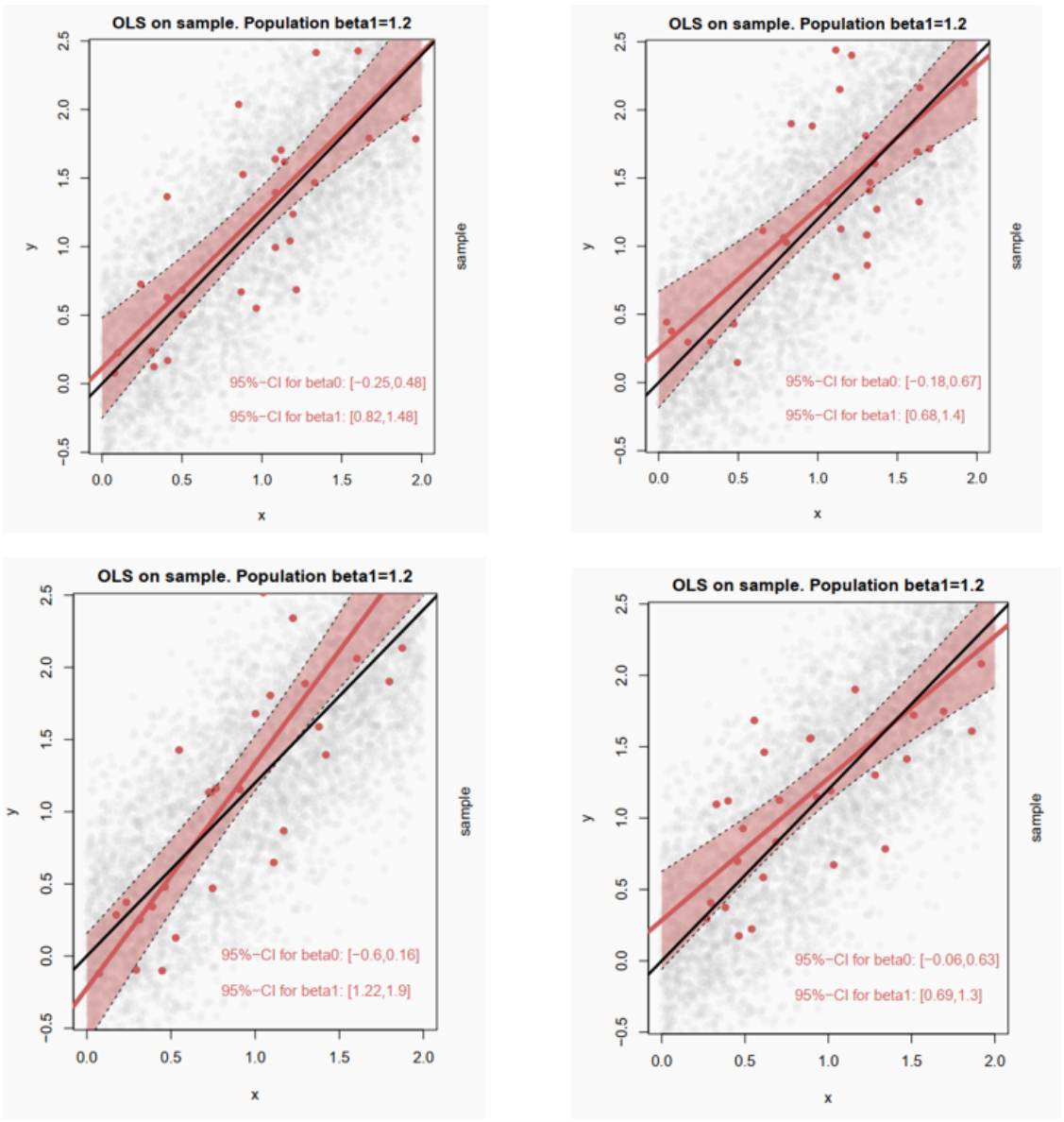
This can be illustrated with the following figures.



Confidence interval of sample means

Source: https://seeing-theory.brown.edu/frequentist-inference/index.html

In the figure, samples were repeatedly drawn from a population with five observations each (orange dots). Each time, the mean is calculated (or estimated because we are using a sample). These point estimates of the mean from each sample are represented by the green or red dots, the confidence interval is the corresponding line to the left and right of it. For the green estimates, the population mean (dashed line) is within the confidence interval, for the red it is not.



Confidence intervals of sample correlations

Source: https://github.com/simonheb/metrics1-public

Here, instead of the mean value, results of a bivariate regression (similar to a correlation) are shown. The bivariate relationship between x and y in the population is represented by the black slope line (regression coefficient $\beta_1 = 1.2$). The grey dots are the units in the population. From this population, four random samples are drawn and the observations in the sample are represented by the red dots. The point estimate of the correlation is represented by the red slope line. The confidence interval is the light red area around the estimate. In samples 1, 2, and 4, the confidence interval includes the population parameter; only in sample 3 (bottom left) is this no longer the case for parts of the confidence interval. Analogously to the thought experiment regarding means above, 95% of estimates from repeated samples would include the population mean, and 5% would not.

How to calculate and interpret the confidence interval

The idea of confidence intervals refers to the sampling distribution, i.e. many (up to infinitely many) repeated random samples from the population. For a 95% confidence interval, the interpretation would be: for repeated samples, 95% of the samples will contain the population parameter, 5% will not.

In practice, we only have one sample and can therefore only calculate confidence intervals once. Whether the population parameter actually falls within the value range of the confidence interval of our one sample or not cannot be answered conclusively. What can be said is that the likelihood of this is high. For example, with the 95% confidence interval, 95 out of 100 samples would have to contain the population parameter and it is likely that our one sample is one of the 95 and not one of the 5. (At the same time, some interpretations made in textbooks such as "with 95% probability the population parameter is in the interval" are strictly speaking not true).

We can therefore interpret the confidence intervals primarily as **the likely or plausible range of values of the unknown population parameter we are interested in (given the characteristics of our sample).** The confidence interval thus gives us, at the same time as the point estimate, a range of variability to be expected if we were to repeat studies similar to ours more often. This is indeed important information, and for many researchers, it is also more intuitive than significance values from null hypothesis tests.

In addition, there is a relationship between confidence intervals and statistical significance, a concept related to null hypothesis tests. We will look at this relationship at the end of this case study. In short, confidence intervals show a possible range of values of the population parameter *and* at the same time provide information about statistical significance.

To compute a confidence interval, we generally use the following formula:

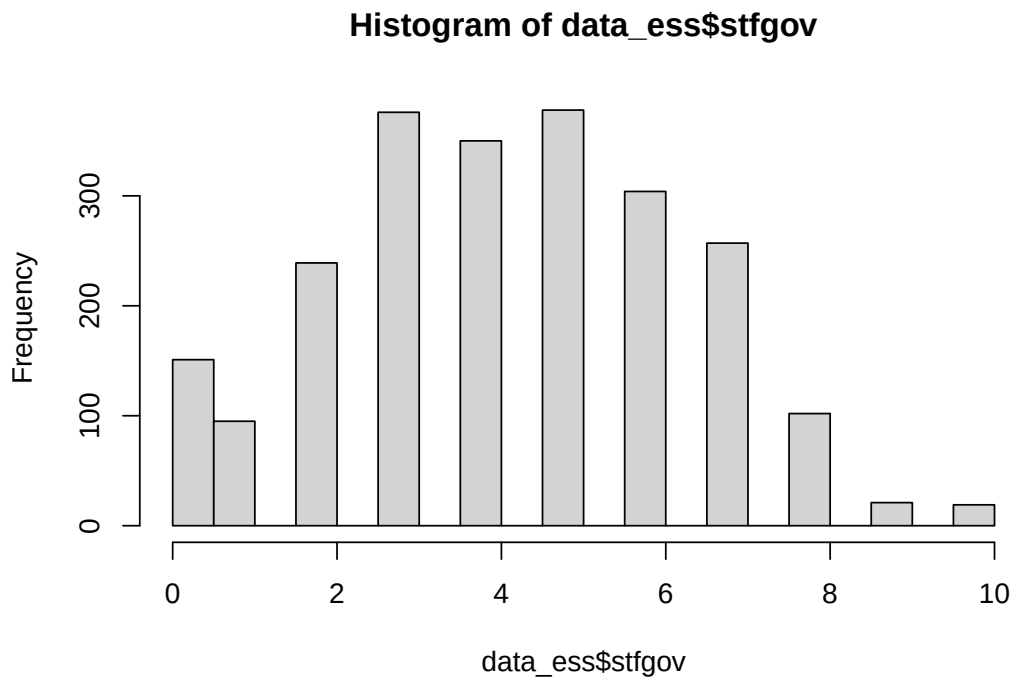
$pointestimate \pm z_\alpha \times standarderror$

z_α refers to the critical Z-value (i.e., a given value of the z-distribution that corresponds a set probability of error, this will be explained in greater detail below).

Mean For the lower and upper 95% confidence intervals of a mean, we use: $95\%CIs = [\bar{x} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}]$

Let us look into this with real data from the European Social Survey (German sample). First, we read the data and get an overview of the distribution of the variable satisfaction with government `stfgov`. `stfgov` was surveyed using an 11-point scale ranging from 0 "extremely dissatisfied" to 10 "extremely satisfied."

```
data_ess <- read_dta("data/ESS9_DE.dta", encoding = "latin1")

hist(data_ess$stfgov, breaks = "FD")
```




```
summary(data_ess$stfgov)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    3.00    4.00    4.28    6.00   10.00      66
```

Next, we calculate the 95% confidence interval in several steps:

```
# We store the mean, the standard deviation and the number of observations as objects,
# because this way we can refer to them later on
n <- 2292
xbar <- mean(data_ess$stfgov, na.rm = TRUE)
s <- sd(data_ess$stfgov, na.rm = TRUE)

# Set confidence level with 1-alpha (alpha = our willingness to be wrong in repeated
# samples)
conf.level <- 0.95

# Calculating the critical z-value for a two-sided test
z <- qnorm((1 + conf.level) / 2)

# Calculate the confidence interval
lower.ci <- xbar - z * s / sqrt(n)
upper.ci <- xbar + z * s / sqrt(n)

# Print confidence intervals
cat("The", conf.level * 100, "% confidence interval for the population mean is (",
    round(lower.ci, 2), ",", round(upper.ci, 2), ").\n")
```

```
## The 95 % confidence interval for the population mean is ( 4.19 , 4.37 ).
```

Computing confidence intervals is even easier with the "t.test" command. While we are interested in the z-distribution, t- and z-distribution correspond to each other when the number of observations is high enough (usually already >30).

Let's try it out:

```
t.test(data_ess$stfgov, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data_ess$stfgov
## t = 92.783, df = 2291, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.189216 4.370120
## sample estimates:
## mean of x
##  4.279668
```

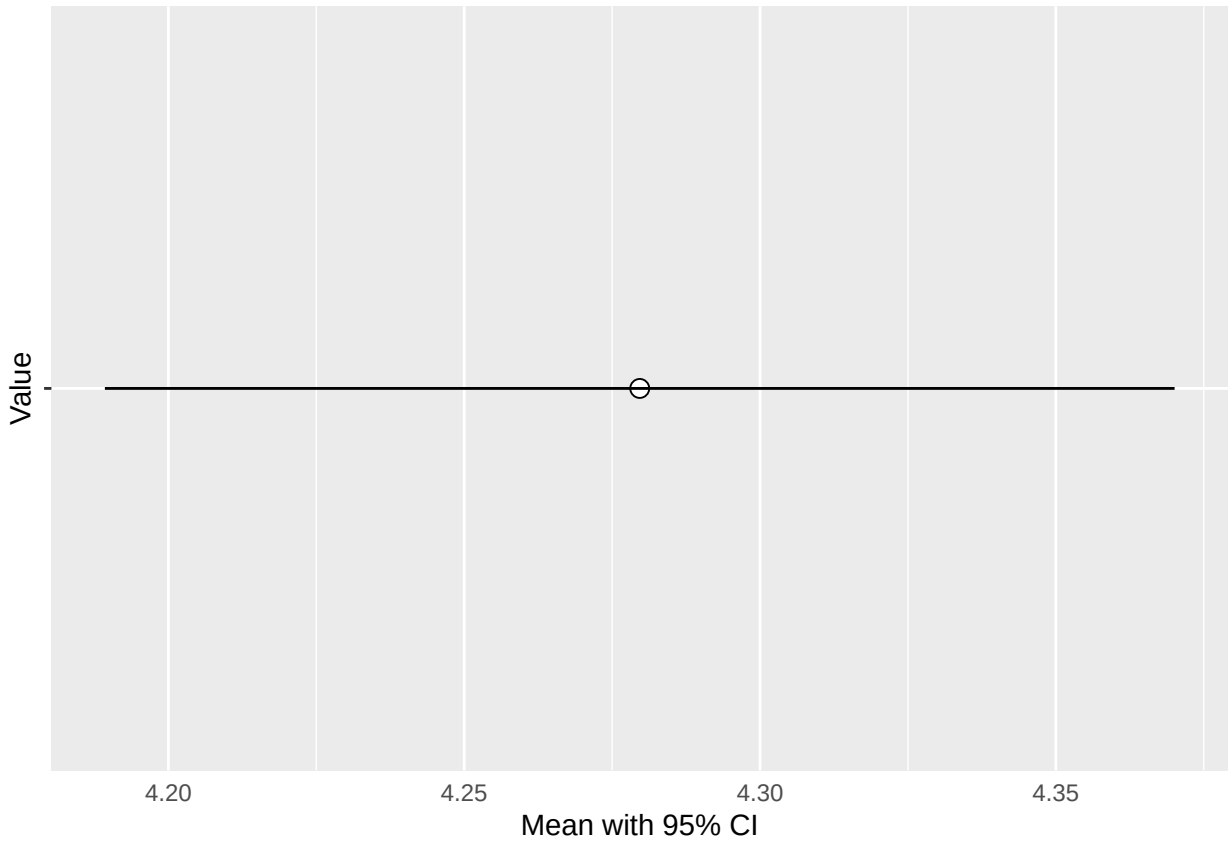
Question: Interpret the confidence interval.

Solution:

> ***Interpretation:*** *Likely values for the population mean are within the limits of 4.2 and 4.4.*

There is also the possibility to display the confidence interval as a plot:

```
df <- data.frame(xbar, lower.ci, upper.ci)

ggplot(df, aes(x = 1, y = xbar)) +
  theme(axis.text.y = element_blank(),
        axis.ticks.x = element_blank()) +
  geom_point(size = 3, shape = 21, fill = "white", colour = "black") +
  geom_errorbar(aes(ymin = lower.ci, ymax = upper.ci), width = 0.2) +
  coord_flip() +
  labs(x = "Value", y = "Mean with 95% CI") +
  scale_x_continuous(breaks = seq(0, 10, by = 1), limits = c(1, 1))
```



Proportion Besides the mean, we can calculate a confidence interval for a proportion value of a sample. The formula is slightly different because the calculation of the standard error is different. It is the square root of the variance divided by the number of observations. The variance is obtained by multiplying the probability of an event p times the counter probability $1 - p$:

$$95\%CIs = [p - 1.96 \times \sqrt{\frac{p \times (1-p)}{n}}, p + 1.96 \times \sqrt{\frac{p \times (1-p)}{n}}]$$

Let's try this by using a dichotomized version of the government satisfaction variable with an initial 11-point rating scale. In this example, values of 5 or less are assigned to 0 ("little or not satisfied") and values above 5 are assigned to 1 ("satisfied"). We can now focus on the proportion of those respondents who are satisfied or rather satisfied with the government (variable value = 1). Note that with a 0/1 coding, the mean of a binary variable corresponds to the proportion of observations with a variable value of 1.

```
# Recode of variable
data_ess <- data_ess %>%
  mutate(stfgov_di =
           case_when(stfgov <= 5 ~ 0,
                     stfgov > 5 ~ 1))

# Proportion via table command
prop.table(table(data_ess$stfgov_di))
```

```
##
##        0        1
## 0.693281 0.306719
```

```
# Correspondence with the mean of the summary command
summary(data_ess$stfgov_di)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.3067  1.0000  1.0000      66
```

```r
# Set confidence level with 1-alpha
conf.level <- 0.95

# Calculating the critical z-value for a two-sided test
z <- qnorm((1 + conf.level) / 2)

# Calculation of the confidence interval
lower.ci <- 0.3067 - z * sqrt((0.3067 * (1 - 0.3067)) / n)
upper.ci <- 0.3067 + z * sqrt((0.3067 * (1 - 0.3067)) / n)

# Print the confidence intervals
cat("The", conf.level * 100, "% confidence interval for the proportion value is (",
    round(lower.ci, 2), ",", round(upper.ci, 2), ").\n")
```

```
## The 95 % confidence interval for the proportion value is ( 0.29 , 0.33 ).
```

```r
# Test using t.test
t.test(data_ess$stfgov_di, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data_ess$stfgov_di
## t = 31.837, df = 2291, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2878265 0.3256115
## sample estimates:
## mean of x
##  0.306719
```

Question: Interpret the confidence interval.

Solution:

> ***Interpretation:*** *Likely values for the population proportion are within the limits of 0.29 and 0.33.*

Correlation coefficient Next, we look at how a confidence interval for a correlation is computed. For this, we use the following formula: $95\%CIs = r \pm 1.96 \times SE$, where the standard error SE is calculated as follows: $SE_r = \sqrt{\frac{(1-r^2)}{n-2}}$. To illustrate that, we let the software compute the confidence interval of the correlation between household income and satisfaction with the government. `hinctnta` measures respondents' net household income (after deductions) across 10 quantiles ("deciles" 1-10) - the reason for this measurement is that this way income becomes adjusted to a country's income distribution and is thus comparable across countries. Regarding the correlation, we assume that - in line with economic voting theory - people with a higher income are more satisfied with the government than people with a lower income.

```r
cor <- cor.test(data_ess$hinctnta, data_ess$stfgov)
cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_ess$hinctnta and data_ess$stfgov
## t = 1.805, df = 2047, p-value = 0.07122
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.003445968  0.083023781
## sample estimates:
##       cor
## 0.03986354
```

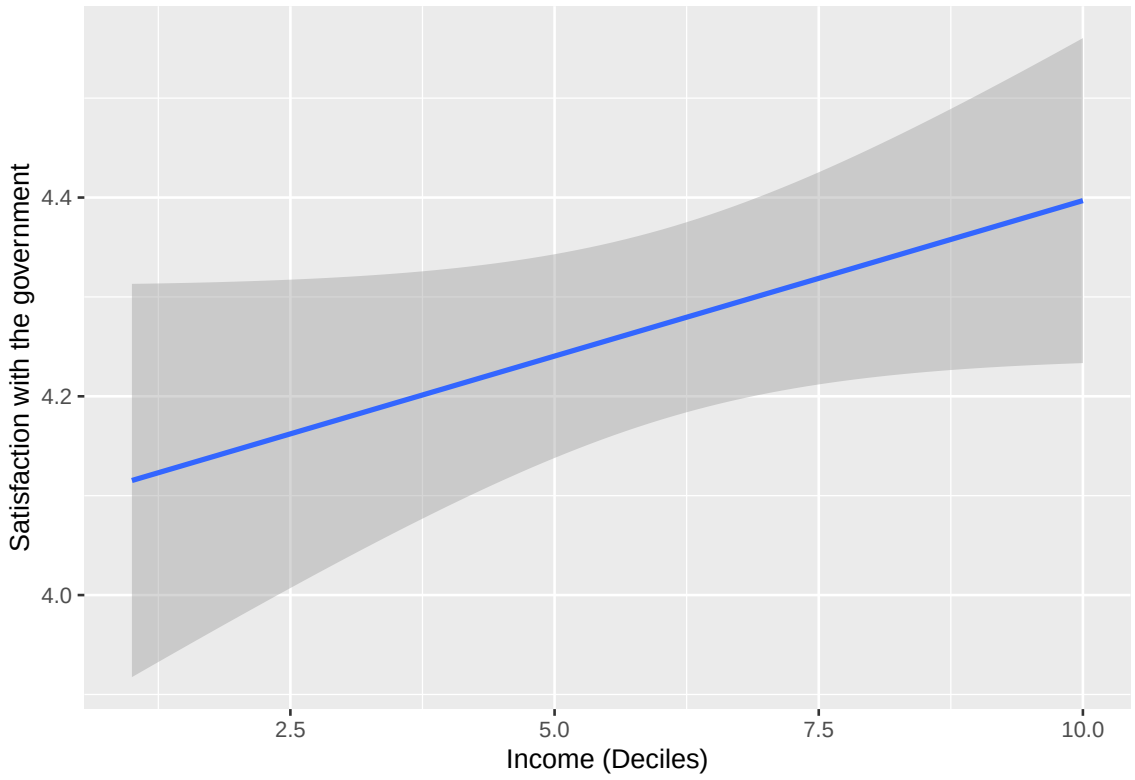
Question: Interpret the confidence interval.

Solution:

> ***Interpretation:*** *Likely values for correlation in the population are within the limits of -0.003 and 0.08.*

The graphical representation of the confidence interval of a correlation is two-dimensional. The width of the confidence interval can vary along the values of the x-axis. The value from the previous test is, so to say, the average confidence interval.

```r
ciplot <- ggplot(data=data_ess, aes(x = hinctnta , y = stfgov)) +
  geom_smooth(method = lm, se = TRUE, level = 0.95) +
  xlab("Income (Deciles)") +
  ylab("Satisfaction with the government")

ciplot
```



Question: The shown figure displays 95% confidence intervals. Would 90% confidence intervals be narrower or wider?

Solution:

The 90% confidence intervals would be narrower. Going back to the probability-world interpretation: If we only want to capture the true population parameter 90 percent of the time (instead of 95), this means that we could narrow down the range of values in which the parameter will fall. In contrast, if we decrease our readiness to err to 1% (i.e., a 99% confidence interval), the range of values to fulfill this assumption would increase - the confidence interval would get wider.

Null Hypothesis Testing

Basic idea and procedure

Hypothesis testing allows us to test how likely it is that a sample estimate corresponds to another value, more specifically the sampling distribution of another value. We typically use for this purpose the distribution that is obtained under the null hypothesis- that is, for example, the assumption that the difference in means between two groups or the correlation between two variables is zero in the population. The distribution of a test statistic under the assumption that the null hypothesis is true is sometimes labeled "null distribution." We know from the central limit theorem that such a sampling distribution would be normal given repeated sampling and a high enough sample size (>30), which implies that we can use a normal distribution for our statistical tests.

The goal usually is to test the sample estimate we obtained against the null distribution. If the estimate is so "extreme" that it becomes very unlikely to obtain such a result although the null hypothesis is true, then this would provide evidence against the null hypothesis and in favor of our alternative hypothesis (e.g., the

correlation between two variables is positive in the population). Hence, the procedure of Null Hypothesis Testing is a form of *proof by contradiction*. We assume the opposite of what we want to show and then collect as much evidence as possible against that assumption.

To begin with, we first formulate the so-called **null hypothesis** H_0 (e.g., there is *no* or zero correlation in the population). Next, we formulate an **alternative hypothesis** H_A (e.g., the correlation in the population is positive or negative) and then want to show that the correlation we found is strong enough so that it becomes very unlikely that H_0 is true. This means that we are trying to disprove H_0, which in turn is evidence for H_A. Note that we do not assign probabilities for H_A to be true but probabilities to observe a given estimate given that H_0 is true (i.e., $P(\text{correlation} \mid H_0)$).

The following steps illustrate how to do Null Hypothesis Testing:

- **Step 1:** Formulate the null hypothesis and alternative hypothesis.
- **Step 2:** Determine the probability of error (α, i.e., our willingness to err in repeated sampling, the convention is typically 5% or less).
- **Step 3:** Compute the value of the test statistic for your estimate (=estimator/standard error).
- **Step 4a:** Determine a critical value c (e.g., from an online table)
- **Step 5a:** If the value of the test statistic falls within the rejection range (i.e., exceeds the critical value), reject H_0.

Alternatively:

- **Step 4b:** Calculate p-value (use statistical software).
- **Step 5b:** If p-value $<$ probability of error, then reject H_0.
- **Step 6:** Interpret the result of the hypothesis test in substantive terms.

Hypothesis test of a difference in means

Differences in means usually refer to a comparison of the mean of a variable between two groups. To illustrate that, we use the example of income and government satisfaction again. We leave the variable government satisfaction `stfgov` in its original metric form so that we can calculate the mean. However, we dichotomize the variable income `hinctnta` into two groups: low earners and high earners.

```
data_ess <- data_ess %>%
  mutate(hinctnta_di =
           case_when(hinctnta <= 6 ~ 0,
                     hinctnta > 6 ~ 1))
summary(data_ess$hinctnta_di)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.4794  1.0000  1.0000     270
```

Question: Formulate the null and alternative hypotheses if we assume that persons with a high-income are more satisfied with the government.

Solution:

Typically, it is easier to start with formulating the alternative hypothesis.

H_A: High-income individuals are more satisfied with government than low-income individuals (i.e., positive relationship between income and satisfaction with government).

Formal: $\mu(satisfaction|highincome) > \mu(satisfaction|lowincome)$ or (equivalent) $\mu(satisfaction|highincome) - \mu(satisfaction|lowincome) > 0$

H_0: High-income individuals are equally or less satisfied with government than low-income individuals (i.e., null or negative relationship between income and satisfaction with government).

Formal: $\mu(satisfaction|highincome) \leq \mu(satisfaction|lowincome)$ or (equivalent) $\mu(satisfaction|highincome) - \mu(satisfaction|lowincome) \leq 0$

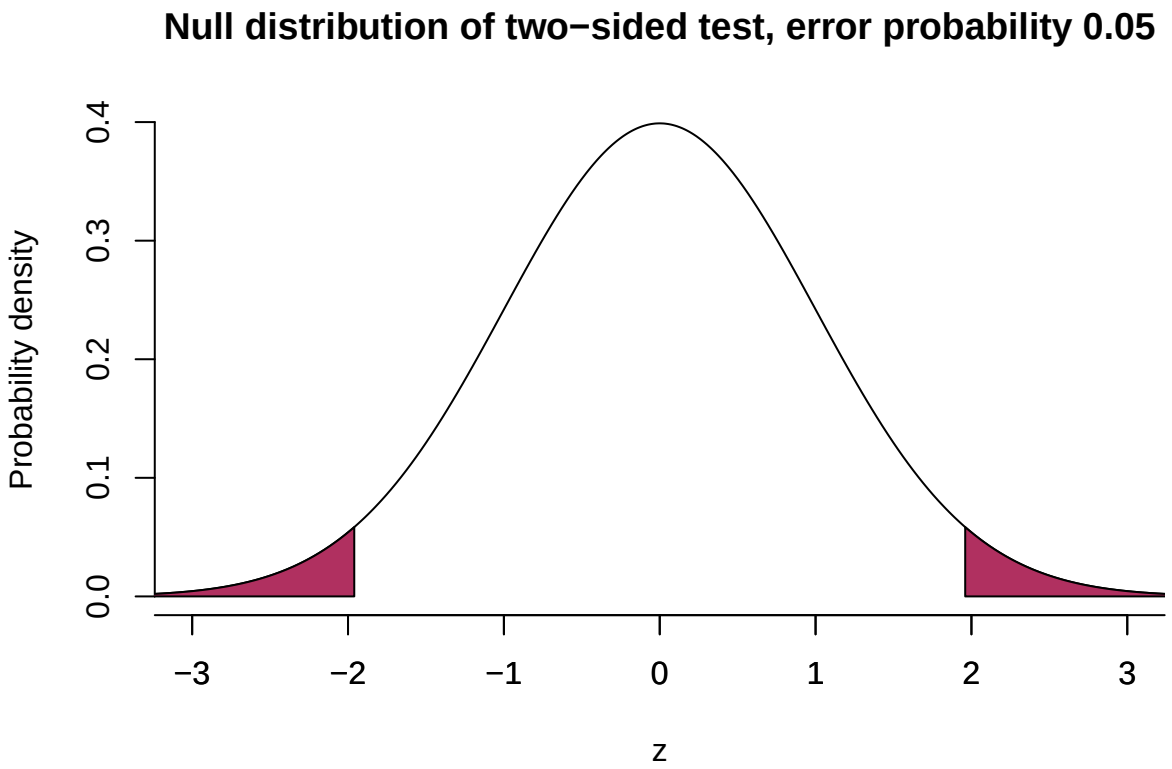
Note that it is important that all possible outcomes are covered by both hypotheses. This means that when formulating a directed hypothesis, the null hypothesis must include the case of no relationship (i.e., "$\leq$" instead of just "$<$").

Let us then assume an error probability of 5% or 0.05 (Step 2). Next, we calculate the test statistic (Step 3). Generally, this is given as $\frac{estimate-truevalue}{standarderror}$. Since we test against the null hypothesis, the true value is 0. The formula for the so-called z-test statistic is given as:

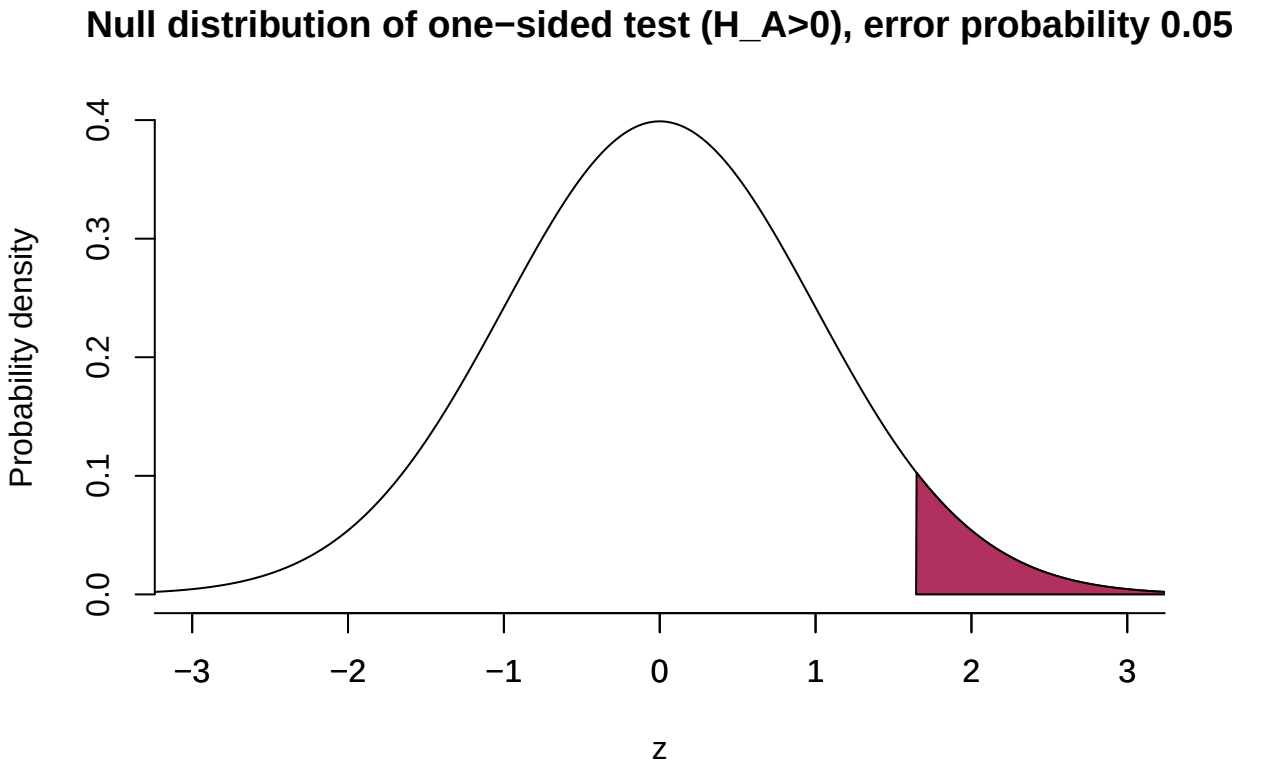
$z = \frac{Estimation}{StandardError}$

Steps 4 and 5 are about the actual test. Remember that per the alternative hypothesis, we assumed a positive mean difference (i.e., high-income earners are more satisfied with the government than low-income earners). Next, we want to know where the mean difference is located on the null distribution. If it is close to the center of the distribution, we have no reason to reject the null hypothesis. However, if it is at the margins of the distribution, this would provide a reason for rejecting the null hypothesis (and favoring the alternative hypothesis). We use the z-distribution with a mean of 0 and a standard deviation of 1. Remember the properties of this distribution from above. If we set the critical value to be lower than -1.96 and higher than +1.96, we would be in the 5% range of the distribution if our test statistic exceeds those critical values. If this was the case, we can conclude that it would be very unlikely to sample such a value even though H_0 holds. Note that we set up a positive and negative area of rejection if we formulate undirected hypotheses (i.e., H_A: the mean difference can be higher or lower than zero, H_0: the mean difference equals zero) and conduct two-sided tests. For a directed hypothesis (and one-sided tests), like in our example, the rejection area is only in the negative or positive range of the distribution (depending on whether H_A expects a pos. or neg. statistic).

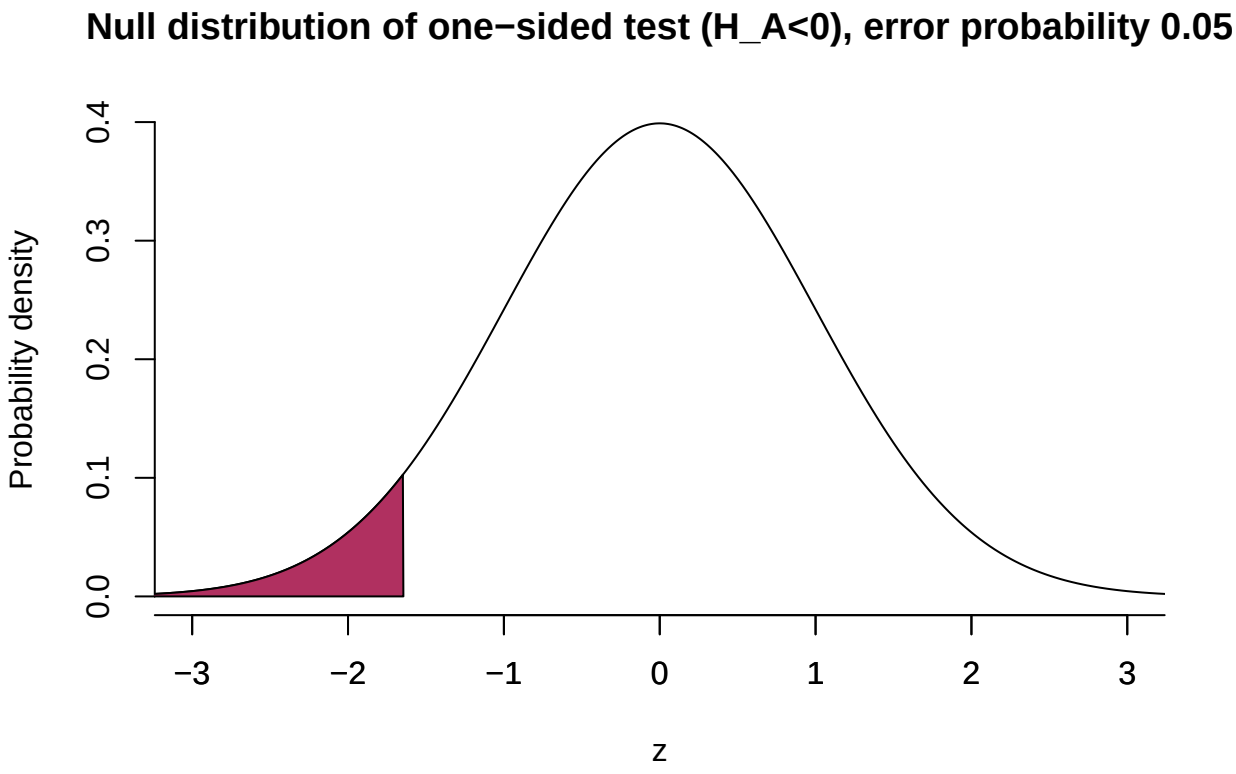
```
# Draw a standard normal distribution:
z = seq(-4, 4, length.out = 1001)
x = rnorm(z)

# Null distribution of two-sided test
plot(x = z, y = dnorm(z), bty = 'n', type = 'l',
     main = "Null distribution of two-sided test, error probability 0.05",
     ylab = "Probability density", xlab = "z", xlim = c(-3, 3))
axis(1, at = seq(-4, 4, by = 1))
polygon(x = c(z[z <= qnorm(0.025)], qnorm(0.025), min(z)),
        y = c(dnorm(z[z <= qnorm(0.025)]), 0, 0),
        col="maroon")
polygon(x = c(z[z >= qnorm(0.975)], max(z), qnorm(0.975)),
        y = c(dnorm(z[z >= qnorm(0.975)]), 0, 0),
        col = "maroon")
```




```r
# Null distribution of one-sided test (H_A>0)
plot(x = z, y = dnorm(z), bty = 'n', type = 'l',
     main = "Null distribution of one-sided test (H_A>0), error probability 0.05",
     ylab = "Probability density", xlab = "z", xlim = c(-3, 3))
axis(1, at = seq(-4, 4, by = 1))
polygon(x = c(z[z >= qnorm(0.95)], max(z), qnorm(0.95)),
        y = c(dnorm(z[z >= qnorm(0.95)]), 0, 0),
        col = "maroon")
```




```r
# Null distribution of one-sided test (H_A<0)
plot(x = z, y = dnorm(z), bty = 'n', type = 'l',
     main = "Null distribution of one-sided test (H_A<0), error probability 0.05",
     ylab = "Probability density", xlab = "z", xlim = c(-3, 3))
axis(1, at = seq(-4, 4, by = 1))
polygon(x = c(z[z <= qnorm(0.05)], qnorm(0.05), min(z)),
        y = c(dnorm(z[z <= qnorm(0.05)]), 0, 0),
        col = "maroon")
```



Question: How would the rejection area change if we test one-sided? What would change if we set a 1% probability of error (two-sided test)?

Solution:

Assuming a positive mean difference or correlation, the rejection area would be located only on the right side of the distribution. Since we still use an error probability of 5%, this area would be larger than in the two-sided case. Consequently, the critical value c thus moves a bit in the direction of the center of the distribution. For the z-distribution, this value is +1.645 for a positive alternative hypothesis (and -1.645 for a negative hypothesis).

If we change the probability of error to 1% and test two-sided, we have a rejection range that is smaller compared to the 5% case. Our estimate and the corresponding z-statistic need to be larger to reject the null hypothesis compared to the 5%-case. In the 1% case, the critical values are -2.326 and +2.326.

Critical values at which one can reject the null hypothesis can be obtained from here: https://www.criticalvaluecalculator.com/

General remarks on P-values P-values **indicate the probability of finding an estimate as extreme (or more extreme) as we got from our sample (e.g., a difference in means), even though the null hypothesis is true in the population.** In formal terms, that is the probability of the estimate given the null hypothesis $P(\text{estimate} \mid H_0)$.

Some features of p-values:

- The smaller the p-value, the more "statistically significant" the result (or the more reason we have to reject the null hypothesis, which, in turn, provides evidence in favor of the alternative hypothesis).

- Contingent upon the probability of error, if p is smaller than α, then the result is statistically significant (usually < 0.05 or 5%).
- If the test statistic associated with the estimate is less than or greater than the critical z- or t-values, respectively, then $p < 0.05$.
- Correspondence with confidence intervals:
  - If the confidence interval includes the null hypothesis value (usually 0), this is consistent with a high p-value, indicating weak evidence against the null hypothesis.
  - If the confidence interval does not include 0, it is consistent with a low p-value, suggesting strong evidence against the null hypothesis.
  - For example: A mean difference of 0.5 with a 95% confidence interval of [0.35; 0.65] is statistically significant at the α=0.05 level (reject H_0). However, a mean difference of -0.1 with a 95% confidence interval of [-0.25; 0.05] is statistically **not** significant at the α=0.05 level (confidence intervals include 0, fail to reject H_0).
- To calculate the exact p-value, we rely on tables or statistical software.
- A p-value says **nothing** about the probability that H_0 or H_A is true; remember $P(\text{estimate} \mid H_0)$!
- When we formulate one-sided hypotheses but use a two-sided test (which is the norm in social science research), we are especially careful and do not want to make a 1st kind error (H_0 is true, but we reject it); we can divide the p-value by 2 and end up with the p-value we would have gotten if we tested one-sided.

Test implementation To actually test the difference in the means of satisfaction with the government between high and low earners, the formula for this is given as $z = \frac{(\bar{x}_1 - \bar{x}_2)}{standarderror}$ (note that the standard error of the mean difference is not simply the difference of two standard errors).

We now calculate the mean difference and conduct a corresponding hypothesis test:

```
mean(data_ess$stfgov[data_ess$hinctnta_di==1], na.rm = TRUE) -
  mean(data_ess$stfgov[data_ess$hinctnta_di==0], na.rm = TRUE)
```

```
## [1] 0.1543566
```

```
t.test(data_ess$stfgov[data_ess$hinctnta_di==1], data_ess$stfgov[data_ess$hinctnta_di==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data_ess$stfgov[data_ess$hinctnta_di == 1] and data_ess$stfgov[data_ess$hinctnta_di == 0]
## t = 1.5882, df = 2046.8, p-value = 0.1124
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03624397  0.34495717
## sample estimates:
## mean of x mean of y
##  4.354545  4.200189
```

Question: Interpret the result of the statistical test in detail with reference to the specific p-value.

Solution:

The probability of obtaining such a result from the sample, although H_0 is true in the population, is 0.11 or 11%. Since this exceeds the threshold of our willingness to err of 5%, we cannot reject the null hypothesis (and thus find no evidence in favor of H_A).

Even if we take into account that we formulated a directed hypothesis and rely on a one-sided test (we can thus divide the p-value by 2), the p-value still exceeds our assumed probability of error (p-value of $0.055 > \alpha$ of 0.05). Hence, even with a one-sided test, we are not able to reject the null hypothesis.

Hypothesis test of a correlation

Instead of a difference in means, we now focus on hypothesis tests of correlations. The procedure is analogous to the one before. We first formulate the null and alternative hypotheses, specify the probability of error (5% in this case), calculate the test statistic ($z = \frac{r}{SE}$), calculate the p-value, and interpret the result.

Question: Assume that higher income is either positively or negatively related to satisfaction with government. Formulate the null and alternative hypotheses (undirected).

Solution:

H_A: Verbal: Income is either positively or negatively related to satisfaction with government.

Formal: $r \neq 0$

H_0: Verbal: Income is unrelated to satisfaction with government.

Formal: $r = 0$

We now calculate the test statistic and the p-value.

```
cor <- cor.test(data_ess$hinctnta, data_ess$stfgov)
cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_ess$hinctnta and data_ess$stfgov
## t = 1.805, df = 2047, p-value = 0.07122
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.003445968  0.083023781
## sample estimates:
##        cor
## 0.03986354
```

Question: Interpret the result of the statistical test (two-sided test) with reference to the t-value and p-value.

Solution:

The probability of obtaining such a result from the sample, although H_0 is true in the population, is 0.07 or 7%. Since this exceeds the threshold of our willingness to err of 5%, the result is not statistically significant and we cannot reject the null hypothesis (and thus find no evidence in favor of H_A).

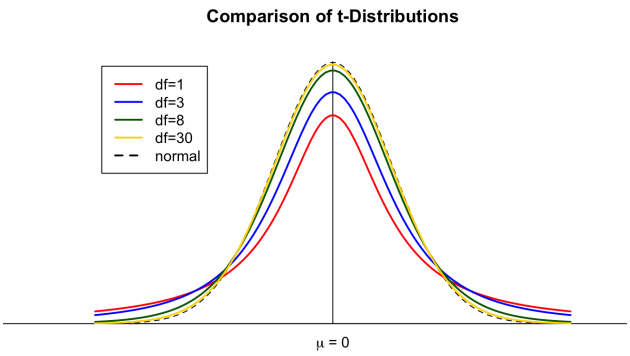
Outlook

This case study reviewed the basics of probability theory, some relevant statistical laws, confidence intervals, and hypothesis testing. In the examples, we have mainly used the z-distribution, which is the way to go if we deal with large or reasonably large samples (>30 is already sufficient). For small samples and other purposes, we use other distributions for statistical testing that are briefly introduced here.

t-distribution

The t-distribution corresponds to the z-distribution for larger samples. For small samples, the distribution is flatter and therefore has wider margins. The critical t-values (above or below we can reject the null hypothesis) are therefore larger on the right and lower on the left side. This makes it more difficult to reject the null hypothesis. In other words: We need more extreme results in small samples to reject the null hypothesis. This accounts for the fact that some outliers have greater weight in small samples. This helps to avoid the so-called alpha error (or type I error), which refers to erroneously inferring statistically significant results although the null hypothesis is true in the population. The alpha error is usually seen as a greater threat to scientific progress than failing to reject the null hypothesis although we should have done so (beta error or type II error).

Df refers to degrees of freedom, which is usually the number of cases - 1 (for estimation).



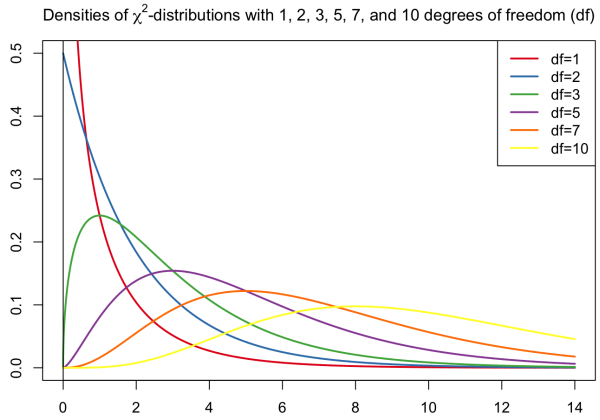
t-distribution

Source: https://www.geo.fu-berlin.de/en/v/soga-py/Basics-of-statistics/Continous-Random-Variables/Students-t-Distribution/index.html

Chi-square distribution

The chi-square distribution is used, for example, for testing variances. One field of application is in cross tabulations. Chi-square values are never negative, so only one-sided tests are suitable.

Df refers to degrees of freedom, which in the case of crosstabs would be the table size, for example.

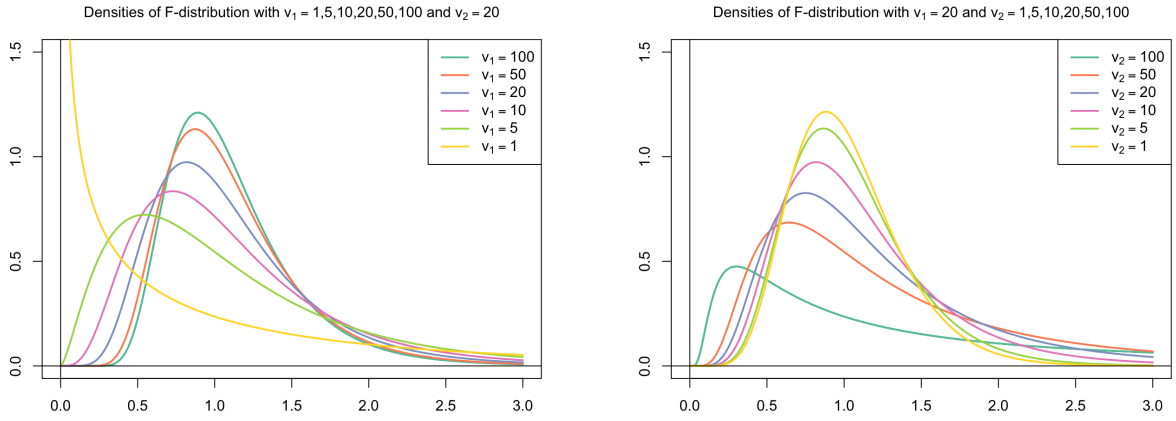


Chi-Square-Distribution

Source: https://www.geo.fu-berlin.de/en/v/soga-py/Basics-of-statistics/Continous-Random-Variables/Chi-Square-Distribution/index.html

F-Distribution

The F-distribution is basically the ratio of two chi-squared distributed random variables (e.g., two variances). The F-distribution is also always positive and the functional form depends on the degrees of freedom of the two variables under consideration. F-tests are used, for example, in comparing the model fit of regression models.



Source: https://www.geo.fu-berlin.de/en/v/soga-py/Basics-of-statistics/Continous-Random-Variables/F-Distribution/index.html

Statistical versus substantive significance

The procedures discussed here relate to the concept of hypothesis testing and statistical significance. They are essentially concerned with quantifying uncertainty and inferring from that whether or not one can assume

that a result from a random sample can be generalized to the population the sample has been drawn from. Although these are important concepts, statistical significance is different from scientific or substantive significance, which refers to the degree to which a result is relevant against the background of other research findings or alternative explanatory factors. Thus, a statistically significant result may not be substantial or, conversely, a substantial result may not be statistically significant. Statistical significance, for example, is highly dependent on the number of observations that are incorporated in calculating the standard error. A measure of the substance of, for example, a regression coefficient can be obtained by using standardized effect sizes such as beta or Cohen's D.

In general, we are looking for empirical results that are both statistically and substantively significant.

Regression Analysis - Case Study Attitudes Toward Justice

In this case study, we use survey data from the 2021 German General Population Survey of the Social Sciences (German: ALLBUS). Specifically, we examine the relationship between **income** and **attitudes toward social justice**.

(1) First, we prepare the data and analyze the bivariate relationship between income and attitudes toward social justice.

(2) Second, we control for possible alternative explanations in a multiple regression model. Here we focus on the "confounding variables" gender of respondents and whether or not respondents have been unemployed in the past 10 years.

(3) Third, we take a look at regression diagnostics, how to model nonlinear relationships between variables, and interactions between variables.

Data preparation and bivariate regression

First, we read the data set and get an overview of the variables in the data set. For reasons of space, we only show a selection of the variables in the table.

```
data_allbus <- read_dta("data/allbus2021_reduziert.dta")

# Here all cases are deleted, which have missing values on any of the variables.
# This is only recommended for reduced data sets with few variables, otherwise you would
# "throw out" observations based on non-relevant variables
data_allbus <- na.omit(data_allbus)

# We also delete three observations that indicated "diverse" for the gender variable, as
# this simplifies hypothesis generation
data_allbus <- subset(data_allbus, sex != 3)

# We now recode variables
data_allbus$income <- data_allbus$incc

data_allbus <- data_allbus %>%
  mutate(female =
           case_when(sex == 1 ~ 0,
                     sex == 2 ~ 1))

data_allbus <- data_allbus %>%
  mutate(east =
           case_when(eastwest == 1 ~ 0,
                     eastwest == 2 ~ 1))

data_allbus <- data_allbus %>%
  mutate(unemployed =
           case_when(dw18 == 1 ~ 1,
                     dw18 == 2 ~ 0))

# Summarize and overview data
summary(data_allbus)
ft4 <- flextable(head(select(data_allbus, !c(im19:im21, incc))))
autofit(ft4)
```

```
##     eastwest          sex             im19            im20          im21
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000
##  Median :1.000   Median :1.000   Median :3.000   Median :3.00   Median :3.000
##  Mean   :1.308   Mean   :1.486   Mean   :2.771   Mean   :2.82   Mean   :2.984
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.:4.000
##  Max.   :2.000   Max.   :2.000   Max.   :4.000   Max.   :4.00   Max.   :4.000
##       dw18            incc          income         female
##  Min.   :1.000   Min.   : 1.0   Min.   : 1.0   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:13.0   1st Qu.:13.0   1st Qu.:0.0000
##  Median :2.000   Median :15.0   Median :15.0   Median :0.0000
##  Mean   :1.812   Mean   :15.3   Mean   :15.3   Mean   :0.4859
##  3rd Qu.:2.000   3rd Qu.:19.0   3rd Qu.:19.0   3rd Qu.:1.0000
##  Max.   :2.000   Max.   :26.0   Max.   :26.0   Max.   :1.0000
##       east          unemployed
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.3079   Mean   :0.1884
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

eastwest	sex	dw18	income	female	east	unemployed
1	2	2	9	1	0	0
1	1	2	17	0	0	0
1	2	2	7	1	0	0
1	2	2	21	1	0	0
1	1	2	21	0	0	0
1	2	2	4	1	0	0

Next, we build an index from the variables `im19` (agreement: income differences increase motivation), `im20` (agreement: rank differences between people are acceptable) and `im21` (agreement: social differences are fair). Since high values on those variables indicate disagreement, we can label the index **`morejustice`** for which high values represent disagreement with social inequality (i.e., advocacy of more justice).

Question: Could we have recoded the variables differently? What would change? And what would happen if, for example, we recoded only two of the three variables and then built an index with all three variables?

Solution:

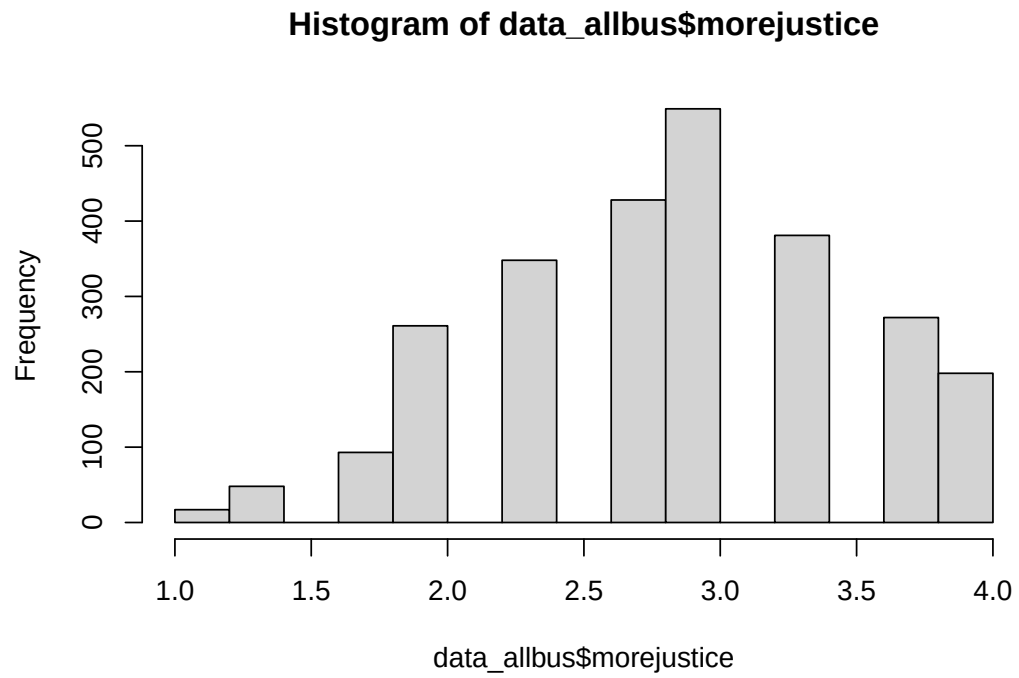
Yes, if we were to recode all items, that would be just as possible. We would then have to label an index as "less justice" or "inequality," since high values represent acceptance of social inequality. Recoding only a few variables is inadmissible, the index would no longer be valid.

For building scales or indices of items, it is always a good idea to check whether there are substantial correlations between the items. There are also other ways to look at the reliability of an index, for example, by calculating Cronbach's alpha. We go with correlations and also look at the distribution of the index variable.

```r
# Do the items correlate sufficiently high for the index? (>.4 would be desirable)
# Calculate correlation and output
vars <- c("im19", "im20", "im21")
cor.vars <- data_allbus[vars]
rcorr(as.matrix(cor.vars))
```

```
##      im19 im20 im21
## im19 1.00 0.46 0.33
## im20 0.46 1.00 0.51
## im21 0.33 0.51 1.00
##
## n= 2595
##
##
## P
##      im19 im20 im21
## im19       0    0
## im20  0         0
## im21  0    0
```

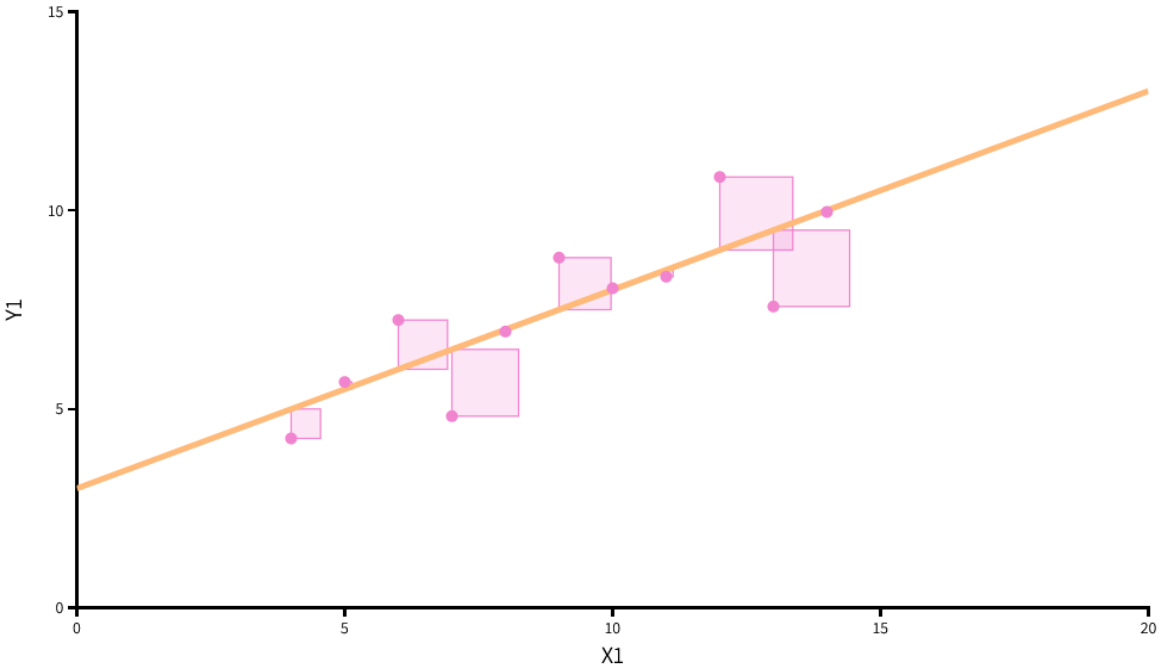
```r
# Index creation
data_allbus$morejustice <- (data_allbus$im19 + data_allbus$im20 + data_allbus$im21) / 3
hist(data_allbus$morejustice)
```



The correlations show that at least most of the indicators are correlated with each other by r > .4. The histogram suggests that the variable values are approximately normally distributed.

The foundations of (bivariate) regression

In the next step, we will work our way into bivariate regression analysis. In order to calculate the relationship between two variables, a so-called slope line is calculated using the least squares method. The following figure shows the basic idea. These are fictitious data points.



Source: https://seeing-theory.brown.edu/regression-analysis/index.html

Basically, this method allows us to find a straight line that optimally travels through the dots of observations. The smaller the squares, the better the line fits the data. We have found the optimal solution if any different slope of the line would lead to an increase in the sum of squared deviations. Hence, this estimator is called **ordinary least squares - the OLS estimator**. If we had more than one variable in the regression model, another spatial dimension would be added (a z-axis) and the straight line would become a surface. With more than two independent variables, we are dealing with a multidimensional surface that is no longer depictable. The least squares method remains the same in such a case, except that the computation becomes more complex (done with a software algorithm).

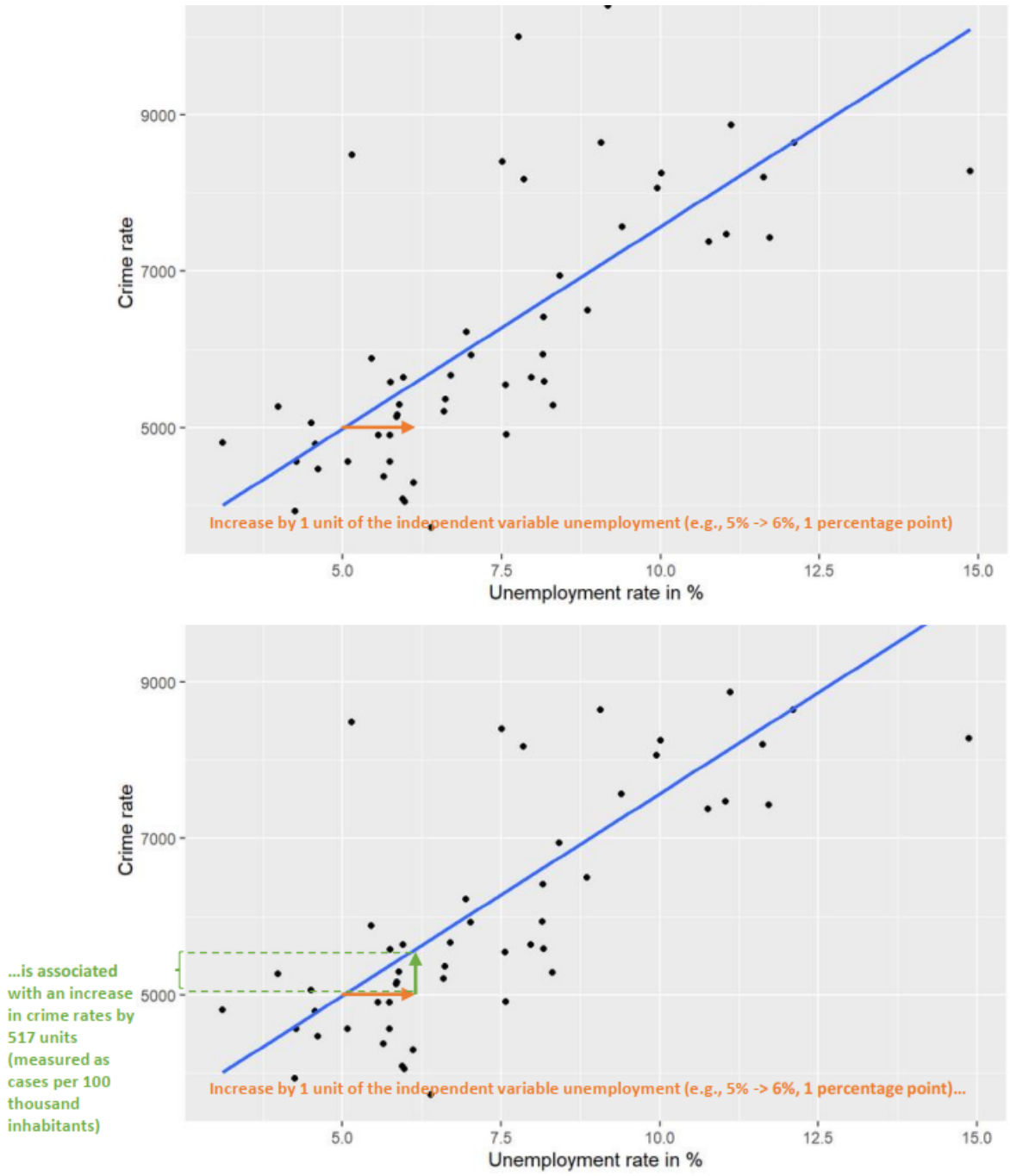
The software program computes a so-called regression coefficient for each independent variable (e.g., β_1 for the corresponding variable X_1). In the bivariate case (one independent variable), the regression coefficient represents the "steepness" of the slope in the following way:

If X_1 increases by one unit, then Y increases or decreases by β_1 units (depending on the sign).

Let us now briefly remember the example from case study on univariate statistics, where we focused in the outlook on the empirical relationship between unemployment and crime rates in German counties. There we found a positive bivariate relationship between the two variables. The table below shows the coefficient estimate and the figure that depicts the positive association. If unemployment increases by one unit (here: 1 percentage point), then crime rates increase by 517.46 units (here: cases per 100T inhabitants), on average.

```r
data_nrw <- read_excel("data/inkar_nrw.xls")
model1 <- lm(crimerate ~ 1 + unemp, data = data_nrw)
stargazer(model1, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                crimerate
## ------------------------------------------------
## unemp                         517.460***
##                                (69.412)
##
## Constant                     2,398.594***
##                                (543.248)
##
## ------------------------------------------------
## Observations                      53
## R2                               0.521
## Adjusted R2                      0.512
## Residual Std. Error      1,240.129 (df = 51)
## F Statistic            55.575*** (df = 1; 51)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```



We can also think about what would happen if the dots (i.e., the observed values on the two variables) were different. If the dots and the corresponding slope line were steeper (and thus the coefficient larger), then an increase in unemployment by one unit would lead to an even higher increase in crime rates. If the dots and the line indicated a flat pattern, then this would mean that there is no change in crime rates when unemployment increases by one unit. If the line (and the coefficient) was negative, then a one-unit increase in unemployment would be associated with a *decrease* (by β_1 units) in crime.

In addition to the coefficients for the independent variables, the regression output displays the so-called constant (or intercept), which is abbreviated as a, α, or β_0. It provides the predicted value for the outcome variable Y if all independent variables in the model have the value of 0. Usually, this information is not of much interest and thus ignored.

Question: What is the value of the constant in the unemployment-crime example (Model 1)? What information content does it carry?

Solution:

The value of the constant in Model 1 is 2399. This indicates the average crime rate (Y) in a municipality with 0% unemployment. In our case, the constant is uninformative because we don't have observations with 0% unemployment in our sample.

Bivariate regression of income and attitudes toward justice

We now estimate a bivariate regression for the relationship between income and attitudes toward justice. Our working hypothesis H_A is that we assume a *negative* relationship - that is, individuals with a higher income advocate *less* strongly social justice compared to individuals with a lower income (e.g., because they fear a loss of income due to redistributive policies).

Formally, H_A can be represented with reference to either a correlation coefficient $r < 0$ or a regression coefficient $\beta_1 < 0$. Accordingly, the null hypothesis would be $\beta_1 \geq 0$.

Question: Would it also make sense to formulate a different hypothesis H_A?

Solution:

Yes: For example, the higher the income, the higher the preference for social justice (since wealthier people can afford it). Which hypothesis makes sense depends on theory and previous research.

In a next step, the bivariate regression model is estimated. To do so, we use the index `morejustice` as the outcome and the variable `income` as an independent variable (or predictor), which measures income classes in 26 levels (e.g., 1 = under 200 EUR monthly net income; 14 = 1750 - 1999 EUR; 26 = 10000 EUR and more).

```
model_biv <- lm(morejustice ~ 1 + income, data = data_allbus) # 1 refers to the constant
                                                              # or intercept
summary(model_biv)
```

```
##
## Call:
## lm(formula = morejustice ~ 1 + income, data = data_allbus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12251 -0.49371  0.06216  0.46936  1.33916
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.140981   0.045227  69.449  < 2e-16 ***
## income      -0.018467   0.002833  -6.518 8.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6574 on 2593 degrees of freedom
```

```
## Multiple R-squared:  0.01612,    Adjusted R-squared:  0.01574
## F-statistic: 42.48 on 1 and 2593 DF,  p-value: 8.532e-11
```

```
stargazer(model_biv, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              morejustice
## ------------------------------------------------
## income                        -0.018***
##                                (0.003)
##
## Constant                      3.141***
##                                (0.045)
##
## ------------------------------------------------
## Observations                    2,595
## R2                              0.016
## Adjusted R2                     0.016
## Residual Std. Error       0.657 (df = 2593)
## F Statistic           42.484*** (df = 1; 2593)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Task: Interpret the output from the regression model using the following heuristics.

"The empirical relationship between income and attitudes toward justice is positive/negative."

"The relationship is/is not statistically significant."

Use the technically correct interpretation of p-values: "The probability of finding such a relationship in the population, . . . "

"Thus, the null hypothesis can be rejected/not be rejected."

The detailed interpretation of the regression coefficient for income is: ". . . "

Interpret additionally the constant and R2 from the model output. What is the value of the constant in the unemployment-crime example (Model 1)? What information content does it carry?

Solution:

The empirical relationship between income and attitudes toward justice is negative.

The found relationship is statistically significant (given a probability of error of 0.01 or 1% due to ***).

The probability of finding such a relationship in the population, even though the null hypothesis is true, is less than 1%. Thus, the null hypothesis can be rejected.

The detailed interpretation of the regression coefficient for income (b=-0.018) is as follows: If income increases by one unit (income category), then (variant 1:) attitudes toward justice decrease (on average) by 0.018 units (scale points) / (variant 2:) changes by -0.018 units.

Constant: An interpretation is tedious because the income variable we used has no zero point. (If we would center the income variable at the mean value, then 0 indicated an average income, and the constant was the predicted level of justice attitudes for individuals with an average income.)

R2 is a measure of model fit and indicates how well the included independent variables explain the outcome. It can be interpreted as the ratio of explained variance by the model to the total variance: R2=0.016 -> 1.6% of the total variance of the outcome variable can be explained by the predictor variables in the model. If R2 = 0.3, then the variance explained by the model would be 30%.

Note: The so-called *adjusted R2* is used to compare models with a different number of predictor variables. While it can no longer be interpreted as % of variance explained, it can be compared across models (better model with higher Adj-R2).

On the correspondence of correlation coefficient and standardized regression coefficient in the bivariate case

Pearson's correlation coefficient is the standardized covariance between two variables and corresponds to the so-called *standardized* regression coefficient from bivariate regression. To obtain the standardized regression coefficient, we can either z-transform the variables ($z = \frac{(x-\bar{x})}{S_x}$) before they enter the regression model or transform the coefficient after estimation by multiplying the unstandardized regression coefficient by the standard deviation of X and dividing the result by the standard deviation of Y: $\beta_s = \frac{(\beta * S_y)}{S_x}$. In the present case, we obtain the standard deviations and then calculate the standardized coefficient by hand, as well as use the `scale` function in R that automatically z-transforms variables and then estimates the regression model. We will check whether both procedures and the correlation correspond to each other.

```
# Standardized regression coefficient by hand
sdx <- sd(data_allbus$income)
sdy <- sd(data_allbus$morejustice)

beta_s <- (-0.018281 * sdx) / sdy
beta_s
```

```
## [1] -0.1256859
```

```
# Standardized regression coefficient with model
model_biv_s <- lm(scale(morejustice) ~ 1 + scale(income), data = data_allbus)
summary(model_biv_s)
```

```
##
## Call:
## lm(formula = scale(morejustice) ~ 1 + scale(income), data = data_allbus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2030 -0.7450  0.0938  0.7083  2.0209
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.553e-17  1.948e-02   0.000        1
## scale(income) -1.270e-01  1.948e-02  -6.518 8.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9921 on 2593 degrees of freedom
## Multiple R-squared:  0.01612,    Adjusted R-squared:  0.01574
## F-statistic: 42.48 on 1 and 2593 DF,  p-value: 8.532e-11
```

```
# Calculate correlation
vars <- c("income", "morejustice")
```

```r
cor.vars <- data_allbus[vars]
rcorr(as.matrix(cor.vars))
```

```
##             income morejustice
## income        1.00       -0.13
## morejustice  -0.13        1.00
##
## n= 2595
##
##
## P
##             income morejustice
## income              0
## morejustice  0
```

> **Interpret the standardized regression coefficient as follows: If X (income) increases by one *standard deviation*, then Y increases/decreases by β_s-*standard deviations*. Here, it is a decrease by 0.13 *standard deviations*.**

Standardized regression coefficients are also widely used in multiple regression (i.e., regressions with more than one independent variable). The advantage is that we "free" the variables from their original scale of measurement and use "standard deviations" as a comparable scale of measurement. We can thus directly compare regression coefficients. If a coefficient is larger, then it also has a larger "effect" on the outcome variable. For example, an increase of one unit on a scale of income that is measured on a scale ranging between 0 and 10,000 would result in a smaller change in the outcome variable than if it was measured on a scale with only 12 categories. Using standardized versions of each variable and using these two versions in two separate regression models should lead to comparable results. As an example of competing explanations that are represented by two independent variables in the same model: If income has a standardized coefficient of -0.13 and age has a standardized coefficient of -0.05, then income has a "greater effect" (on the outcome variable) than age.

Note that we typically report z-standardized coefficients only for metric (but not binary) predictor variables and that there are other routines for comparing coefficients, such as the min-max normalization where all variables have a minimum of 0 and a maximum of 1 (and ordinal or metric variables additional values in between 0 and 1).

Multiple Regression

Basic Idea

The goal of multiple regression is to explain the outcome variable with multiple independent variables. This is useful when we want to test different explanations for a phenomenon (e.g., attitudes toward justice). The multiple regression model produces *adjusted* estimates because the independent variables control for their mutual influences. Recall the example of unemployment and crime from the first case study, where we also included population density as an additional variable. The interpretation of the influence of unemployment (on crime) is in this case adjusted for the influence of population density. For unemployment, we obtained a coefficient estimate of $\beta_1 = 215.4$. As a thought experiment: If we think of municipalities with the same population density and compare them with each other, then the municipalities with one percentage point more unemployment have on average a 215 higher crime rate per 100T inhabitants. The influence of population density on crime can be interpreted in a corresponding way - also here, the influence of the other factor (unemployment) is "held constant" or is controlled for.

This mutual "holding constant" or "controlling" in the multiple regression model means that we can account for alternative explanations with additional variables. In the example of attitudes toward justice, we use the variables gender (`female` with the categories male = 0 and female = 1), whether or not respondents were unemployed in the last 10 years (`unemployed` with the categories not unemployed = 0 and been unemployed =

1), and whether respondents live in West or East Germany (`east` with the categories West Germany = 0 and East Germany = 1) as additional variables. Regarding the influence of gender, we would assume that women are more supportive of social justice than men due to their socialization and societal norms and/or because they earn less than men, on average, due to the gender pay gap. Hence, it is important to control for gender so that income does not partially "transport" a gender effect. Controlling for gender in the multiple regression model allows us to obtain a "net" income effect that is separated from the effect of gender that otherwise could confound the relationship between income and attitudes toward justice.

Question: How relate past or current unemployment and living in East Germany to both variables of interest income and attitudes toward justice?

Solution:

Unemployment: If someone was/is unemployed, he/she is expected to advocate social justice more strongly (e.g. due to self-interest) than someone lacking such an experience. If someone is unemployed, he/she has a lower income, on average, compared to someone who is employed.

East Germany: East Germans are expected to advocate social justice more strongly than West Germans because of a greater sense of disadvantage or their socialization in the socialist GDR. At the same time, residents of East Germany have a higher probability to earn less money in comparable jobs due to a less developed economic situation.

Formulating hypotheses

The multiple regression model with four independent variables can be represented with the following formula:

$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + e$

Y refers to the outcome variable, β_0 is the constant, the other β_i are the regression coefficients of the corresponding predictor variables x_i. e refers to the residual or the part of the variance of the outcome variable that is not explained by the predictor variables in the model.

If we use the variable names for this, the formula for the specific model would be the following:

$morejustice = \beta_0 + \beta_1 * income + \beta_2 * female + \beta_3 * unemployed + \beta_4 * east + e$

Question: Formulate the null and alternative hypotheses for each independent variable.

Solution:

Income H_A: $\beta_1 < 0$ (Income is negatively related to attitudes toward justice)

H_0: $\beta_1 \geq 0$ (Income is unrelated or positively related to attitudes toward justice)

Female H_A: $\beta_2 > 0$ (Being female is positively related to attitudes toward justice, which is the same as women advocate social justice more strongly than men)

H_0: $\beta_2 \leq 0$ (Being female is unrelated or negatively related to attitudes toward justice)

Unemployed H_A: $\beta_3 > 0$ (Unemployment is positively related to attitudes toward justice, which is the same as people who experienced unemployment advocate social justice more strongly than those without such an experience)

H_0: $\beta_3 \leq 0$ (Unemployment is unrelated or negatively related to attitudes toward justice)

East Germany H_A: $\beta_4 > 0$ (Living in East Germany is positively related to attitudes toward justice, which is the same as people from East Germany advocate social justice more strongly than those from West Germany)

H_0: $\beta_4 \leq 0$ (Living in East Germany is unrelated or negatively related to attitudes toward justice)

Estimation and result interpretation

We now estimate the corresponding regression model. The bivariate model has already been estimated above and is included for comparison.

```
model_mult <- lm(morejustice ~ 1 + income + female + unemployed + east,
                 data = data_allbus)

stargazer(model_biv, model_mult, type = "text")
```

```
##
## ===============================================================
##                                  Dependent variable:
##                      ---------------------------------------
##                                       morejustice
##                             (1)                    (2)
## ---------------------------------------------------------------
## income                   -0.018***              -0.011***
##                           (0.003)                (0.003)
##
## female                                           0.128***
##                                                  (0.028)
##
## unemployed                                       0.066**
##                                                  (0.033)
##
## east                                             0.075***
##                                                  (0.028)
##
## Constant                  3.141***               2.936***
##                           (0.045)                (0.059)
##
## ---------------------------------------------------------------
## Observations               2,595                  2,595
## R2                         0.016                  0.027
## Adjusted R2                0.016                  0.026
## Residual Std. Error  0.657 (df = 2593)      0.654 (df = 2590)
## F Statistic         42.484*** (df = 1; 2593) 18.263*** (df = 4; 2590)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

We can now insert the regression coefficients into the two formulas from above:

General form: $y = 2.936 - 0.011 * x_1 + 0.128 * x_2 + 0.066 * x_3 + 0.075 * x_4 + e$

Specific form: $morejustice = 2.936 - 0.011 * income + 0.128 * female + 0.066 * unemployed + 0.075 * east + e$

Question: Interpret the coefficients from the multiple regression model in terms of (a) direction of association, (b) significance, and (c) effect size.

Solution:

Income

> *Negative relationship between income and attitudes toward justice, which is statistically significant ($p < 0.01$). An increase in income by one unit is associated with a* ***decrease*** *in attitudes toward justice by 0.011 scale units (while keeping the influence of the other variables in the model constant). In terms of effect size, moving from the lowest to the highest income category would be associated with a decrease in attitudes toward justice by 0.286 scale units (26 categories* 0.011 for the coefficient = 0.286, which corresponds to the effect of a min-max normalized income variable).*

Female

> *Positive relationship between being female and attitudes toward justice, which is statistically significant ($p < 0.05$). An increase in the variable female by one unit (= the difference between women and men) is associated with an* ***increase*** *in attitudes toward justice by 0.128 scale units (while keeping the influence of the other variables in the model constant).*

Unemployed

> *Positive relationship between being unemployed in the last 10 years and attitudes toward justice, which is statistically significant ($p < 0.05$). An increase in the variable unemployment by one unit (= the difference between those who were unemployed and those who were not) is associated with an* ***increase*** *in attitudes toward justice by 0.066 scale units (while keeping the influence of the other variables in the model constant).*

East Germany

> *Positive relationship between living in East Germany and attitudes toward justice, which is statistically significant ($p < 0.05$). An increase in the variable east by one unit (= the difference between people from East vs. West Germany) is associated with an* ***increase*** *in attitudes toward justice by 0.075 scale units (while keeping the influence of the other variables in the model constant).*

In terms of effect comparison, the variables female, unemployed, and east are measured on a comparable scale (0/1). The min-max effect of income is also comparable as this corresponds to recoding the variable to range between 0 and 1. Hence, income is comparatively the strongest predictor in the multiple regression model.

Question: What changed between the bivariate and the multiple model?

Solution:

In the multiple model (compared to the bivariate model), the coefficient of income (or the relationship between income and attitudes toward justice) is smaller. The underlying reason is that in the bivariate model, the coefficient partly represented unobserved factors, which is corrected in the multiple regression model (at least with respect to the variables included in the model).

In terms of model fit, the multiple model has a better fit to the data (since adj. R2 is higher).

OLS assumptions and model diagnostics, non-linear relationships between variables, and interactions

OLS assumptions

To obtain unbiased and efficient coefficient estimates from OLS regression models, several assumptions must be met. These assumptions refer to three areas.

1. The first area ensures that the **estimation method** works at all. The first assumption is that the *coefficients* are additive because otherwise, the coefficients would not be estimable by the least squares method. This assumption is trivial and usually holds. Note that some statistic books erroneously cite this assumption that the empirical relationship between the independent and dependent variables must be linear. This is not the case and multiple linear regression can easily address non-linear relationships by multiplying *variables* with each other (see also below). The second assumption relates to the collinearity of independent variables. This means that two independent variables must not be perfectly correlated. Otherwise, the influence of one variable would be completely displaced by the influence of the other variable, which inhibits the reliable estimation of coefficients. It is rather rare that variables are too highly correlated with each other, and if there are variables in the dataset that measure similar constructs (and are thus highly correlated), one would build an index from those variables, anyway.

Assumption	Definition	Why?
Linearity in parameters	$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k + x_{ik} + e_i$	Estimation method
No perfect collinearity	x_{ik} not constant, no linear function of $x_{il}, k \neq l$	Estimation method
Exogeneity of predictor variables	$E(e\|x_1, \ldots, x_k) = 0$	Unbiased estimates
Data generating process	Random sample	Foundation for statistical inference
Uncorrelated errors (no serial correlation)	$E(e_i, e_s\|x_1, \ldots, x_k) = 0, \quad i \neq s$	Statistical inference (efficiency)
Normally distributed errors	$e \sim N(0, \sigma^2)$	Statistical inference (efficiency)
Homoscedasticity	$Var(e\|x_1, \ldots, x_k) = \sigma^2 = const.$	Statistical inference (efficiency)

2. The second area concerns the unbiasedness of coefficient estimates. The underlying assumption is **exogeneity of independent variables**. If a variable is exogenous, this ensures that the coefficient of this variable reflects the actual relationship and does not represent something else. This is defined as the error term (i.e., the variance of the outcome left unexplained which represents possible unobserved factors that yield an influence on the outcome) is unrelated to the predictor variables. In a randomized experiment, the treatment variable is an exogenous predictor (per design due to the randomization) that is unrelated to unobserved third variables. In a model with observational data (e.g., from a survey), the exogeneity assumption requires the exclusion of alternative explanations by including relevant control variables. If we have all relevant control variables in the model, this assumption would be met. However, if there are unobserved variables that are correlated with the predictor and the outcome variable (which is the rule rather than the exception), we have a problem: the exogeneity assumption is not met and the regression coefficient is biased. We cannot empirically determine the extent of this bias: *There are no diagnostic tests for this assumption.* This means that we need to do good theory work to get clarity about causal relationships between variables and to include relevant control variables in the regression

model. Another way to go is - as already mentioned - using experimental designs to rule out alternative explanations.

Assumption	Definition	Why?
Linearity in parameters	$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k + x_{ik} + e_i$	Estimation method
No perfect collinearity	x_{ik} not constant, no linear function of $x_{il}, k \neq l$	Estimation method
Exogeneity of predictor variables	$E(e\|x_1, \ldots, x_k) = 0$	Unbiased estimates
Data generating process	Random sample	Foundation for statistical inference
Uncorrelated errors (no serial correlation)	$E(e_i, e_s\|x_1, \ldots, x_k) = 0, \quad i \neq s$	Statistical inference (efficiency)
Normally distributed errors	$e \sim N(0, \sigma^2)$	Statistical inference (efficiency)
Homoscedasticity	$Var(e\|x_1, \ldots, x_k) = \sigma^2 = const.$	Statistical inference (efficiency)

3. The third area refers to the process of **statistical inference**. If the assumptions are violated, the standard errors and thus the test statistics (e.g., t/z statistics), and significance tests are biased. The unbiasedness of the coefficients (as in 2.) operates independently of this. The first assumption in this area relates to whether we work with a random sample. Without one, there is no basis for inferential statistics, although in practice inferential statistics are done even without a random sample. (Here one uses, for example, significance tests as information about how "reliable" or precise an empirical relationship is estimated.) The assumption of **uncorrelated errors** means that each observation must carry a unique piece of information, independent of the information content of other observations in the sample. If we use, for example, cross-national survey data from different countries, observations within countries are possibly similar to each other regarding the outcome variable. For panel data, where the same units are observed multiple times over time, the similarity between observations from the same units is labeled serial correlation. If such a similarity between observations is not addressed by specific modeling choices (e.g., using multilevel models) or the variables included in the regression model, standard errors are biased. There are formal tests of uncorrelated errors, such as the Durbin-Watson Test. The assumption that the **errors must be normally distributed** often serves as a rationale for assessing the distribution of variable values using a histogram and, if necessary, for transforming a variable so that its distribution is more "normal" (e.g., logarithmitize a right-skewed variable). This can be a useful strategy. At the same time, it is important to note that essentially the assumption refers to the distribution of errors (i.e., the outcome after accounting for variables in the model) and that with a sufficiently large number of observations, a violation becomes less problematic in terms of biased standard errors. Finally, the assumption of **homoscedastic standard errors** states that the distribution of errors must be the same (or constant) for different values of the independent variable(s). If the assumption is violated and *heteroscedasticity* is present, then the standard errors are biased. Typically (but by no means always), there is an underestimation of the standard errors, which in turn is associated with overly optimistic significance tests (and type I errors). That is, we find a statistically significant result (and reject the null hypothesis) even though it is not actually a significant result. As a remedy, so-called robust standard errors can be applied.

Assumption	Definition	Why?
Linearity in parameters	$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k + x_{ik} + e_i$	Estimation method
No perfect collinearity	x_{ik} not constant, no linear function of $x_{il}, k \neq l$	Estimation method
Exogeneity of predictor variables	$E(e\|x_1, \ldots, x_k) = 0$	Unbiased estimates
Data generating process	Random sample	Foundation for statistical inference
Uncorrelated errors (no serial correlation)	$E(e_i, e_s\|x_1, \ldots, x_k) = 0, \quad i \neq s$	Statistical inference (efficiency)
Normally distributed errors	$e \sim N(0, \sigma^2)$	Statistical inference (efficiency)
Homoscedasticity	$Var(e\|x_1, \ldots, x_k) = \sigma^2 = const.$	Statistical inference (efficiency)

Model diagnostics

Collinearity between variables -> VIF To illustrate how to employ several existing tests of OLS assumptions, we use the above-estimated model *model_mult*.

The variance inflation factor (VIF) provides information on whether variables in the model are excessively correlated with other variables (this would also be registered in a correlation matrix of the variables, and one should be cautious if variables correlate with each other by more than 0.85). Regarding the VIF, a VIF above 5 is already a bit suspicious and a value above 10 indicates strong collinearity.

```
# Regression diagnostics
vif(model_mult)
```

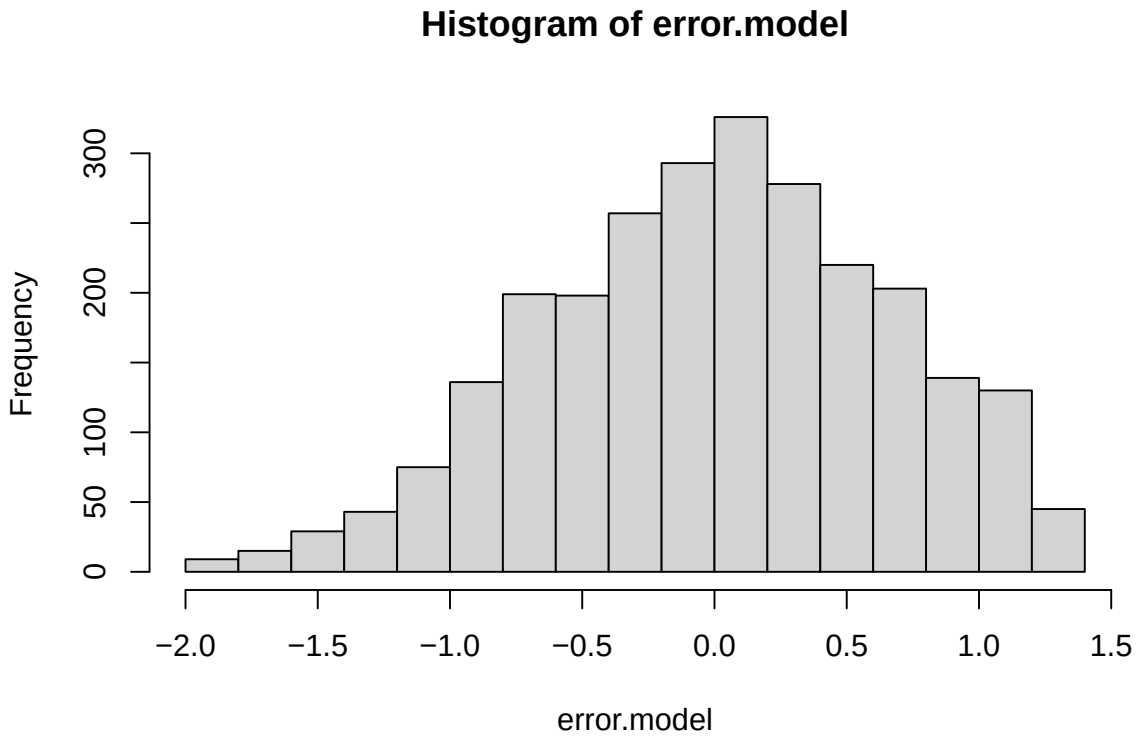
```
##     income     female unemployed       east
##   1.215933   1.175482   1.037561   1.014182
```

The VIF is modest (< 5) for all variables in the model. We can conclude that there is no multicollinearity present.

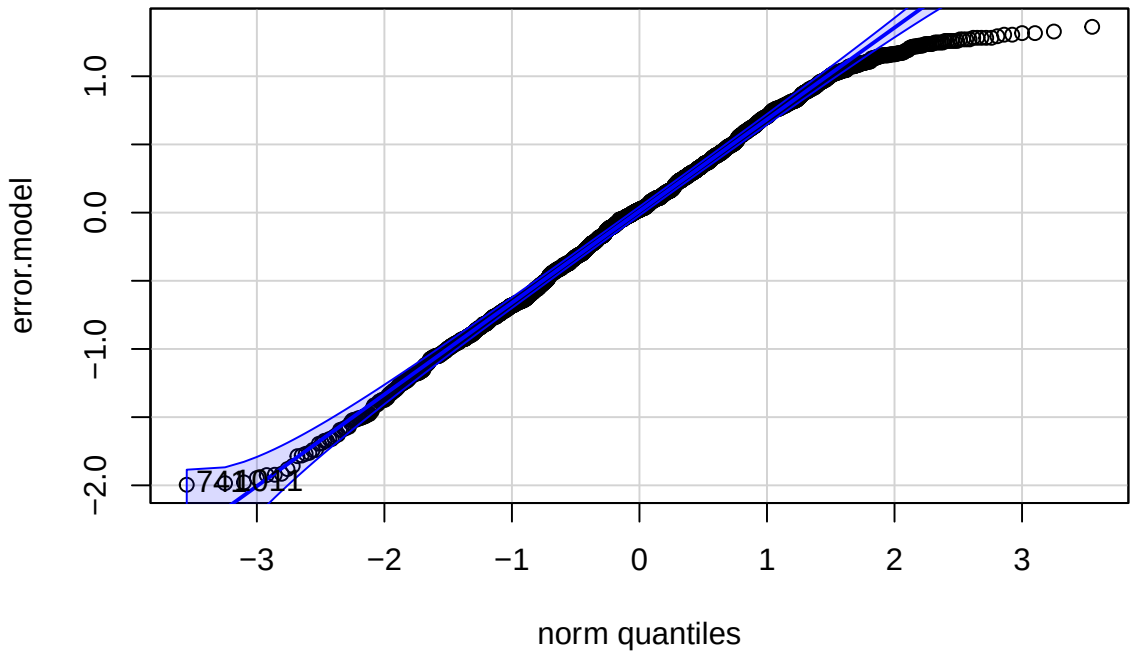
Uncorrelated errors Since we are using a one-stage random sample, this assumption is most likely met.

Normality of the residuals As a test for normality, we can (a) look at the distribution of residuals using a histogram, (b) plot a so-called Q-Q plot and interpret it, or (c) look at the distribution of variables in the model using a histogram (note that this can only give an indirect cue about the distribution of errors).

```
predict.model <- fitted(model_mult)
error.model <- residuals(model_mult)

hist(error.model)
```




```
qqPlot(error.model)
```



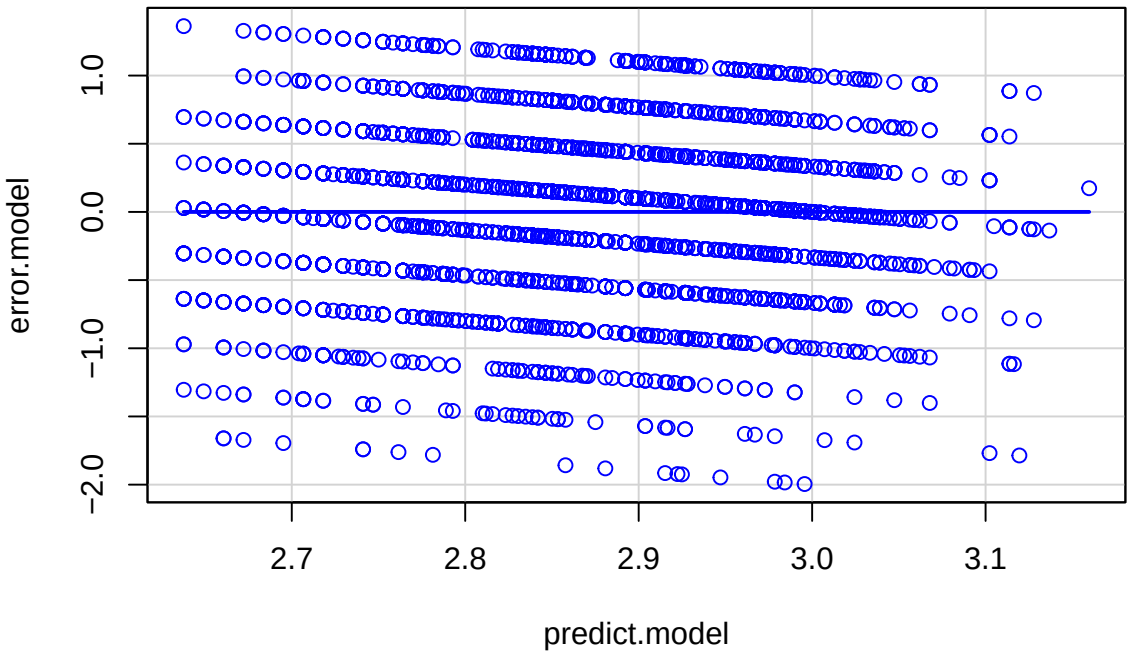

```
## [1]  741 1011
```

From the visual inspection, the distribution of errors looks similar to a normal distribution. The center of the distribution is about zero and the distribution is only slightly left-skewed.

Regarding the Q-Q plot, the residual values shown should be on (or close to) the line as best as possible. A bit of "fraying" for high or low values is acceptable. We see slight deviations on the right, so there are a little too few values in the positive extreme range. Otherwise, it looks fine.

Homoscedasticity Next, we inspect the assumption of homoscedasticity. We compare the predicted values from the model with the residual values. The more evenly the dots are scattered around the blue line, the more likely we can assume that the assumption is met.

```
scatterplot(error.model ~ predict.model, regLine = TRUE, smooth = FALSE, boxplots = FALSE)
```



In our case, the errors scatter quite evenly. This means that the model fits similarly at low, medium, and high prediction values. To be certain, we could plot the (standardized) residuals for all independent variables in the model and their ranges of values. Furthermore, there is a statistical test for homoscedasticity: the *Breusch-Pagan test*. The null hypothesis here would be that homoscedasticity exists (i.e., the residuals are independent of the predictors). A non-significant result is therefore the desired result of the test.

```
bptest(model_mult)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_mult
## BP = 5.4228, df = 4, p-value = 0.2466
```

The non-significant result indicates that the assumption of homoscedasticity appears to be met. If this was not the case, we would have to work with robust standard errors.

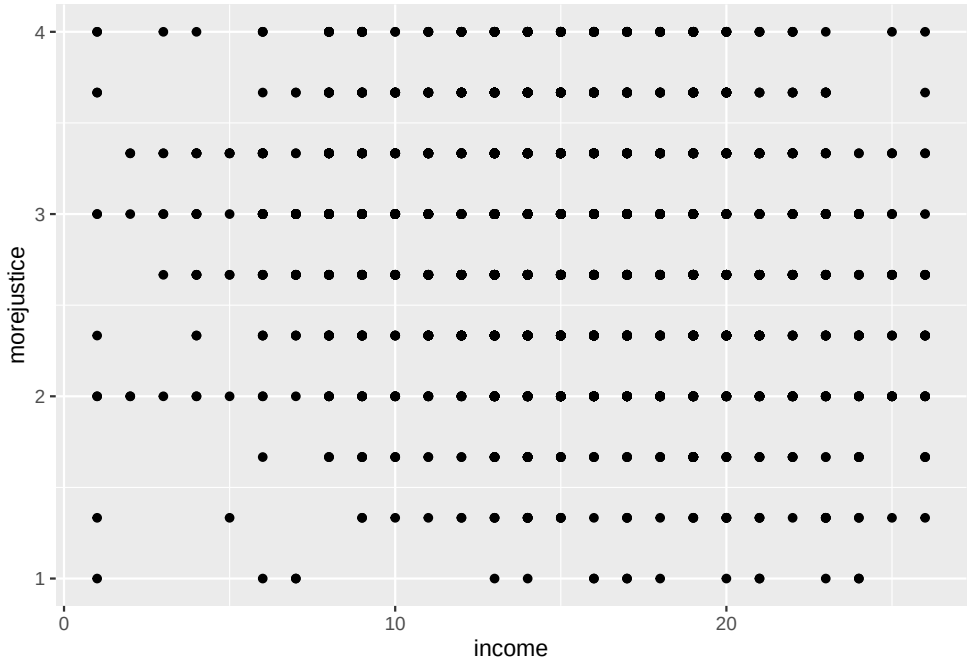
Non-linear effects

Including a non-linear variable relationship in a regression model can be based on several motives. One reason may be that we may notice from a scatter plot that the relationship between two variables is u-shaped or inversely u-shaped. Another reason might be that we can theoretically reason that the relationship is non-linear. A third motivation is to include non-linear relationships to reduce heteroskedasticity.

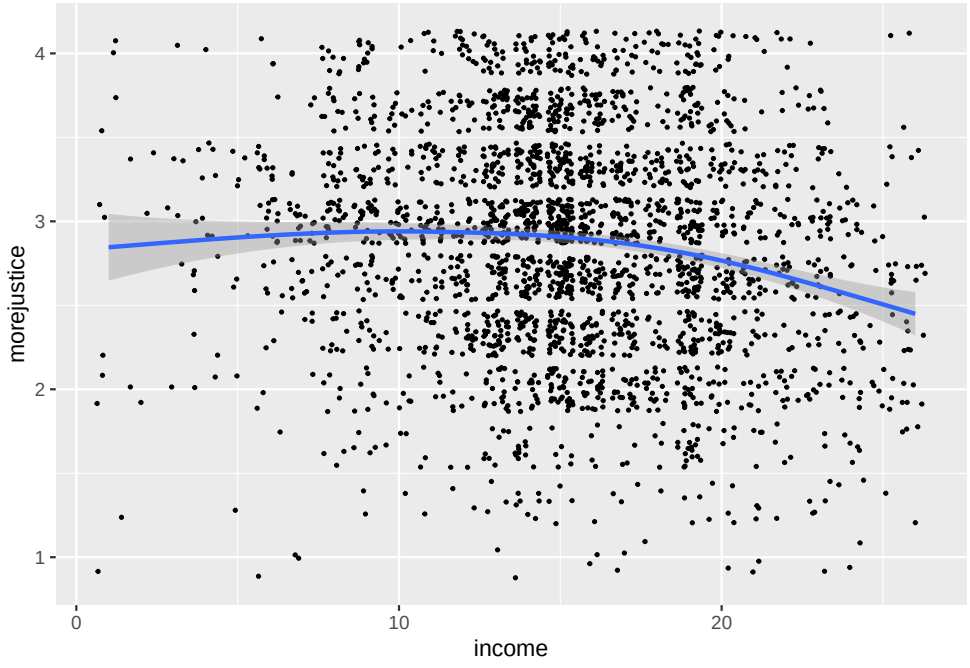
Below, we test whether or not income has a non-linear effect on attitudes toward justice. Specifically, we assume that particularly middle-class individuals prefer justice and poor and rich individuals do so less. For rich people, the reason could be self-interest, while people with a low income might oppose redistributive policies because they are afraid of welfare state overuse by newcomers such as immigrants. The latter phenomenon has been discussed in the literature as "welfare chauvinism."

Let us first look at the relationship graphically.

```
# Create and print scatter plot
sc1 <- ggplot(data = data_allbus, aes(income, morejustice)) +
  geom_point()

sc1
```




```
sc2 <- ggplot(data = data_allbus, aes(income, morejustice)) +
  geom_jitter(aes(income = morejustice), size = 0.5) + geom_smooth()

sc2
```



In the first figure, one cannot see a pattern because many points are simply stacked on top of each other. To make the density of the points more visible, we can add a small random fluctuation. This makes the plot somewhat less accurate, but we can now see the pattern, which is also illustrated by the smoothed correlation function (and confidence intervals) represented by the blue line. There is some visual evidence for a non-linear relationship.

We now test for the non-linear relationship in the regression model. To do so, we add the squared variable for income as an additional predictor in the multiple model. The idea here is that the effect of income now depends on the value of "another" variable, namely income itself. In other words, the effect of income on attitudes toward justice is now allowed to vary depending on where you are on the income scale.

```
data_allbus$income_squared <-  data_allbus$income * data_allbus$income

model_nlin <- lm(morejustice ~ 1 + income + income_squared + female + unemployed + east,
                 data = data_allbus)

stargazer(model_biv, model_mult, model_nlin, type = "text", omit.stat = "f")
```

```
##
## ================================================================================
##                                        Dependent variable:
##                          ----------------------------------------------------
##                                             morejustice
##                                (1)               (2)               (3)
## --------------------------------------------------------------------------------
## income                      -0.018***         -0.011***         0.048***
##                              (0.003)           (0.003)           (0.013)
##
## income_squared                                                  -0.002***
##                                                                 (0.0004)
##
## female                                        0.128***          0.125***
##                                               (0.028)           (0.028)
##
## unemployed                                    0.066**           0.055*
##                                               (0.033)           (0.033)
##
## east                                          0.075***          0.058**
##                                               (0.028)           (0.028)
##
## Constant                    3.141***          2.936***          2.546***
##                              (0.045)           (0.059)           (0.101)
##
## --------------------------------------------------------------------------------
## Observations                  2,595             2,595             2,595
## R2                            0.016             0.027             0.036
## Adjusted R2                   0.016             0.026             0.034
## Residual Std. Error 0.657 (df = 2593) 0.654 (df = 2590) 0.651 (df = 2589)
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:

- *First, we look if the coefficient for the squared variable* `income_squared` *is statistically significant ->* ***yes, so we can infer a non-linear relationship***
- *Next, we look at the sign of the coefficient of* `income` *->* ***positive***
- *And the sign of the coefficient of* `income_squared` *->* ***negative***
- *To summarize: The effect of income on justice attitudes is* ***positive*** *and* ***decreases*** *with higher values of income.*

Question: How would the effect be interpreted if both signs were negative?

Solution:

The effect of income on attitudes toward inequality would be negative and decrease for high values of income.

Interaction between two different variables

Another important tool in regression analysis is to model interactions between variables. The idea is that the effect of one variable on the outcome variable depends on the *value* of another variable in the model. Specifically, the question we ask here is whether the effect of gender on attitudes toward justice differs between West and East Germany. The assumption would be that in East Germany the difference between men and women in justice attitudes is less pronounced than in West Germany. On the one hand, this may be because men in East Germany are more supportive of justice (and thus have similarly high levels of support compared to women). On the other hand, it could be (but this is implausible) that women in the East have similarly "low" values as men.

Let us look at this as an interaction in the regression model. For this, it is important to have the two variables `female` and `east` in the model, as well as a variable that is the product of both variables (interaction). This variable can be built beforehand or we build it "on the fly" in R by adding a ":" between two existing variables ("*" also works).

```r
# Letting R know that the two nominal variables "female" and "east" are indeed nominal
# or "factor variables" is important for the graphical representation of the interaction
# below

data_allbus$female <- factor(data_allbus$female)
data_allbus$east <- factor(data_allbus$east)

# The interactions can be specified with ":" or "*".
model_interact <- lm(morejustice ~ 1 + income + unemployed + female + east + female:east,
                     data = data_allbus)

stargazer(model_interact, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                              morejustice
## ------------------------------------------------
## income                        -0.011***
##                                (0.003)
##
## unemployed                     0.065*
##                                (0.033)
##
## female1                       0.163***
##                                (0.033)
##
## east1                         0.128***
##                                (0.040)
##
## female1:east1                  -0.106*
##                                (0.056)
##
## Constant                      2.912***
##                                (0.060)
##
## ------------------------------------------------
## Observations                    2,595
## R2                              0.029
## Adjusted R2                     0.027
## Residual Std. Error       0.654 (df = 2589)
## F Statistic           15.351*** (df = 5; 2589)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

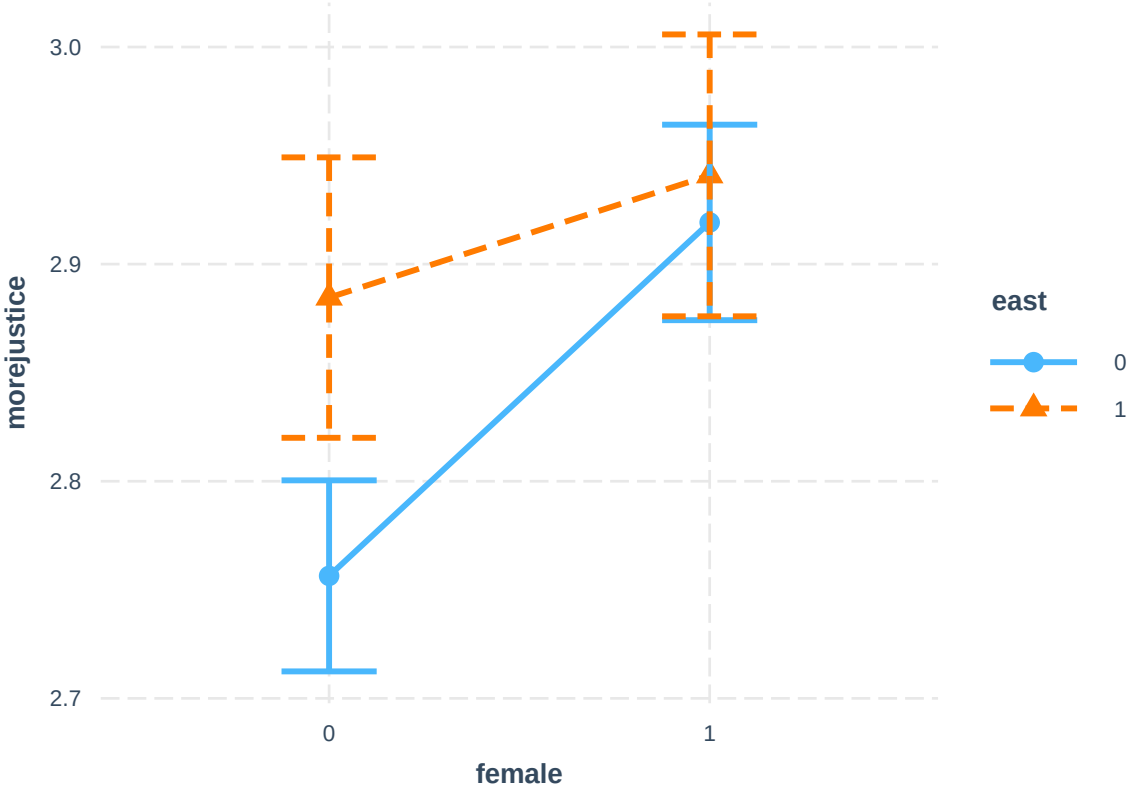
Interpretation:

- *First, we look to see if the coefficient for the interaction variable* `female1:east1` *is statistically significant ->* ***Yes, so we can assume that the effect of*** `female` ***on*** `morejustice` ***depends on the value of*** `east`***. In other words, the gender difference in justice attitudes differs between the two parts of Germany. (Note: It would also be correct to interpret the interaction effect symmetrically: The effect of*** `east` ***on*** `morejustice` ***depends on the value of*** `female`***. For simplicity, we do not pursue this interpretation here.)***

- *Then, we look at the sign and significance of the coefficient of* `female` *->* ***positive*** *(= positive effect of* `female` *if* `east` *= 0, i.e., in West Germany)*
- *Then, we look at the sign of the coefficient of the interaction term* `woman:east` *->* ***negative***
- *To summarize: The* ***positive*** *effect of being a woman on justice attitudes (= gender difference) is less strong ("increasingly* ***negative****") at higher values of* `east`*, i.e., in East Germany*

Below, we see a graphical representation of the package `interactions` (see https://interactions.jacob-long.com/index.html for more examples).

```
cat_plot(model_interact, pred = female, modx = east, geom = "line", point.shape = TRUE,
         vary.lty = TRUE)
```



The blue line represents West Germany, and the orange line East Germany. The lines with the bars at the end are 95% confidence intervals, thus indirectly showing how "significant" the respective effects are estimated to be.

Three things stand out:

- Women and men in East Germany are more supportive of social justice than women and men in West Germany (orange triangles compared to blue dots). The difference is statistically significant for men in West Germany (confidence intervals on the left side do not overlap). To know whether the overall level of support (women and men combined) is statistically significantly higher in the East, we have to go back to **model_mult** and look at the *average* relationship ($\beta_4 = 0.075, p < 0.01$).
- Looking at the lines and thus the effect of gender separately for both parts of the country, the difference between women and men is statistically significant in West Germany (see coefficient for `female`, 0.163). In East Germany, the difference is not significant. (This can also be visually inferred by comparing the orange and the blue estimates to each other.)
- The effect of being a woman is larger in West Germany than in East Germany (the blue line is steeper than the orange line). This is represented by the coefficient for the interaction term.

We can now insert the regression coefficients into a regression formula:

$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \beta_5 * x_3 * x_4 + e$

$morejustice = 2.912 - 0.011 * income + 0.065 * unemployed + 0.163 * female + 0.128 * east - 0.106 * female * east + e$

We can also derive the formulas for the so-called marginal effects by taking the first derivative and entering the variable values in a combination that informs us about the effects for the different groups.

1. Marginal Effect for the variable `east` (first derivative) $dy/d(east) = 0.128 - 0.106 * female$

- We can substitute 0 for the variable female, i.e., the effect of `east` for men: $b(east = 1, female = 0) = 0.128 - 0.106 * 0 = 0.128$.
- This refers to the difference between East and West in men (plot: distance between blue dot and orange triangle on the left).
- If we substitute 1 for the variable female: i.e., the effect of `east` for women: $b(east = 1, female = 1) = 0.128 - 0.106 * 1 = 0.022$.
- This refers to the small difference between East and West in women (distance between blue dot and orange triangle on the right side).

2. Marginal Effect for the variable `female` (first derivative) $dy/d(female) = 0.163 - 0.106 * east$

- We can substitute 0 for the variable east, i.e., the effect of `female` for West Germans: $b(female = 1, east = 0) = 0.163 - 0.106 * 0 = 0.163$.
- This refers to the difference between women and men in West Germany (plot: distance between both blue dots).
- If we substitute 1 for the variable east, i.e., the effect of `female` for East Germans: $b(female = 1, east = 1) = 0.163 - 0.106 * 1 = 0.057$.
- This refers to the difference between women and men in East Germany (distance between both orange triangles).

3. The effect difference of women (different in both slopes) is represented by the coefficient for the interaction, i.e., –0.106.

Thank you for engaging with the Introduction to Statistics and Data Analysis - A Case-Based Approach!

Feedback to conrad.ziller@uni-due.de is much appreciated.
