Quantitative Methods 1, ZHAW

Jürgen Degenfellner

2024-11-01

Contents

Chapter 1

Introduction

This script is a collection of notes and exercises for the course "Quantitative Methods 1" (Codename for: Statistical foundations) at ZHAW in Winterthur, Switzerland. It is permanently evolving and the github repository is public.

The **vision** is to write this script in **collaboration with students**. If you find any errors, have suggestions, or want to contribute, feel free to contact me. Which (online) content (video, book, blog…) helped you understand the topic at hand better? We should link those resources in the script!

The **goal** of this script is to provide a starting point for further reading and learning. It should function as an initial start to get you going. We are all learners.

Feel free to use any content for your own purposes.

At the end of each chapter, you will find a list of exercises. Last year's exam will be uploaded in time in the respective Moodle folder to help you prepare.

The most important lessons you'll learn are not part of the final exam: Intellectual honesty and humility.

A lot of content relevant for this course can be found in the ZHAW_teaching folder, which contains many R scripts (among other stuff).

We will focus a lot on descriptive statistics and basic concepts as it is a very important part of understanding data. Statistical modeling will be covered later.

One thing to consider for health sciences with respect to quantitative methods is the following: We do not *have* to use quantitative methods for each and every question arising in applied research (physiotherapy, midwifery, nursing, occupational therapy). Small sample sizes (e.g., n=10) often do not warrant the use of inferential statistics or estimation or at least make it considerably harder to answer the question at hand. *If* one decides to answer questions in a statistical way, we are bound to the rules of the game.

GPT4o and Github Copilot where used for writing this script.

1.1 Books I can (highly) recommend:

- FILL in HERE: book recommendation for Probability and descriptive statistics….
- Statistical Rethinking, YouTube-Playlist: Statistical Rethinking 2023
- (Free) Understanding Regression Analysis: A Conditional Distribution Approach
- (Free) Doing Bayesian Data Analysis
- Applied Regression Analysis and Generalized Linear Models

These books are well-written, approachable, and not overly technical.

For more theoretically advanced approaches, I can recommend:

- (Free) Bayesian Data Analysis
- (Free) Elements of Statistical Learning
- (Free) Unterstanding Advanced Statistical Methods

1.2 R

In this course, we use R as our main tool for data analysis. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS.

You may want to give Hadley Wickham's book a try:

- (Free) R for Data Science

Especially, the chapter on exploratory data analysis is very helpful.

This free book seems also very helpful for beginners:

- (Free) R for non-programmers R4NP

Further resources:

(Example) Introduction to R

In this video, many basic commands are explained. There are many more R introductions. My tip: Watch the beginning of a few different ones and see which explanations work best for you individually.

Overview of introductory resources

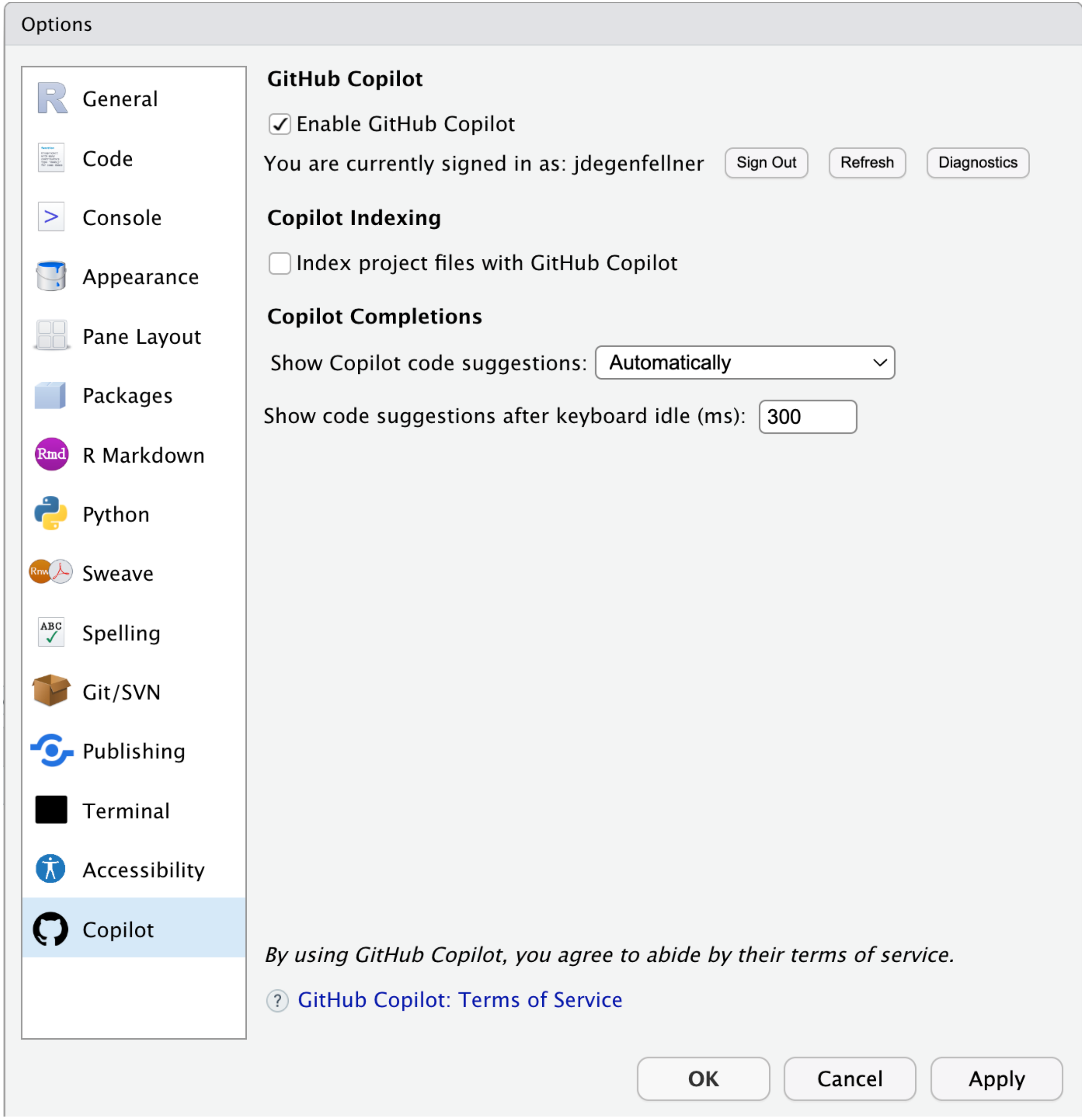
1.3 Additional Tools

Currently I use a combination of Github Copilot (paid), GPT4o (paid), and RStudio for writing code and this script. You can use Github Copilot directly in RStudio to make code suggestions, which is often increasing productivity. I can highly recommend using these tools in combination. GPT4o can process images as well, which is a game changer. As far as I know, there is no self-debugging version of GPT4o in combination with RStudio available yet (in the style of AutoGPT or similar approaches). Maybe, we'll work on that soon as a side project. For the more technically inclined among you, you can also use VSCode with the Github Copilot extension to write your R-Code.

Large Language Models (LLM) like GPT4o enable you to write/adapt code using natural language. Among other tasks, they help you create complicated, aesthetic plots. Very often, debugging attempts get stuck. They are far from perfect yet, but an impressive feat of engineering.

1.4 Workflow suggestion

You could use RStudio to write and execute your code. In the global options of RStudio (below), you can add your github copilot account. Additionally, you can have your GPT4o(1) window open and copy-paste images of your plots or code in order to achieve your current objective.



GitHub Copilot

☑ Enable GitHub Copilot

You are currently signed in as: jdegenfellner

Copilot Indexing

□ Index project files with GitHub Copilot

Copilot Completions

Show Copilot code suggestions: Automatically

Show code suggestions after keyboard idle (ms): 300

By using GitHub Copilot, you agree to abide by their terms of service.

? GitHub Copilot: Terms of Service

Chapter 2

Probability

Probability is a measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1 (or 0 to 100%), where 0 indicates impossibility and 1 indicates certainty, although we will see later that a probability of 0 does not necessarily mean that such an event can never occur. The higher the probability of an event, the more likely it is that the event will occur.

Why is probability important in our field of study (applied health sciences)?

Quantative research methods (often a code name for statistics) use probability theory to make statements about a larger population or a data generating process (DGP), as it should be more appropriately called.

In observational studies, we often make statements about associations between variables.

In experimental studies (e.g., a randomized controlled trial), we often try to make statements about the effect of an intervention on a certain outcome - for instance if a therapy lowers pain by at least 1 point better compared to usual therapies.

Probability theory has its roots in gambling and betting. Blaise Pascal wrote a letter to Pierre de Fermat in 1654 when a French essayist Antoine Gombaud, intrigued by gambling, sought to solve "the problem of points," first posed by Luca Paccioli in 1494. The problem asked how to fairly divide the winnings if a game is interrupted before its conclusion. Gombaud approached mathematician Blaise Pascal, who collaborated with Pierre de Fermat. Together, they laid the groundwork for modern probability theory. Fermat's method involved listing all possible outcomes and calculating each player's chance of winning, while Pascal developed a backward induction algorithm to assign probabilities. Their work revolutionized mathematics and influenced fields like economics and actuarial science.

Philosophically speaking, we could distinguish between two flavors of probability: Probabilities for events that are repeatable respectively have already happened, and probabilities for events that haven't happened yet.

An example for a repeatable event is getting a 6 when throwing a fair die. We can do this experiment right now by fetching a die and throwing it.

An example for the latter is the probability of a patient dying within the next 5 years after a certain diagnosis. It is hard to argue that this experiment would be repeatable under (almost) identical conditions since every patient is different whereas the dice are typically much more similar. Here, we could at least put forward that other similar patients have a certain proportion of dying within 5 years.

There are of course events that have not happend ever before, like the creation of artificial general intelligence (AGI). Nevertheless, one can still try to assign probabilities when such an event would happen.

2.1 Frequentist vs. Bayesian statistics

There are two main schools of thought in statistics: Frequentist and Bayesian. Often one hears that there is a “war” between the two.

It is *not* our place to say which one is better. Both have their strengths and weaknesses and are used in different contexts.

I would consider the rapant misuse of p-values and the cookbook-like application of frequentist statistics as a weakness of this approach (in its widely used form at least). Of course, this is not the method’s fault but the fault of the user.

Bayesian statistics is often considered more intuitive and flexible. It is also more computationally demanding and requires prior knowledge which is argued to be subjective. Computation time is sometimes still an issue in comparison for instance in regression modelling when using an end user laptop. It is also argued that for large sample size frequentist and Bayesian statistics converge to the same result.

There are very smart proponents on both sides and we will try to use and contrast both techniques throughout this script whenever convenient.

Especially one of the early eminent statisticians, Ronald Fisher, was an oponent of Bayesian statistics, or as he called it: “inverse probability”.

The only thing we are interested in is the practical application of both methods in the field of applied health sciences. How well can we describe data and make predictions, how well can we learn from data in our field?

2.1.1 Frequentist statistics

Frequentist statistics is based on the idea that probability is the long-run frequency of events. For instance, if I throw a fair die 1000 times, the frequency of getting a 3 is (approximately) $\frac{1}{6}$. In the limit, if I throw the die infinitely many times, the frequency of getting a 3 will converge to $\frac{1}{6}$. In mathematical notation, we would write

$$\mathbb{P}(\text{getting a 3}) = \lim_{n\to\infty} \frac{\text{Number of 3s in } n \text{ throws}}{n} = \frac{1}{6},$$

where $\mathbb{P}$ is the probability measure which we will define more formally later (see Exercise 1).

More genereally, in frequentist statistics, we are looking for a fixed but unknown parameter from an underlying data generating process (DGP). In the dice example, the process of repeatedly throwing the die is the data generating process. Basically, we could estimate the parameter of interest arbitrarily well by reapeated drawing from the DGP if we had enough data.

Example: Throw your (fair or unfair) die often enough and you will get a good estimate of the probability of getting a 3.

Example: We could try to estimate the mean birth weight of all babies from smoking parents born in Switzerland in 2022. We would draw a (random) sample of birthweights and calculate the mean. With a sample large enough, we could estimate this parameter fairly well. With all birthweights, we would know the true mean of the population of interest (for that year alone).

2.1.2 Bayesian statistics

Bayesian statistics, on the other hand, is based on the idea that probability is a measure of our uncertainty about an event or a parameter. Here, we use *prior* (i.e., before/outside of our experiment) knowledge about a parameter and update this knowledge with new data using the famous Bayes’ theorem:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta) \cdot p(\theta)}{p(\text{data})},$$

where:

- $p(\theta|\text{data})$ is the **posterior probability**: the updated probability of the parameter θ given the observed data.
- $p(\text{data}|\theta)$ is the **likelihood**: the probability of observing the data given a certain value of the parameter θ.
- $p(\theta)$ is the **prior probability**: the initial belief about the parameter θ before seeing the data.
- $p(\text{data})$ is the **marginal likelihood** or **evidence**: the probability of observing the data under all possible parameter values.

2.1.2.1 Example in applied health sciences (physiotherapy)

Suppose you're a physiotherapist trying to estimate the probability that a new therapy improves the mobility of patients with chronic back pain (Improvement Yes/No). You already have some prior knowledge (based on previous studies or expert opinions) that suggests the therapy works for 30% of patients. This is your **prior knowledge**: $\theta = 0.30$, where θ is the probability that the therapy is effective. Your colleagues are not convinced that the therapy is effective and argue that the probability is 40%. Now, you run a small trial with 50 patients and observe that 22 of them showed a clinically relevant improvement in mobility (self-reported from the patient). This new data (the result of the trial) *updates* your belief about the effectiveness of the therapy. Using Bayes' theorem (Exercise 2), you combine the prior knowledge $\theta = 0.30$ with the likelihood of the new data $p(\text{data}|\theta)$, and you calculate the **posterior probability**, $p(\theta|\text{data})$, which reflects your updated belief about the effectiveness of the therapy after observing the trial data. We could assign the probability of $\theta = 0.3$ or $\theta = 0.4$ equally: $p(\theta = 0.3) = p(\theta = 0.4) = 0.5$.

Bayesian analysis allows you to update your estimates as new evidence becomes available, providing a flexible framework for decision-making in health sciences.

2.2 Foundations of probability theory

We need to know some basic concepts of probability theory in order to dive in deeper. We will try to introduce them playfully and find formality as we go along. As stated above, in the frequentist sense, we are interested in the long-run frequency of events. How often does an event occur if we repeat the random experiment many times?

Heureka?

Let's imagine we are in a research department with 1000 researches all trying to answer the same question: Does the new physiotherapy work (e.g., reduce pain by 1 point better than the usual treatment)? Let's assume (unrealistically) that they are all working on this one question and they are not talking about their experiments or their research methodology to each other (assumption of independence). The statistician in the department has calculated (due to the variability of such treatment effects in the relevant population and theoretical considerations) that even under the assumption of the therapy is not working *at all* - **which we will assume for the time being -** , one would see an effect just *by chance* in 4% of the study results.

What would be considered a discovery under these cicumstances?

We now conduct an experiment. All 1000 ressearches are conducting a study with 50 patients to answer the same question. This is our random experiment (instead of throwing dice). Instead of throwing a fair die, we do a round of "research" with 1000 researchers. You as observer give the assignment to the researchers and come back as soon as all 1000 researeches have finished their experiments. Again, the are not taking to each other and we can (unrealistically) assume that their results will be not influenced by each other.

Now we could ask different questions:

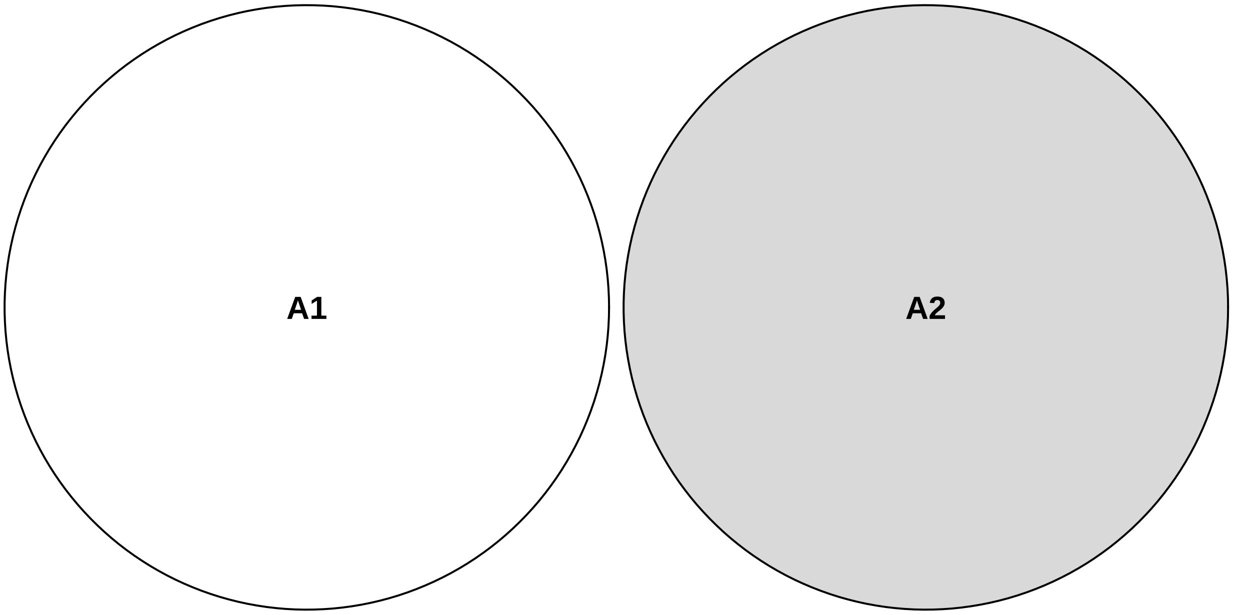
2.2.1 Questions about the 1000 researcher-experiment (among many others):

1. If you had to bet, how many experiments showed a treatment effect if you assume that the therapy is not working at all?
2. If you get 137 results showing a treatment effect, would you be surprised? Would you reject the assumption, that the therapy is not working at all? Why?
3. How many experiments (would you expect) showed a treatment effect if you assume that the therapy is "working" (positive result by chance) in 12% (instead of 4%) of the patients?

4. Assuming that you have 47 results showing a treatment effect and your marketing lead is asking you to write a press release stating that 47 out of 50 studies showed a treatment effect. What is the problem?
5. Assuming one very motivated researcher has tested 65 (secondary) hypotheses in her experiments and found 4 results that are difficult to explain by chance alone. What is the problem?
6. Suppose there are many large research departments in the world with 1000 researchers. How strongly would the number of positive results vary between these large departments?

We will try to answer these questions below.

First, it seems intuitive that Probability *within* an experiment should add up *if* the events are **disjoint**. The event A_1 = "only researcher 45 gets a positive result" and A_2 = "only researcher 897 gets a positive result" are mutually exclusive. If only researcher 45 finds an effect, then researcher 897 does not find an effect and vice versa. They cannot happen at the same time within that one experiment. Hence, the two events are said to be disjoint. If we add up the probabilities of all mutually exclusive events, we should get 1, or 100%. We say that the probability of all elementary events (called ω) sums to 1. Let's look at a Venn diagram to illustrate the concept of being mutually exclusive (disjoint).



Again, this refers to being mutually exclusive within our 1000-researcher experiment. Both events cannot happen at the same time in this context, so we assign 0 to the event that both occur simultaneously: $\mathbb{P}(A_1 \cap A_2) = 0$. The $\cap$-Symbol refers to all elementary events that are in both sets. In our case we have the sets

$$A_1 = \{(\dots, R_{45} = pos, \dots, R_{897} = neg, \dots)\}$$

$$and$$

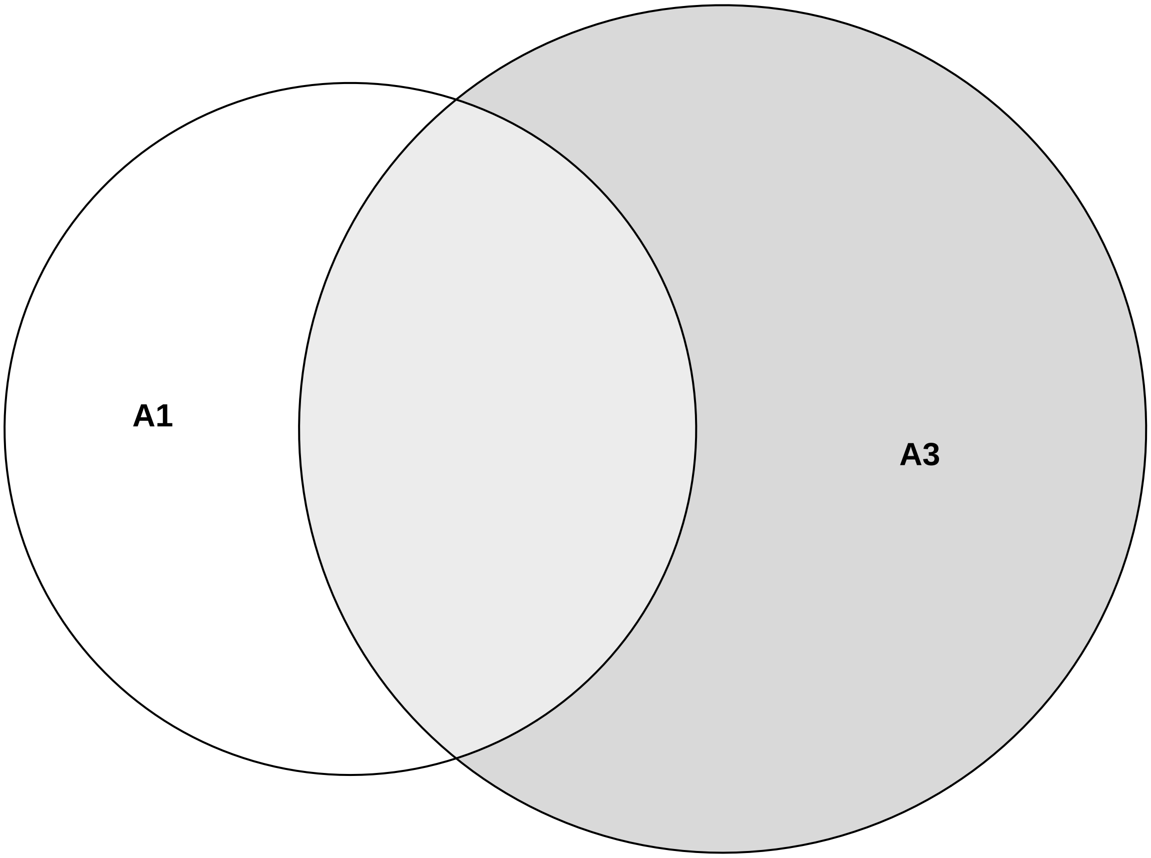
$$A_2 = \{(\dots, R_{45} = neg, \dots, R_{897} = pos, \dots)\}.$$

An example of **non-disjoint** events (within our 1000-researcher experiment) would be the event A_1 = "only researcher 45 gets a positive result" and the event A_3 = "only researcher 45 or only reasearcher 67 gets a positive result". Which researchers got a positive result in both events? The answer is: Researcher 45. Hence, the two events are said to be non-mutually exclusive. We can't just add up the probabilities (of events A_1 and A_3) here, since we would count the probability of researcher 45 twice. The sets look like this:

$$A_1 = \{(\dots, R_{45} = pos, \dots)\}$$

and

$$A_3 = \{(\dots, R_{45} = pos, \dots), (\dots, R_{67} = pos, \dots)\}.$$



How many elementary events are in the set of all possible outcomes of our 1000-researcher experiment? For every researcher, there are two possible outcomes: positive or negative result. Hence, we have $2 \cdot 2 \cdot 2 \cdots = 2^{1000}$ elementary events in our set of all possible outcomes. This is a very large number ($\sim 10^{300}$) - more than there are particles in the universe ($\sim 10^{80}$).

We call the set of all elementary events Ω (the Greek letter Omega):

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_{2^{1000}}\}.$$

Note, that **we collect elementary events to form events** like we just did for event A_3.

Note, that the 2^{1000} **elementary events in the 1000 researcher experiment are also disjoint**. Why? For every elementary event, certain researchers found something and others did not. The combinations are all different from each other. Hence, all the elementary events cannot happen at the same time within that one experiment. All of them are disjoint.

The probability of the event " " (nothing occurred) should be zero ($\mathbb{P}(\emptyset) = 0$), were " " denotes the event that no researcher gets a positive or negative result ($= \emptyset$, the so-called empty set). This is impossible due to the design of the experiment. We would therefore define this probability as zero and (if we can count the number of different outcomes) this event can indeed *never* happen.

Obviously, the probability of an event should at a mininum be zero and at a maximum be one:

$$0 \leq \mathbb{P}(A) \leq 1$$

.

2.2.2 Axioms of probability theory

We can summarize these informally stated properties more formally (Kolmogorov's axioms):

1. $\mathbb{P}(\emptyset) = 0$: Probability of the "impossible" event should be zero. (2.1)
2. $\mathbb{P}(\Omega) = 1$: Probability, that any outcome occurs in our random experiment. (2.2)
3. If $A_1, A_2, \ldots$ pairwise disjoint: $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ (2.3)

The ∞-symbol in **axiom 3** comes into play if we are dealing with (potentially) infinitely many events. For instance, we could ask for the number of researchers we need to look at until we see the first positive result (geometric distribution). We could find the first positive result in the first researcher, or the second, etc. There is no upper limit.

As concrete example for law 3 in our example, we can put the following:

$$\mathbb{P}\big(\text{"(only) researchers 34, 56 and 777 get a pos. result" or "(only) researchers 1 and 5 get a pos. results"}\big) = \tag{2.4}$$

$$\mathbb{P}\big(\text{"(only) researchers 34, 56 and 777 get a pos. result"}\big) + \mathbb{P}\big(\text{"(only) researchers 1 and 5 get a pos. results"}\big) \tag{2.5}$$

Since the researchers are working independently from each other, we can simply multiply the probabilities of their individual positive or negative results in our larger 1000-researcher experiment. For example, for the first probability there are exactly 3 positive results (=effect found) and 997 negative results (=no effect found). This can be calculated as: $0.04 \cdot 0.04 \cdot 0.04 \cdot \underbrace{0.96 \cdots 0.96}_{\text{997 times}} = 0.04^3 \cdot 0.96^{997}$, which yields a very small number ($1.350826 \cdot 10^{-22}$) since we are fixating on specific researchers to find the effect. If we relax the question to the *number of researchers* that find an effect, we get much larger numbers. We say, the number X of positive results under H_0 (there is no true effect)) for a positive effect is **binomially distributed**: $X \sim Bin(n = 1000, p = 0.04)$. The YouTube-channel 3Blue1Brown is highly recommended in general. You should watch this video on the binomial distribution to get a clearer picture. This video from KhanAcademy could also be helpful. In our example, the probability that exactly 3 researchers find an effect is $\binom{1000}{3} \cdot 0.04^3 \cdot 0.96^{997} = 2.244627 \cdot 10^{-14}$. Still small, but much higher than before. Of course, the commands in R can be found easily via Google or your favourite large language model (LLM): “Give me the commands for the binomial distribution in R and a nice example too”. Note that the sum of all elementary events (all possible outcomes) indeed adds up to 1 in our 1000-researcher-experiment: $\sum_{i=0}^{1000} \binom{1000}{i} 0.04^i 0.96^{1000-i} = 1$

```r
sum(dbinom(0:1000, prob = 0.04, size = 1000))
```

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

As we will see later, **axiom 1** above does not mean, that the event can never occur. For every continuous random variable (e.g. with a normal or a uniform distribution), the probability of a single point is zero. This video could help.

Axiom 2 is always true. Some result *has* to occur in our random experiment. What is Ω again? In our countable case of researchers, $\Omega = \{\omega_1, \omega_2, \cdots, \omega_{2^{1000}}\}$ would be the set of all possible outcomes if we let 1000 researchers conduct the experiment. Each researcher can either find an effect or not. Hence, we have 2^{1000} possible outcomes of our 1000-researcher experiment. This is a *very* large number. Adding up all these probabilities would sum to 1 according to axiom 3. Combining different elementary events ω from the whole collection of possible outcomes Ω gives us "events" like the ones we used above (A_1, A_2, A_3).

Note that there is a **difference between the elementary experiment of the individual researcher (finding an effect or not) and the whole experiment** of 1000 researchers we are looking at (simulatenously). Do not make the mistake to add the single probabilities of finding an effect (under H_0) of 0.04 to get the probability of finding an effect in the whole experiment: This would result in: $1000 \cdot 0.04 = 40 > 1$, which is hardly a probability anymore.

This leads us to the concept of independence of events.

2.2.3 Independence of events

Two events A and B are independent if the occurrence of one event does not affect the occurrence of the other event. In plain English, the probability of event A happening is the same whether event B happens or not. Mathematically, we can write this as:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

or equivalently:

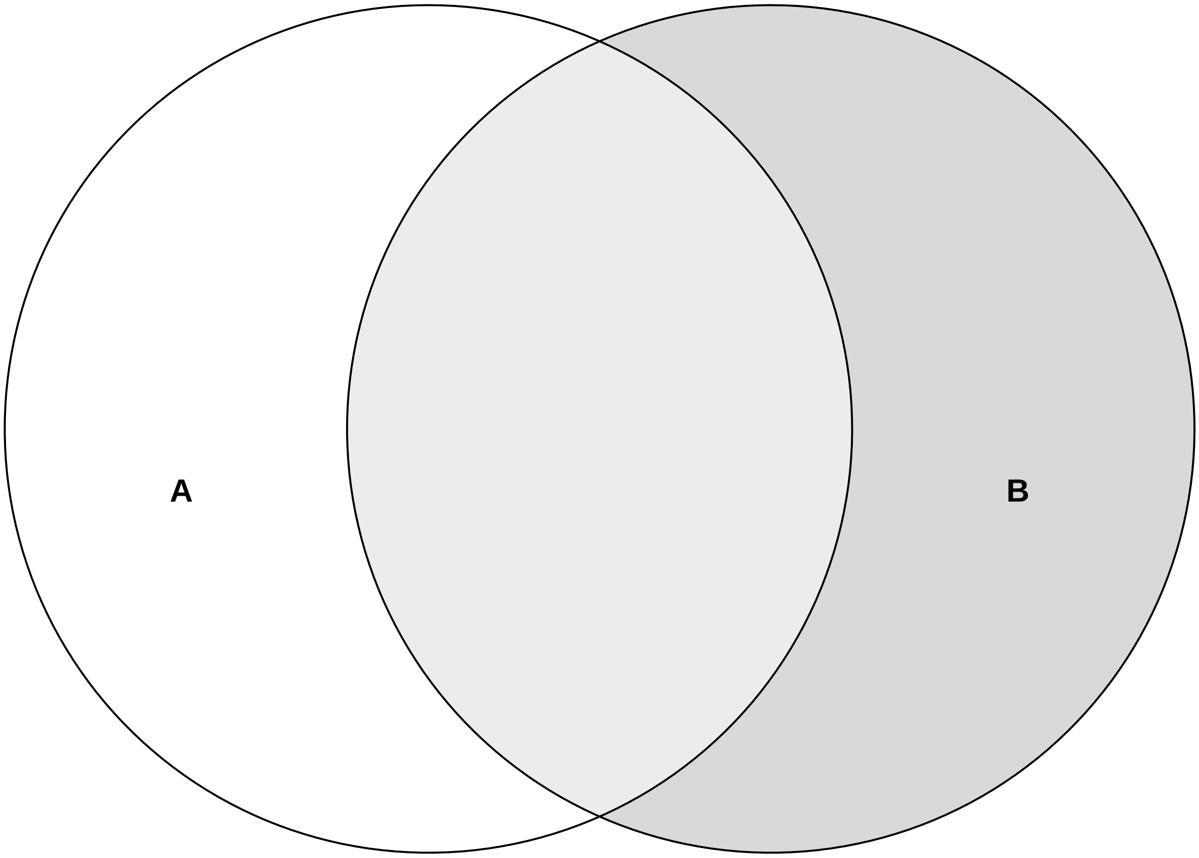
$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

A simple **example in our context**: The probability of researcher 45 finding an effect (event A) is the same whether researcher 67 finds an effect (event B) or not since they are not communicating with each other.

Here, we used the very important concept of **conditional probability**. The probability of event A *given* that event B has occurred (not necessarily chronologically different!) is denoted as

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

This video explains it well.



For the probability of event A, we a now only interested in the light-grey area with respect to the whole area of event B since event B is our reference frame now (as opposed to the whole space Ω before). Note that even if the probability of event A changes when B has happend, B could still have no *causal* effect on A. They could have a common cause, for instance.

Example in our context: Let's assume researchers 45 and 67 would not be independent. We would for instance find that the probability of 45 is higher than 4% if we knew that 67 found the effect. This does not necessarily mean that researcher 67 causes researcher 45 to find an effect. It could might as well be that their statistical training was very similar and they both made the same mistake in their analysis.

2.2.4 Difference between independence and disjointness

There are four possible scenarios when considering two events:

Example 1: disjoint but not independent

- Event A: Patient receives treatment A.
- Event B: Patient receives treatment B (or is in the control group).

These two events, A and B, are disjoint because a patient cannot receive both treatments at the same time. If a patient receives treatment A, they cannot receive treatment B (and vice versa), meaning the events cannot occur together in this setting. Thus, $P(A \cap B) = 0$. However, these events are not independent, because the probability of receiving one treatment depends on not receiving the other. In this setup, if the patient received treatment A, the probability of receiving treatment B is zero: $\mathbb{P}(B|A) = 0$. The probability of the patient receiving therapy B could be 50% (if randomized): $\mathbb{P}(B) = 0.5$. Hence, they are dependent.

Example 2: independent but not disjoint

- Event A: The patient shows a treatment effect during a study.
- Event B: The patient wins the lottery during the study.

These two events are independent because the probability of a patient showing a treatment effect is not influenced by whether they win the lottery or not (at least if we assume that lottery participants do not have different properties compared to non-lottery particiants that are conducive to showing a treatment effect). Also, the probability of winning the lottery is not influenced by whether the patient shows a treatment effect or not. We would probably see a surge in volunteers in our studies. The events are unrelated: one depends on the treatment, while the other is purely a matter of luck. However, these events are not disjoint because both can happen at the same time. A patient could experience the treatment effect and also win the lottery during the study. Thus, $P(A \cap B) \neq 0$, meaning both events can occur together.

Example 3: neither independent nor disjoint

- Event A: The patient shows a treatment effect during a study.
- Event B: The patient is a heavily motivated and self-sufficient.

These two events are neither independent nor disjoint. The patient's motivation could influence the treatment effect (if for instance home exercises are part of the therapy), making the events dependent. However, the patient's motivation is not mutually exclusive with the treatment effect: The patient can be heavily motivated and show a treatment effect at the same time. Hence, the events are not disjoint either. They can occur together.

Example 4: independent and disjoint

See Exercise 4.

2.2.5 Answers to questions about the 1000 researcher-experiment (among many others):

Maybe, we can already answer some of the questions from above using what we have learned so far.

For the first question we would probably bet on the maximum probability of the binomial distribution. The number of positive experiments out of 1000 has to be between 0 and 1000. Each one of them has 0.04 probability of happening. With R, we quickly calculate the maximum probability:

```r
# Calculate the maximum probability using binomial distribution
n <- 1000   # number of researchers
p <- 0.04   # probability of a treatment effect for each researcher

# Calculate the probabilities for each possible number of positive results
probs <- dbinom(0:n, size = n, prob = p)

# Find the number of experiments with the highest probability
# index of the maximum probability starting with 1
max_prob_number <- which.max(probs)

# Show the result
max_prob_number - 1 # since we started with 0
```

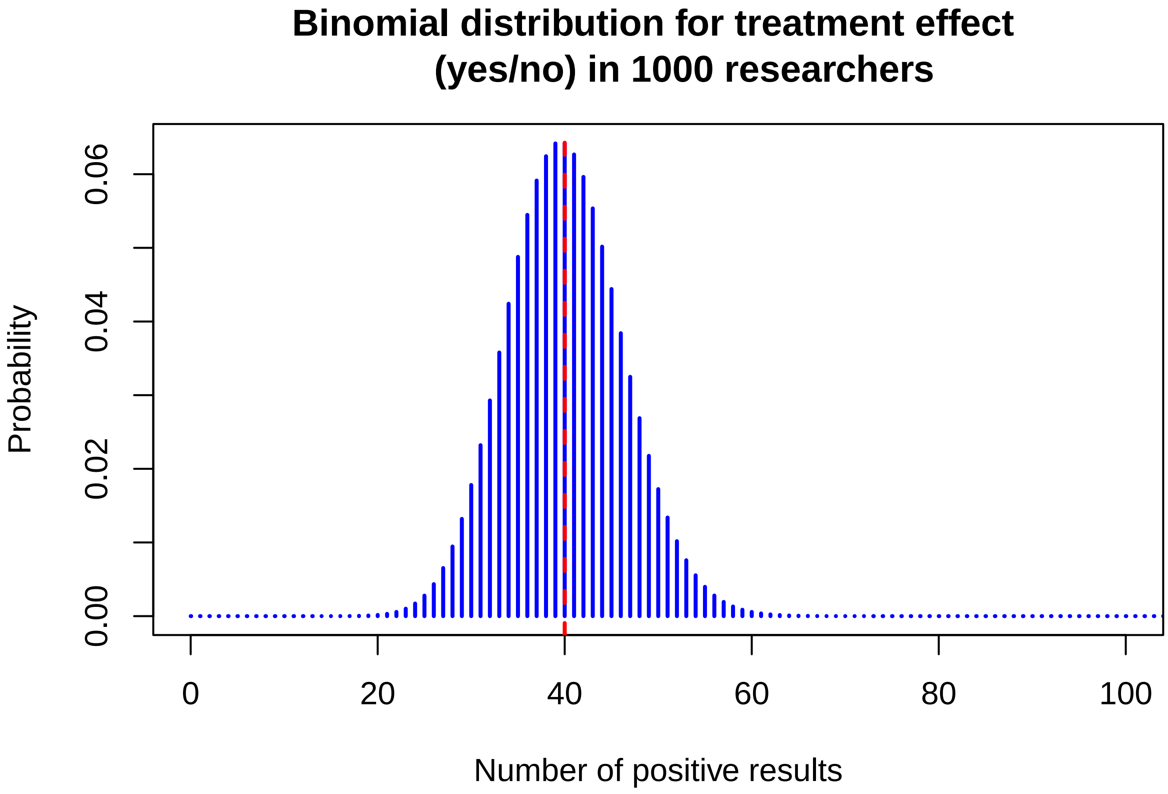
```
## [1] 40
```

```
# 40 is the most likely number of positive results
dbinom(39:41, size = 1000, prob = 0.04)
```

```
## [1] 0.06417798 0.06424483 0.06267788
```

Now, let's visualize the binomial distribution for this case using base R syntax:

```
# Plot the binomial distribution
plot(0:n, probs, type = "h", lwd = 2, col = "blue",
     xlab = "Number of positive results",
     ylab = "Probability",
     xlim = c(0, 100),
     main = "Binomial distribution for treatment effect
     (yes/no) in 1000 researchers")
abline(v = max_prob_number - 1, col = "red", lwd = 2, lty = 2)
```



Note, this form of distribution looks like a bell curve, aka a normal distribution, probably the most important distribution in statistics. One can show formally that the binomial distribution converges to the normal distribution under certain conditions. So, if we would only have one shot to predict the number of researches reporting a treatment effect under the assumption that no treatment exists, we would bet on 40. This guess would also not be too bad considering the whole range (0 to 1000) since we can expect the number of successes above, let's say, 65 and below, let's say, 15 to be very unlikely.

```
sum(dbinom(0:14, size = 1000, prob = 0.04)) # prob of 14 or less
```

```
## [1] 1.384829e-06
```

```
sum(dbinom(66:1000, size = 1000, prob = 0.04)) # prob of 66 or more
```

```
## [1] 7.160623e-05
```

```
sum(dbinom(15:65, size = 1000, prob = 0.04)) # prob of 15 to 65
```

```
## [1] 0.999927
```

The second question asked about observing 137 positively reporting researchers. We can calculate the probability of observing 137 or more positive results using the binomial distribution (plug into the formula): $\mathbb{P}(\text{observing 137 or more}) = \sum_{i=137}^{1000} \binom{1000}{i} 0.04^i (1 - 0.04)^{1000-i}$:

```
# Calculate the probability of observing 137 or more positive results
# (using the complement rule)
1 - sum(dbinom(0:136, size = 1000, prob = 0.04))
```

```
## [1] 5.551115e-16
```

```
# Compare to winning the Swiss lottery
(1 / 31474716) / (1 - sum(dbinom(0:136, size = 1000, prob = 0.04)))
```

```
## [1] 57234507
```

57 million times less likely than winning the Swiss lottery. If this event would happen, we would probably reject the assumption that the therapy is not working at all.

In the calculation above, we used the complement rule to calculate the probability of observing 137 or more positive results: $1 - \mathbb{P}(\text{observing 136 or less})$.

In general, for an event A:

$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A),$$

where A^C comprises all elementary events that are not in A. In our case, the compliment of observing 136 or less is observing 137 or more and vice versa: $\mathbb{P}(0, \dots, 136) = 1 - \mathbb{P}((0, \dots, 136)^C) = 1 - \mathbb{P}(137, \dots, 1000)$.

The third question asked about the expected number of positive results if the therapy is working in 12% of the patients. As you can probably guess by now: We would guess $1000 \times 0.12 = 120$ positive results. This is the so-called **expected value** $\mathbb{E}(X)$ of the binomial distribution. It is not always the maximum probability (the so-called mode) of the distribution though: Consider a binomial distribution $\text{Bin}(10, 0.77)$:

- The **mean** is $\mathbb{E}(X) = 10 \times 0.77 = 7.7$. This number is not an integer and we can therefore not calculate the density at this point.
- The **mode** is 8.

The fourth question asked about the problem of writing a press release stating that 47 out of 50 studies showed a treatment effect. Well, this would be scientific fraud and a case of survivorship bias. You only look at the studies that showed a treatment effect and ignore the ones that did not or you restrict the number of studies to a certain number lower than the true number. This is also relevant in finance. You may want to read this excellent article by John Ioannidis for a humbling big-picture of how relevant published results can be.

The fifth question asked about the problem of multiple testing. If you test many hypotheses, you will find some "significant" results by chance alone. One could also call the practice of testing many hypotheses to find "significant" ones p-hacking. This should be absolutely avoided. Unfornately, it is still common practice in many fields. Often it happens unconsciously. Example: If you test 100 hypotheses simultaneously at a significance level of 4%, you would expect 4 "significant" results by chance alone. If you report those 4 results as legitimate finding, you are p-hacking. When reading a scientific article, watch out for large amounts of p-values and their (over-)interpretation as "significant" (relevant) or "non-significant" (not relevant). This article is recommendable to get away from a too strict dichotomous interpretation of research results.

The sixth question asked about the variation of positive results between large research departments. This demands the very important concept of variance: The expected quadratic deviation from the mean: $\mathbb{V}ar(X) = \mathbb{E}\{(\mathbb{E}(X) - X)^2\}$. In simple terms: How much does the number of positive results vary around the mean of 40 on average? See also Exercise 5. Maybe this video helps as well.

2.2.6 Addition of probabilities

Above in axiom 3, we stated that the probability of the union of pairwise disjoint events is the sum of the probabilities of the individual events. What if the events are not disjoint? For simplicity, let's consider only 2 researchers (doing 2 parallel experiments) and define event A_1 as "researcher 1 finds an effect" and A_2 as "researcher 2 finds an effect". What is the probability that at least one of the researchers finds an effect? Our event space $\Omega = \{(R1pos, R2pos), (R1pos, R2neg), (R1neg, R2pos), (R1neg, R2neg)\}$.

$\sum_{\omega_i} \mathbb{P}(\omega_i) = 0.04^2 + 0.04 \times 0.96 + 0.96 \times 0.04 + 0.96^2 = 1$

$A_1 \cup A_2 = \{(R1pos, R2pos), (R1pos, R2neg), (R1neg, R2pos)\}$

$A_1 = \{(R1pos, R2pos), (R1pos, R2neg)\}$

$A_2 = \{(R1pos, R2pos), (R1neg, R2pos)\}$

$\mathbb{P}(A_1) = 0.04^2 + 0.04 \times 0.96$ (First researcher finds an effect or both find an effect)

$\mathbb{P}(A_2) = 0.04^2 + 0.96 \times 0.04$ (Second researcher finds an effect or both find an effect)

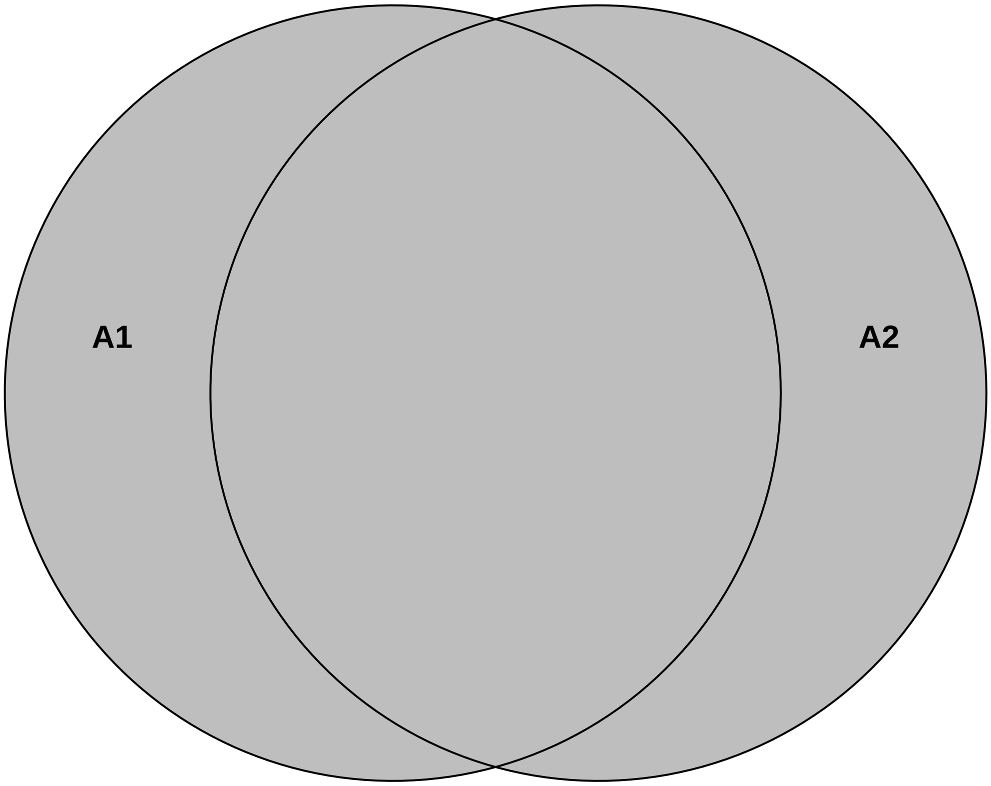
In, general, we can write the probability of the union of two events as: $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$

Put in the values: $0.04^2 + 0.04 \times 0.96 +$ $0.04^2 + 0.96 \times 0.04 - 0.04^2 = 0.04^2 + 0.04 \times 0.96 + 0.96 \times 0.04. = 0.0784.$

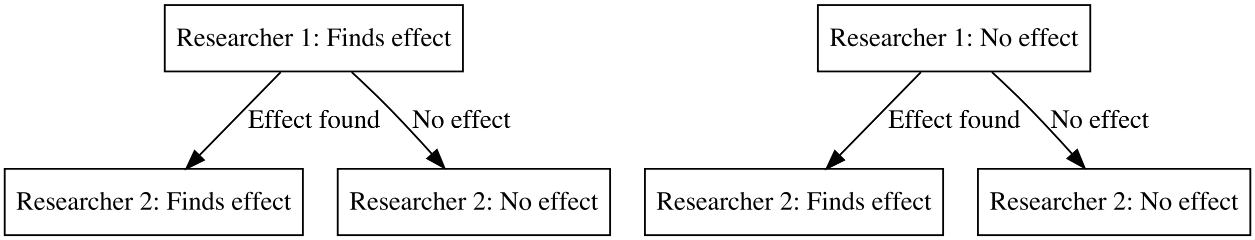
Or simpler with the **complement rule**:

$\mathbb{P}(A_1 \cup A_2) = 1 - \mathbb{P}(\text{neither } A_1 \text{ nor } A_2) = 1 - 0.96^2 = 0.0784.$

See also Exercise 6.



Here is another helpful depiction of the situation:

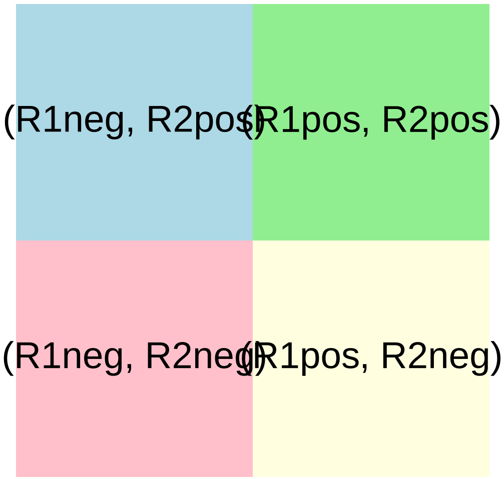


So, the probability of at least one researcher finding an effect is the sum of the probabilities of the individual researchers finding an effect minus the probability of both finding an effect, which is the same as that both or exactly one of them finds an effect.

We can also visualize the 4 disjoint elementary events

$$\Omega = \{(R1pos, R2pos), (R1pos, R2neg), (R1neg, R2pos), (R1neg, R2neg)\}$$

in a Venn diagram. The probabilites of these 4 events in the event space Ω must add up to 1 since they are disjoint and one of them has to happen. There is no “room” left.



2.2.7 Probabilities for health science

We have learned a lot so far: The axioms of probability theory, the difference between independence and disjointness, and the addition of probabilities.

How does probability theory fit into the big picture of statistics for health sciences?

In many health-related studies, we want to perform one or more of the following tasks:

- Estimate proportions (e.g., the proportion of patients with lower back pain. How big is the problem from a public health perspective?),
- Test hypotheses (e.g., whether a new therapy is superior to the standard therapy. How sure can we be that the new therapy is better? What is the probability that the treatment effect is between x and y points on some scale?),
- Estimate therapy effects (e.g., the effect of a new therapy on pain reduction: How many points does the pain decrease? How is the pain reduction distributed? Are there outliers and why? Are there participants that to not benefit from the therapy?)

In all such cases, probability theory is the established tool to answer questions that are afflicted with uncertainty. Would there be no variation in results/effects, we would probably argue differently. In our world, probability theory is the tool to **quantify uncertainty**.

We can always ask ourselves: Where is this entity (proportion, effect, etc.) with which frequency/probability?

2.2.8 Discrete vs. continuous probability distributions

As one of the most prominent examples of a discrete distribution, we have already seen the binomial distribution in our 1000-researcher-experiment. A special case of it is the Bernoulli distribution, where you only throw the coin once or let one researcher conduct the experiment.

As an example of a continuous distribution we have mentioned the normal distribution above. It is the most important distribution in statistics for reasons that become increasingly clear as we go along. One of them is the central limit theorem which we have already mentioned in the introduction slides. Feel free to watch this video. The theorem states that, under appropriate conditions, the distribution of a normalized version of the sample mean ($\bar{X} = \sum_{i=1}^{n} X_i$) converges to a standard normal distribution. By this theorem, we can link **any** distribution to the normal distribution.

Discrete or continuous, the **goal** is the same: We want to now **where** the realization of my **random variable** lands **with what probability** when I do the experiment? How often will I get heads?

- How often will the researcher find an effect?

- With what probability will I get a pain-score reduction of at least 1 point in this patient in front of me given his/her characteristics and history?

- When looking at ZHAW students, female, soccer lovers, what kind of hourly intense sports activity can I expect and does that differ to other groups?

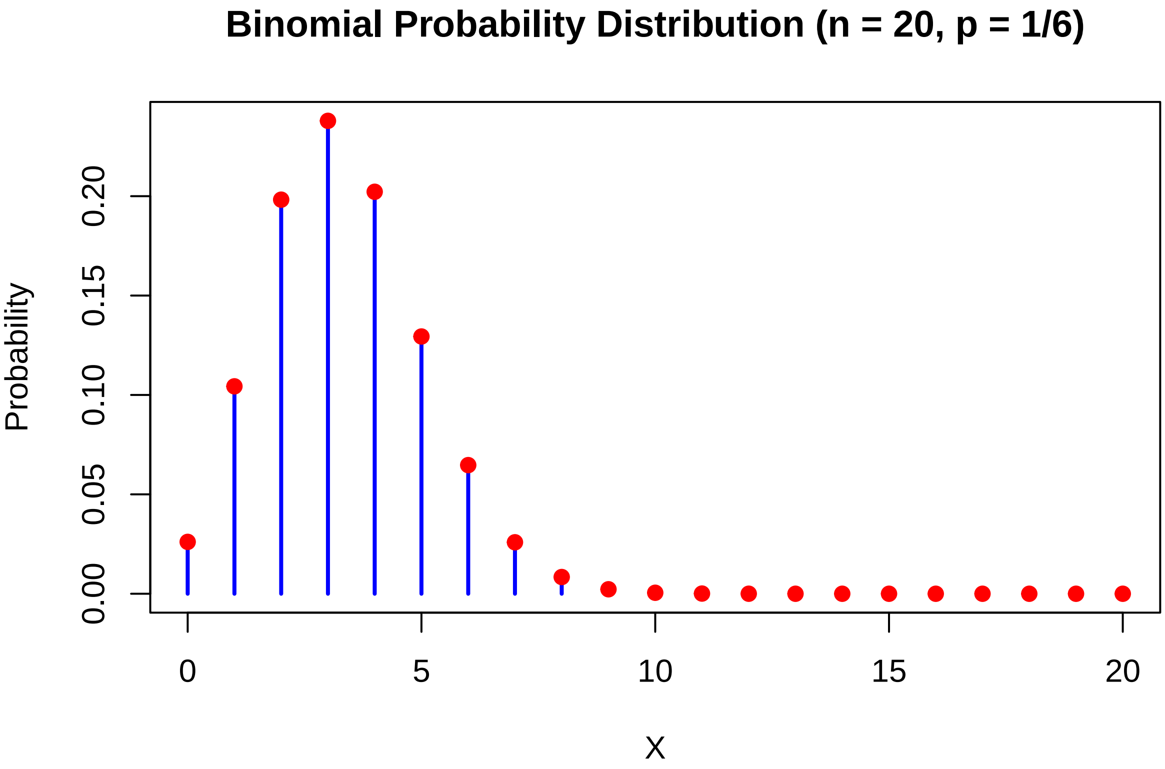
For us, the following definitions should suffice.

2.2.8.1 Discrete probability distrubtions are used when we can count the outcomes, which includes infinitely many.

Some examples of discrete probability distributions are: - **Bernoulli distribution**: A single trial with two outcomes (e.g., find an effect or do not find an effect). - **Binomial distribution**: The number of successes in a fixed number of trials (e.g., the number of false effects found among 1000 researchers). - **Poisson distribution**: The number of events in a fixed interval of time or space. - **Geometric distribution**: The number of trials until the first success. This number has no upper limit.

We always assign probabilites to the countable outcomes of these distributions, like in the example of the binomial distribution when we throw the dice 20 times and are interested in the number of 3s:

```r
# Define parameters for the binomial distribution
x_values <- 0:20
probabilities <- dbinom(x_values, size = 20, prob = 1 / 6)

# Plot the binomial distribution with styling
plot(x_values, probabilities, type = "h", lwd = 2, col = "blue",
     xlab = "X", ylab = "Probability",
     main = "Binomial Probability Distribution (n = 20, p = 1/6)")

# Add points for clarity
points(x_values, probabilities, pch = 19, col = "red")
```



Each outcome has a probability > 0 assigned to it. The sum of all probabilities is 1: $\sum_{i\in\text{Possible outcomes}} \mathbb{P}(X = i) = 1$. For every event, we just add the probabilities of the elementary outcomes that are in the event:

$\mathbb{P}(X \in (3, 8, 9, 14)) = \mathbb{P}(X = 3) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 14)$. This principle is true for all discrete probability distributions. Rather simple and elegant:

$$\sum_i \mathbb{P}(X = x_i) = 1,$$

where X ist the random variable (which takes values x_i when the random experiment is conducted) and x_i are the possible outcomes of X.

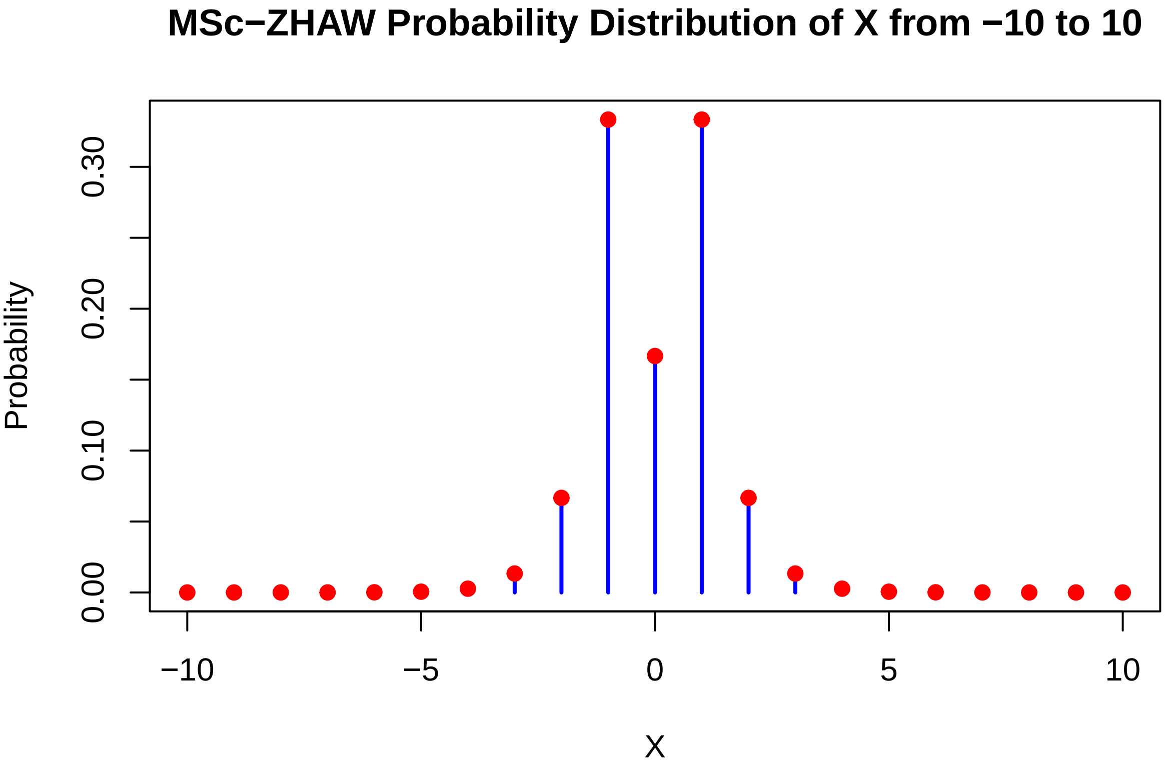
We could **invent our own discrete probability distribution** instantly (see also Exercise 8), we'll call it the MSc-ZHAW-distribution:

Let $X \in \mathbb{Z}$. Every whole number gets the following probability: $\mathbb{P}(X = 0) = 0.1$ and for $x_i \neq 0$: $\mathbb{P}(X = x_i) = 0.2^{|x_i|}$. The sum of all probabilities is: $\sum_{x_i\in\mathbb{Z}} \mathbb{P}(X = x_i) = \mathbb{P}(X = 0) + 2 \cdot \sum_{i\in\mathbb{N}} 0.2^i = 0.1 + 2 \cdot \frac{0.2}{1-0.2} = 0.6$. Hence, we need to divide every probability by 0.6 to get in sum 1. The final definition is then:

$\mathbb{P}(X = 0) = \frac{1}{6}$ and for $x_i \neq 0$: $\mathbb{P}(X = x_i) = \frac{5}{3}0.2^{|x_i|}$.

```
# Define the probability function
P <- function(X) {
  if (X == 0) {
    return(1 / 6)
  } else {
    return((5 / 3) * (0.2^abs(X)))
  }
}

# Create a sequence of X values from -10 to 10
```

```
X_values <- -10:10

# Compute the probabilities for each X value
probabilities <- sapply(X_values, P)

# Plot the probabilities
plot(X_values, probabilities, type = "h", lwd = 2, col = "blue",
     xlab = "X", ylab = "Probability",
     main = "MSc-ZHAW Probability Distribution of X from -10 to 10")

# Add points for clarity
points(X_values, probabilities, pch = 19, col = "red")
```



```
# Check if it sums to 1 (approximately):
x_values <- -1000:1000
sum(sapply(x_values, P))
```

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

Deviations from zero (±1) are highly likely with this distribution. The probability of $X = 0$ is also rather high with $\frac{1}{6}$. Larger deviations from zero are less likely and go exponentially towards zero (very fast). So we would expect almost never to see values outside of ±10. This does of course not mean that we will never see them. Do the experiment often enough and you will seen them with probability 1 (see Exercise 10).

```
x_values <- setdiff(-1000:1000, -10:10) # exclude values from -10 to 10
sum(sapply(x_values, P))
```

```
## [1] 8.533333e-08
```

Expectation $\mathbb{E}(X)$ of a discrete random variable: The expectation of a discrete random variable X is defined as:

$$\mu = \mathbb{E}(X) = \sum x_i \cdot \mathbb{P}(X = x_i),$$

a weighted sum of possible values x_i with their respsective probabilities $\mathbb{P}(X = x_i)$.

The term "expectation" is probably somewhat misleading. It is not necessarily the value we "expect to see" when we do the experiment. For instance, the expected value of a Bernoulli distribution is: $\mu = \mathbb{E}(X) = 0 \cdot (1-p) + 1 \cdot p = p$, which could be 0.5. The individual outcomes are 0 and 1, and not 0.5. But 0.5 would be the mean of many experiments.

The expectation can be interpreted as the center of mass of the distribution. It is the value that the distribution "balances" around.

Maybe this video helps too.

The cool thing is that we can learn the true (but unknown) expectation of a distribution by the sample mean. The more samples we collect, the closer we will be. This is (roughly) the statement of the law of large numbers:

$$\bar{X}_n \to \mu = \mathbb{E}(X) \quad \text{as} \quad n \to \infty.$$

See here for an animated example of this law.

Remember: The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a (really good) estimator for the expectation $\mu = \mathbb{E}(X)$ of a distribution. This is true for discrete and continuous distributions.

The **variance of a discrete random variable** is defined as:

$$\mathbb{V}ar(X) = \mathbb{E}\{(\mathbb{E}(X) - X)^2\} = \sum_i (\mathbb{E}(X) - x_i)^2 \mathbb{P}(X = x_i),$$

the expected squared deviation from the mean. It is a measure of how much the values of the random variable differ from the mean.

Remember: The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an estimator for the variance $\mathbb{V}ar(X)$ of a distribution. This is true for discrete and continuous distributions.

A more natural interpretation of variability is the **standard deviation**:

$$\sigma = \sqrt{\mathbb{V}ar(X)},$$

since it's on the same scale as X.

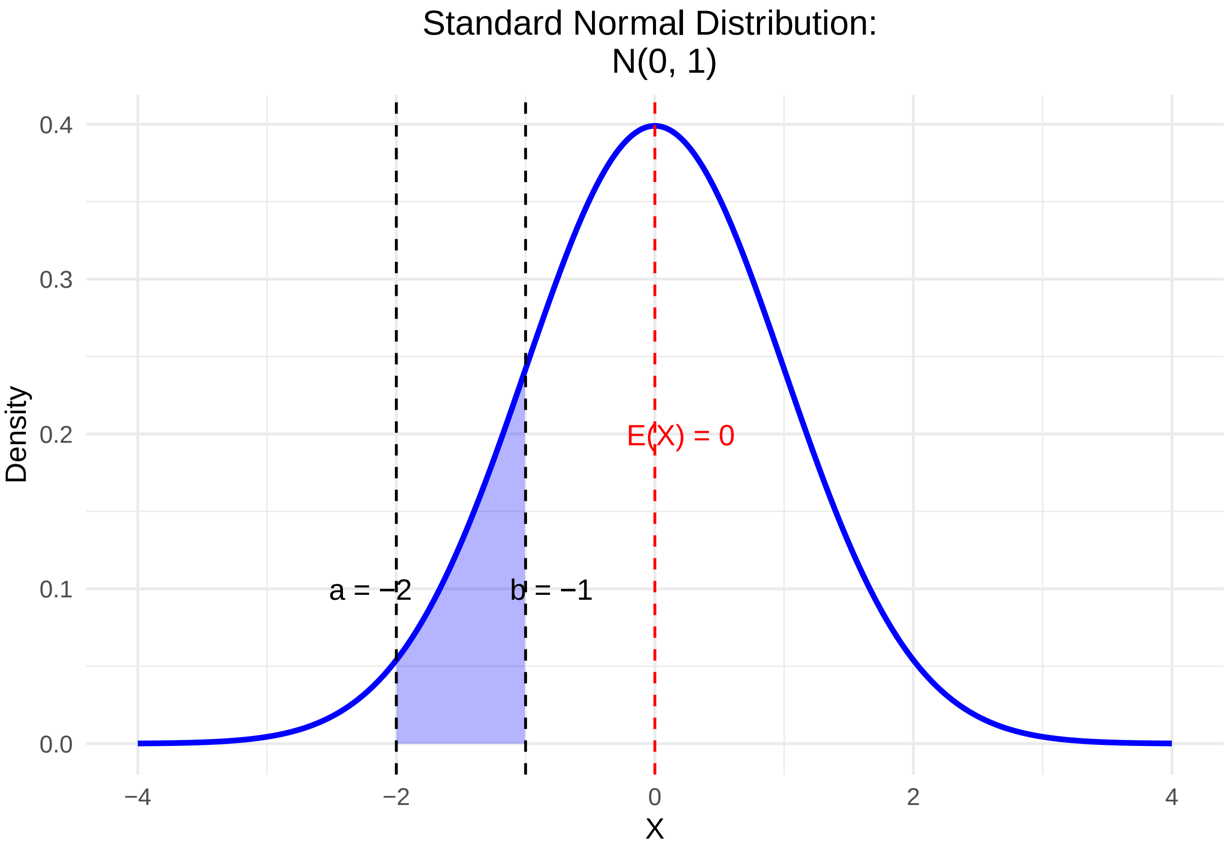
2.2.8.2 Continuous probability distributions are used when we cannot count the outcomes.

The most famous continuous probability distribution is the normal distribution. This video about probability distributions in general might be helpful.

```r
# Load necessary library
if (!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```r
pacman::p_load(ggplot2) # Installs and loads the package at the same time

# Define parameters for the normal distribution
mu <- 0    # Mean
sigma <- 1 # Standard deviation

# Define the limits for the area to be shaded
a <- -2  # Lower bound
b <- -1   # Upper bound

# Create a sequence of x values to evaluate the PDF
x_vals <- seq(mu - 4 * sigma, mu + 4 * sigma, length.out = 1000)

# Compute the corresponding density values using dnorm
y_vals <- dnorm(x_vals, mean = mu, sd = sigma)

# Create a data frame for plotting
df <- data.frame(x = x_vals, density = y_vals)

# Create a subset of the data for shading the area between a and b
df_shaded <- df[df$x >= a & df$x <= b, ]

# Plot the normal density and shade the area between a and b
ggplot(df, aes(x = x, y = density)) +
  geom_line(color = "blue", linewidth = 1) +  # Use linewidth instead of size
  geom_ribbon(data = df_shaded, aes(ymin = 0, ymax = density),
              fill = "blue", alpha = 0.3) +  # Shaded area
  ggtitle(paste("Standard Normal Distribution:
  N(", mu, ", ", sigma^2, ")", sep = "")) +
  xlab("X") +
  ylab("Density") +
  theme_minimal() +
  geom_vline(xintercept = mu, color = "red", linetype = "dashed") +
  annotate("text", x = mu + 0.2, y = max(y_vals) / 2,
           label = paste("E(X) =", mu), color = "red") +
  geom_vline(xintercept = a, color = "black", linetype = "dashed") +
  geom_vline(xintercept = b, color = "black", linetype = "dashed") +
  annotate("text", x = a - 0.2, y = max(y_vals) / 4,
           label = paste("a =", a), color = "black") +
  annotate("text", x = b + 0.2, y = max(y_vals) / 4,
           label = paste("b =", b), color = "black") +
  theme(plot.title = element_text(hjust = 0.5))
```



Here, like in any other "nice" continuous distribution, the area under the curve is 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

The probability of a single point is zero ($\mathbb{P}(\{x_i\}) = 0$). In any continuous distribution, we use the area under the curve to calculate probabilities. The probability of X being between a and b is the area under the curve (blue shade) between a and b: $\mathbb{P}(X \in (a, b))$. Note that the area over a single point would be zero and therefore the probability of a single point is zero.

The graph above is called a probability density function (PDF). Over every point, we express the probability by the height of the curve. See exercise 3 in the next chapter for a practical example for what we will use this in research.

Expectation $\mathbb{E}(X)$ of a continuous random variable: The expectation of a continuous random variable X is defined as:

$$\mu = \mathbb{E}(X) = \int x \cdot f(x)dx,$$

where x are the possible values of X and $f(x)$ is the probability density function of X.

Variance of a continuous random variable: The variance of a continuous random variable X is defined as:

$$\mathbb{V}ar(X) = \mathbb{E}\{(\mathbb{E}(X) - X)^2\} = \int (\mathbb{E}(X) - x)^2 f(x)dx.$$

A more natural interpretation of variability is the **standard deviation**:

$$\sigma = \sqrt{\mathbb{V}ar(X)},$$

since it's on the same scale as X.

Example: Normally distributed Ages of ZHAW students: $\mu = 24$, $\sigma = 3$. For the normal distribution, this means that approx. 68% of the students are between $(24 - 3 =)21$ and $(24 + 3 =)27$ years old.

2.2.9 Examples of prominent probability distributions used in health sciences

The first 2 are absolutely essential.

- The most important one is, as mentioned above, the normal distribution. It is often used to model the distribution of many variables in health sciences, e.g., blood pressure, weight, height, etc. Normality is also a common assumption in many statistical tests and models. This is the reason why you will find many statements like "we have checked normality using the shapiro-wilk test" (Which I would not recommend) in scientific articles. Normal distribution theory is very aesthetic and one is sometimes lead to believe that this is the normal state of nature, which is not the case. See also the history of the normal distribution. A common use of the normal distribution is in linear regression, where the errors and the conditional distribution of the modeled variable in the model are assumed to be normally distributed. We will deal with this in QM2.
- The binomial distribution ($X \sim B(n, p)$) is used to model the number of successes in a fixed number of trials. For example, the number of patients that respond to a therapy in a fixed number of patients. A special case of it is the Bernoulli distribution, which is used to model a single trial with two outcomes (throw the coin once; $X \sim B(1, p)$).
- Logistic Distribution. Underpins logistic regression models, which are used to predict binary outcomes (e.g., the presence or absence of a disease).
- Poisson distribution. Used to model the number of events in a fixed interval of time or space. For example, the number of patients arriving at an emergency department in a fixed time interval.
- Exponential distribution. For instance used in survival analysis to model the time until an event (e.g. refrigerator stops working) occurs.
- Student's t-distribution (small "t" please) generalizes the standard normal distribution. Like the latter, it is symmetric around zero and bell-shaped, but has fatter tails (compared to the normal distribution), i.e., "extreme" values are more likely. It is a very well known distribution underlying the t-test. See Exercise 12 for a practical example.

There are infinitely (!) many more distributions.

Our goal is to learn: How can we describe (the distribution of) what we see in our data? How can we make predictions? How can we make decisions based on our data? Probability theory and statistics are (for us) a very large tool box to answer these questions.

2.3 Exercises

2.3.1 Exercise 1 - Throwing a die very often

- Use your favourite large language model (LLM) to create an R-Script to simulate throwing a fair die 1000 times.
- Try to run the script. If it does not run, try to debug it using the LLM.

- Once, the script runs, let the LLM explain the code and outputs ("Please explain this script in detail…").
- Plot the frequency of each number (1-6) (after 1000 throws) and compare it to the theoretical probability of getting each number ($\frac{1}{6}$).
- Plot the relative frequency of 3s on the y-axis and the number of throws on the x-axis. This should give a converging pattern towards $y = \frac{1}{6}$.
- Which law of probability theory is illustrated by this simulation?

2.3.2 Exercise 2 - Bayes-teaser

Use Bayes' theorem to calculate the posterior probability of the therapy's effectiveness in the physiotherapy example above (Example). For simplicity, let's just test two θ-values: 0.3 (as in the previous study) and 0.4. We assign 50% in the prior knowledge that the parameter $\theta = 0.3$, and 50% to $\theta = 0.4$ since we trust our colleagues as well.

2.3.3 Exercise 3 - Find journals

Note: This is among the most important exercises of the course: Use Google or your favourite search engine to find scientific journals in *your* field (physiotherapy, midwifery, nursing, etc.). Look at the latest articles. We are interested in articles that used statistics (no qualitative studies).

- What was the research question? What where they trying to find out/confirm? Write down at least 10 research questions!
- Which statistical *methods* were used? Write down at least 10 methods!
- Was prior/external knowledge - before the actual model was estimated - used in any of the analysis?
- Where the results presented in a dichotomous way; meaning, was there a "significant"/"non-significant" result or not?

2.3.4 Exercise 4 - Independent and disjoint

Look at the definitions above and try to come up with examples for independent and disjoint events in your field of study.

- Is this possible?
- Why or why not?
- What would that imply?
- Draw a Venn diagram if possible!

2.3.5 Exercise 5 - Variance

- Simulate the number of positive results (found an effect even though there is none) in our 1000-researcher-experiment under the assumption that the therapy is not working at all ($p = 0.04$).
- Do this experiment in R 10,000 times and visualize the results in a histogram.
- How often do you get 65 or more positive results? How often do you get 15 or less positive results?
- Can you find the limits of a 90% interval around the mean (of 40) - using the so-called quantiles - for the number of positive results?
- What is the theoretical variance for our experiment?
- How can you estimate this theoretical (and in reality: unknown) variance from the 10,000 simulations?

2.3.6 Exercise 6 - Three researchers

Above in Addition of probabilites we went through in detail the case of 2 researchers finding an effect. Let's now consider 3 researchers simulatenously conducting the experiment.

- What does the event space Ω look like?
- Which elementary events are in the set of all possible outcomes of our 3-researcher experiment and how many are there?
- Draw the corresponding trinary (?) tree for this experiment.
- Which elementary events are in the following event: "Researcher 3 finds a positive effect"?
- Are the events "only researcher 1 finds an effect" and "only researcher 3 finds an effect" disjoint and/or independent?

2.3.7 Exercise 7 - Conditional probability

Let's consider again the 2 reasearcher situation from above (Addition of probabilites).

$\Omega = \{(R1pos, R2pos), (R1pos, R2neg), (R1neg, R2pos), (R1neg, R2neg)\}$.

- What is the probability that researcher 1 finds an effect given that researcher 2 found an effect?

2.3.8 Exercise 8 - Invent a discrete probability distribution

- Invent your own discrete probability distribution.
- What is the expected value of your distribution?
- What is the variance of your distribution?
- Think of something in the real world that could be modeled by your distribution.

2.3.9 Exercise 9 - Continuous probability distributions

- Invent your own continuous probability distribution.
- What is the expected value of your distribution?
- What is the variance of your distribution?
- Think of something in the real world that could be modeled by your distribution.

Hint: You can use simple shapes for the densitiy function defined by lines. And you can use simulation to answer questions about expected value and variance.

2.3.10 Exercise 10 - MSc-ZHAW-distribution

- Create sufficiently many random numbers (sample) from the MSc-ZHAW-distribution (see above) and see if you can produce values outside of ± 6.
- What is the mode of this distribution and how could we estimate it from the sample?
- What is the interquantile range of this distribution and how could we estimate it from the sample?

2.3.11 Exercise 11 - Independence and disjointness for dice events

Find examples of dice events when throwing a die once that are:

- Not independent and not disjoint.
- Not independent but disjoint.
- Independent but not disjoint.

2.3.12 Exercise 12 - Student's t-distribution

Let's look at a paper, where the t-distribution is used (in the background). The aim of the study was to assess the efficacy of pulmonary rehabilitation in addition to regular chest physiotherapy in non cystic fibrosis bronchiectasis. Table 1 describes the patient characteristics in both groups. Table 2 shows the primary endpoint (incremental shuttle walk test - ISWT) at baseline and follow-up time points. Figure 2 shows the outcomes at baseline, 8 weeks and 20 weeks for both groups.) They want to find out if the ISWT is different between the two groups. (Note, that an arbitrary threshold for the p-value of 0.05 is used to decide if the groups are "significantly" different. One should avoid these formulations. There is no reason not to use a different threshold (like 4.3%).) **The standardized difference of the group means is t-distributed**. This case is a bit more complex than the simple ones, since we have different sample sizes (15 vs. 12) and different variances in the groups. The statistics software will take care of this and use the so-called Welch's t-test.

- What do you think about the baseline values for ISWT in the two groups?
- What is the number in brackets next to the ISWT-values?
- According to the article, the data is normally distributed. Draw 3 normal distributions in one graph with the respective parameters for baseline, 8 weeks and 20 weeks for both groups. Make two graphs, one for each group.
- According to the text, Figure 2 shows the means ± standard errors ($SE = \frac{s}{\sqrt{n}}$) of the ISWT at baseline, 8 weeks and 20 weeks for both groups. Look at Figure 2, a. Does this match the description for instance at 8 weeks in the acappella+pulmonary group? Do the bars make sense?
- Now, let's simulate the differences at week 8 (ISWT) using the parameters given: Group sizes, 15 and 12, means (338.7 and 344.2) and standard deviations (42.2 and 115.5). Draw a histogram of the simulated differences. Calculate the 1.5% and 98.5% quantiles of the differences.

2.4 Solutions

Solutions for this chapter can be found here.

Chapter 3

Descriptive statistics

There are a myriad sources (books, websites, videos) explaining the concepts of descriptive statistics. We do not need to reiterate everything here. You can go through these sources to get started:

- R for non-programmers
- Science direct

The goal is to describe data in a meaningful and honest way. We **summarize** data to make them **more easily digestable** for us humans to answer questions like

- Where are the data points **located**? These questions are answered (at least attempted) by the location measures such as mean, median, and mode.
- How widely are they spread? How much do they **vary**? These questions are answered by the dispersion measures such as variance, standard deviation (root of the variance), interquantile range or just the range; or even Gini's mean difference.
- Are there any **outliers** (rare data points that are far away from the rest [Westfall, 2020, 405].) and why? This is a bit more complicated.

3.1 Example: Descriptive statistics in health sciences

These are birds-eye views on the data. Let's look at a paper which was recently published in the Journal of Physiotherapy in order to get a running start:

Patients with worse disability respond best to cognitive functional therapy for chronic low back pain: a pre-planned secondary analysis of a randomised trial

(This should be open access.)

The research question was "Do five baseline moderators identify patients with chronic low back pain who respond best to cognitive functional therapy (CFT) when compared with usual care?".

In **Table 2** of the paper, the authors present the baseline characteristics of the patients stratified by the treatment group (ususal care vs CFT). We find absolute numbers, percentages, means, and standard deviations for the continuous variables, medians, and interquartile ranges for the ordinal variables. This should give an idea of the sample. In the population paradigm of statistics, we draw a sample from the population of interest and try to make inferences about the population. We want to learn more about the population respectively that data generating process (DGP) producing the data [Westfall, 2020, 6-8]. How did the data come about? Note that this **sample varies everytime we draw from the population**. We can either imagine an infinitely large population or a finite one (e.g. population of Switzerland).

Often, variables in a study are (approximately) normally distributed. We can then efficiently summarize the variable with its mean and standard deviation (location and scale parameter) as is done in the paper for *Age* in years for instance. We do not want to present p-values in such an overview table since we merely *describe* data instead of making inferences about the population or the DGP. It is by no means given that *age* has to be normally distributed in our sample. We could easily have a sample with many young people and few elderly ones. This would result in a (positively) *skewed* distribution. Having many elderly people and few young ones would result in a negatively skewed distribution of course.

3.2 Univarate vs. bivariate statisics

One can distinguish between univariate and bivariate statistics. In univariate statistics, we look at one variable at a time, for instance *Age* in the example above, where we could draw a boxplot or a histogram. Table 2 in the paper is a good example of univariate statistics. We are not so much interested in the relationship between variables.

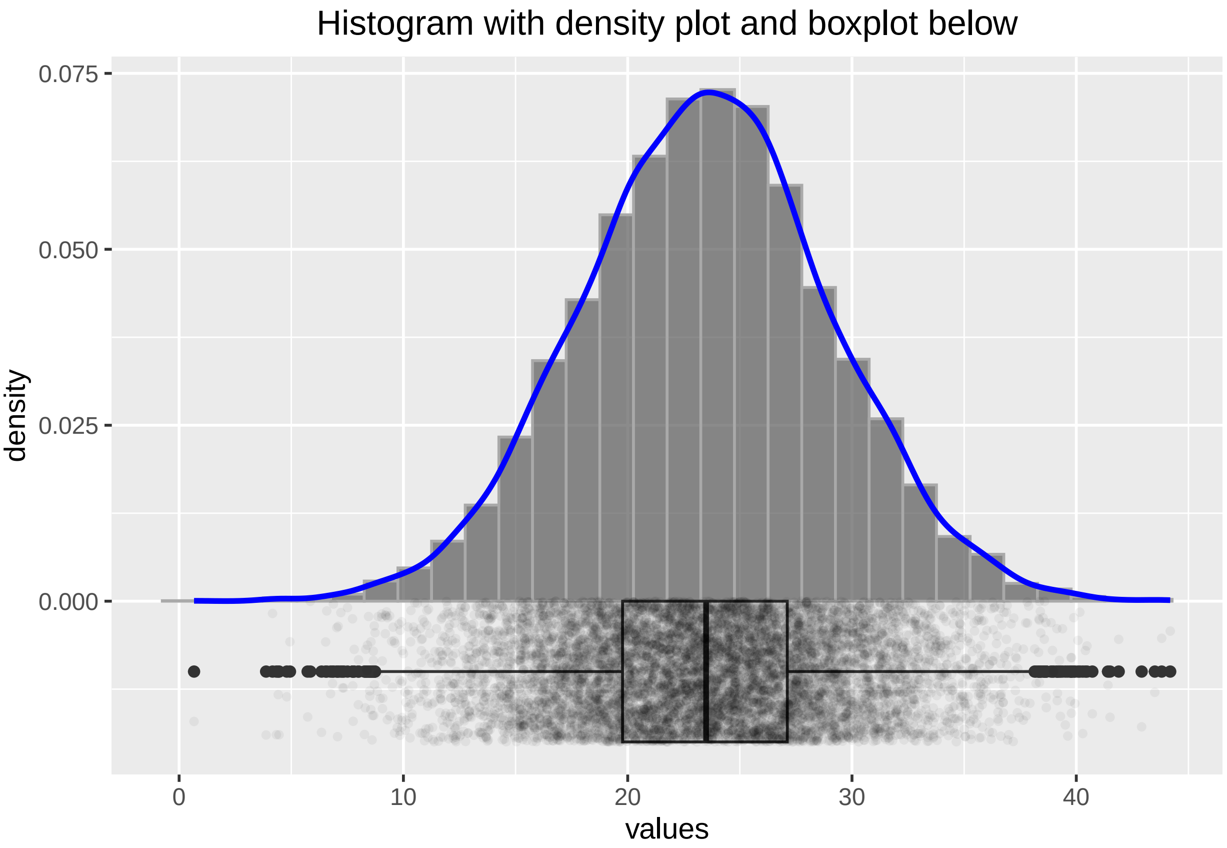
In bivariate statistics, we look at two variables *simultaneously*. An example could a scatter plot of *Age* and *Cognitive flexibility*, where we would possible find a falling relationship.

Multivariate statistics is the next step, where we look at more than two variables at the same time.

3.3 The histogram

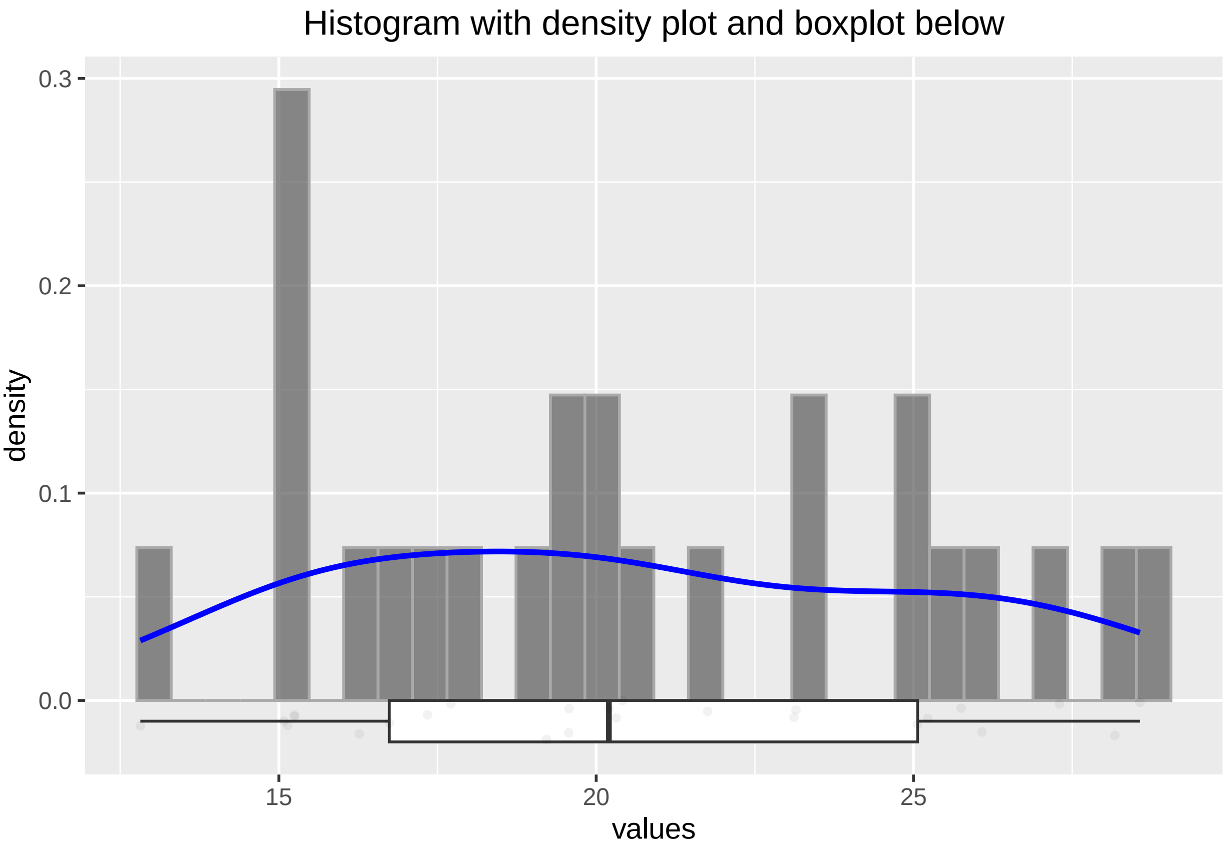
One way to visualize the distribution of a continuous variable is the histogram. This video might be helpful to cover the basics. I recommend plotting the histogram with a boxplot below; this helps to visualize the raw data points as well.

```r
library(pacman)
p_load(tidyverse)

set.seed(4433) # to get the same plot every time

# Generate normally distributed sample
x <- rnorm(10000, mean = 23.4, sd = 5.6)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values),
               width = 0.02, position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



Some researchers would discard the values below 10 and above 40 as outliers, but we know here that the data points are perfectly legitimate. One important thing we should be aware of in connection with small sample sizes ist variability. Let's create not 10000 samples of a normally distributed variable, but only 25 samples:

```
# Generate normally distributed sample
set.seed(1245) # to get the same plot every time
x <- rnorm(25, mean = 23.4, sd = 5.6)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values), width = 0.02,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



We *know* (in this case) that these values come from a normal distribution with a mean of 23.4 and a standard deviation of 5.6. Let's *estimate* the parameters (μ, σ) from the sample and use the often but not recommended Shapiro-wilk test for normality:

```
# Estimate mean and standard deviation
mean(x) # or
```

```
## [1] 20.6017
```

```
1 / length(x) * sum(x)
```

```
## [1] 20.6017
```

```
sd(x) # or
```

```
## [1] 4.622042
```

```
sqrt(1 / (length(x) - 1) * sum((x - mean(x))^2))
```

```
## [1] 4.622042
```

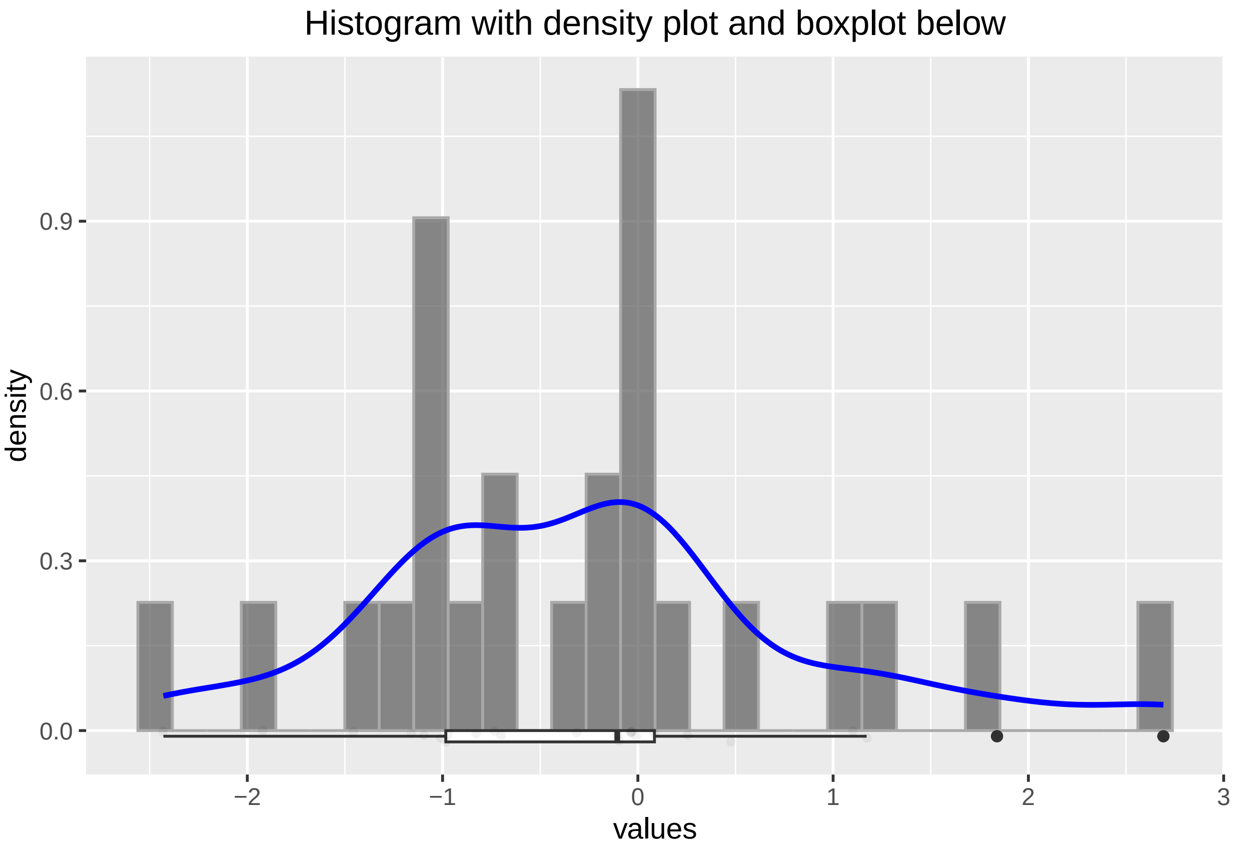
```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95059, p-value = 0.2585
```

Firstly, the histogram looks rather differently when using a different seed. Secondly, we would probability not be able to tell if the data stems from a normal distribution. Thirdly, the sample mean and standard deviation are not all that bad estimators for the true (but unknown) mean and standard deviation. Nicely enough, the Shapiro test would not reject the null hypothesis of normality in this case.

Let's try the Shapiro test with a t-distribution with 3 degrees of freedom, which is not normal:

```
set.seed(1245) # to get the same plot every time
x <- rt(25, df = 3) # random numbers from t-distribution
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values), width = 0.02,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95827, p-value = 0.3811
```

Some researchers would argue, that the distribution is normal (since the p-value is rather large), but we know it is not. We will be cautious with such ugly rule of thumbs.

3.3.1 Example in the wild

We will try to catch a histogram in the *wild* (= in a research paper): We find one here. Figure 1 shows a histogram often the foot posture index (FPI-6) scores from the participants (3217 healthy children aged 3 to 15). "The FPI score may range from −12 (highly supinated) to +12 (highly pronated)". There are a few things to note here:

- Acccording to the Methods-section: "Testing for normality using a Kolmogorov-Smirnov test, found non-normal distribution of all data…". Especially the BMI could be asymmetrically distributed, or also age. Unfortunately, we do not have histograms or boxplots for these variables. Nevertheless, the authors present the mean and standard deviation for these variables which implies (when not reading the methods section) that the data is (at least sufficiently) normally distributed.
- Strictly speaking, an FPI score cannot be normally distributed since it takes discrete values which are bounded between -12 and 12 (normal distribution can take values between $-\infty$ and ∞). But that

should not be a problem with so many levels of an ordered categorical variable. It should be a sufficient approximation.

- "The FPI was analysed as continuous data, rather than as z-score data". z-scores are standardized scores:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

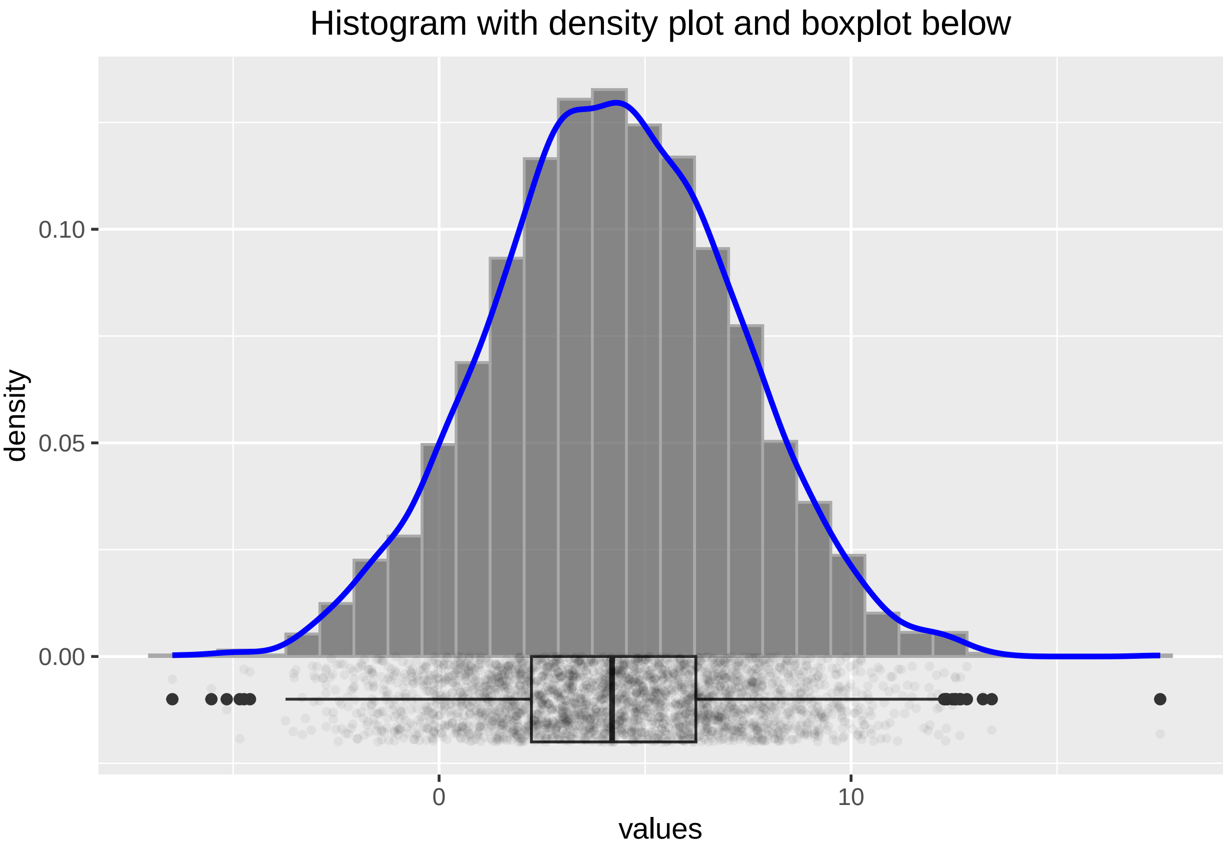
Doing this for the FPI scores would give (FPI right, see Table 1):

$$z_i = \frac{FPI_i - \overline{FPI}}{SD(FPI)} = \frac{FPI_i - 4.20}{3.00}$$

$\overline{FPI}$ ist the arithmetic mean of the FPI scores, $SD(FPI)$ the standard deviation of the FPI scores. μ_X is the expectation of X and σ_X the standard deviation of X. So we measure not in FPI units anymore, but in **standard deviations from the mean** which makes it easier to compare different variables with each other. A nice property is: If X is normally distributed ($X \sim N(\mu, \sigma)$), then the z-scores are standard normally distributed ($Z \sim N(0, 1)$). See Exercise 4. Of course *both*, z-scores as well as FPI scores are considered "continuous" (in this context).

- With so many obsvervations ($n = 3217$), how would a truly normal distribution with the parameters $FPI \sim N(4.20, 3.00)$ look like? We can simulate this with the following code:

```
set.seed(8345) # to get the same plot every time
x <- rnorm(3217, mean = 4.20, sd = 3.00)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values), width = 0.02,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



Comparing these histograms (using different seed-values), we could assume that for truly normally distributed FPI scores, the histogram might look smoother than the one in Figure 1 in the paper. Especially the values around 6 seem to deviate from the normal distribution.

3.4 Correlation

The correlation between two variables is a measure of the strength and direction of the **linear** relationship between them. It does **not** (directly) a measure of other kinds of relationships (for instance monotonic or polynomial). See also Anscombe's quartet.

Correlation is often denoted by the Greek letter

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

for the population parameter (the true but unknown)

where:

- $\text{Cov}(X, Y)$ is the covariance between variables X and Y,
- σ_X and σ_Y are the population standard deviations of X and Y, respectively.

and

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

for the sample estimate.

Remember: We use r to estimate ρ using the sample.

The correlation coefficient can take values between -1 and 1.

- A value of 1 indicates a perfect positive linear relationship between the variables.
- A value of -1 indicates a perfect negative linear relationship between the variables.

See here and here for a nice illustration. Study this in detail. A correlation of ±1 just means that the data points lie on a straight line. It does not say anything about the steepness of the line.

You can watch these videos to get a better understanding of the concept:

- StatQuest - correlation
- StatQuest - covariance
- Animated correlation

Correlation is an often (over-)used measure in research. It is important to note that correlation does not imply causation. For instance, chocolate consumption is positively correlated with the number of Nobel laureates in a country. This does - unfortunately - not mean that eating chocolate makes you smarter.

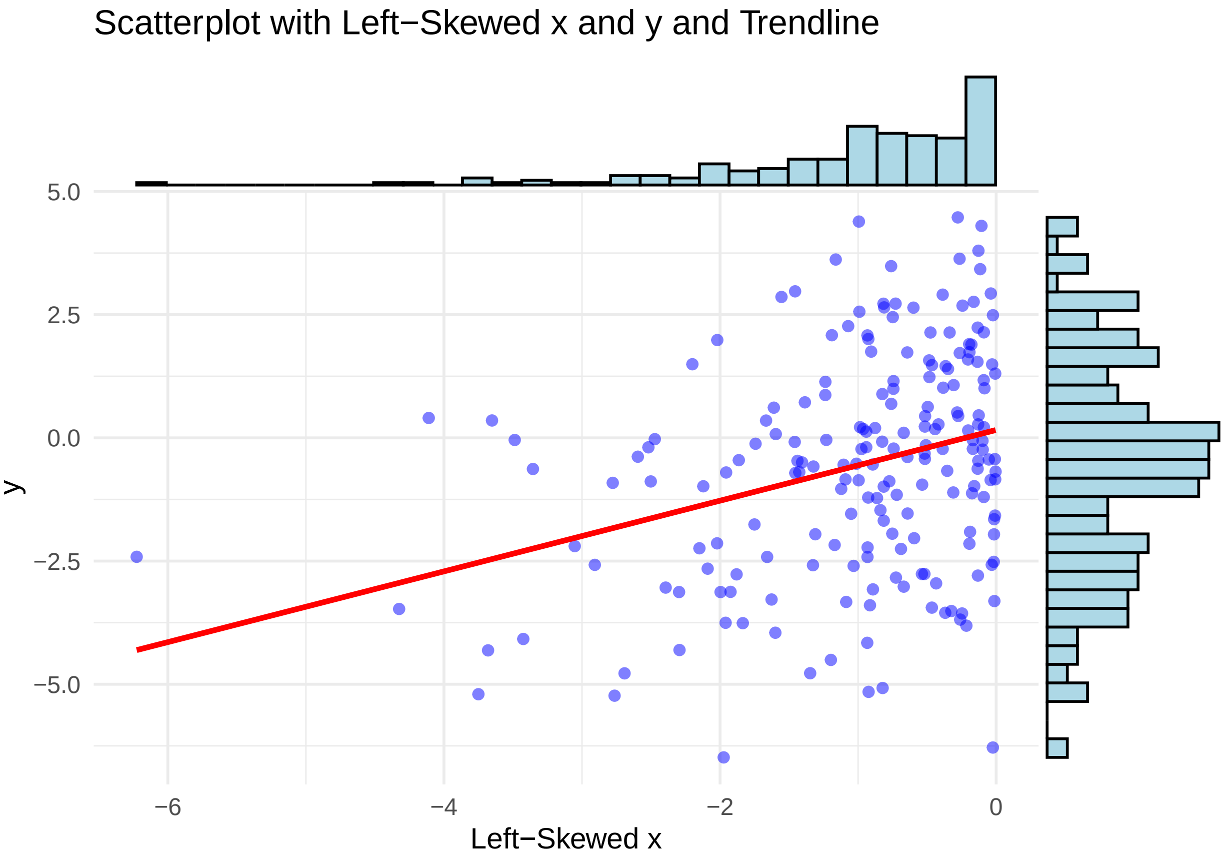
One thing many of us think when they see a high correlation (near 1) is that if we know the value of one variable, we can predict the value of the other variable. This is not necessarily true. This section explains why. Study the visualization here. It depicts the reduction of the prediction interval for Y given X as the correlation increases (assuming that the variables are jointly normally distributed). **How much smaller is the interval where my next observation of Y will fall if I know the value of X?** As you can see, this curve is relatively flat in the beginning, meaning, on an individual level, the correlation does not tell us much about the value of Y given X. Of course with correlation 1, we can predict the value of Y perfectly, but even with a correlation of 0.5 (which is considered high in many areas), the prediction interval for Y is only 13% smaller.

One often reads that the two variables used for calculation of a Pearson correlation coefficient should be normally distributed. This is not necessary, at least not for descriptive purposes. The correlation coefficient is a measure of the linear relationship and we can think of an example with skewed data where the correlation coefficient is still meaningful:

```r
# Load required libraries
library(ggplot2)
library(ggExtra)

# Set seed for reproducibility
set.seed(1234)

# Parameters
n <- 200  # Sample size
beta <- 0.7  # Desired beta

# Generate left-skewed variable x (negative exponential distribution)
x <- -rexp(n, rate = 1)

# Generate correlated left-skewed variable y
error <- rnorm(n, mean = 0, sd = 2)
```

```
y <- beta * x + error

# Create a data frame
data <- data.frame(x = x, y = y)

# Create scatterplot with marginal histograms and trendline
p <- ggplot(data, aes(x = x, y = y)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Scatterplot with Left-Skewed x and y and Trendline",
       x = "Left-Skewed x", y = "y") +
  theme_minimal()

# Add marginal histograms using ggExtra
ggExtra::ggMarginal(p, type = "histogram", fill = "lightblue", color = "black")
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



If X and Y are bivariate normally distributed, then one can use the variable $t = r\sqrt{\frac{n-2}{1-r^2}}$ as test statistic for the null hypothesis (next chapter) $H_0 : \rho = 0$. Note: It is not sufficient for X and Y to be individually normally distributed in order to be jointly normally distributed.

3.4.1 Example in the wild

Let's again visit the previous paper. In the results, correlations are presented in a dichotomous way with regards to p-values, which should be avoided. We will expand on this in the next chapter. For large sample sizes, even small (true) correlations will be "significant", especially, if one decides to use an arbitrary threshold for the p-values like 0.05. There is not much information in this fact. I would argue that all p-values for such large sample sizes and larger correlations do not carry much information. See also Exercise 5.

That was all nice, what could go wrong?

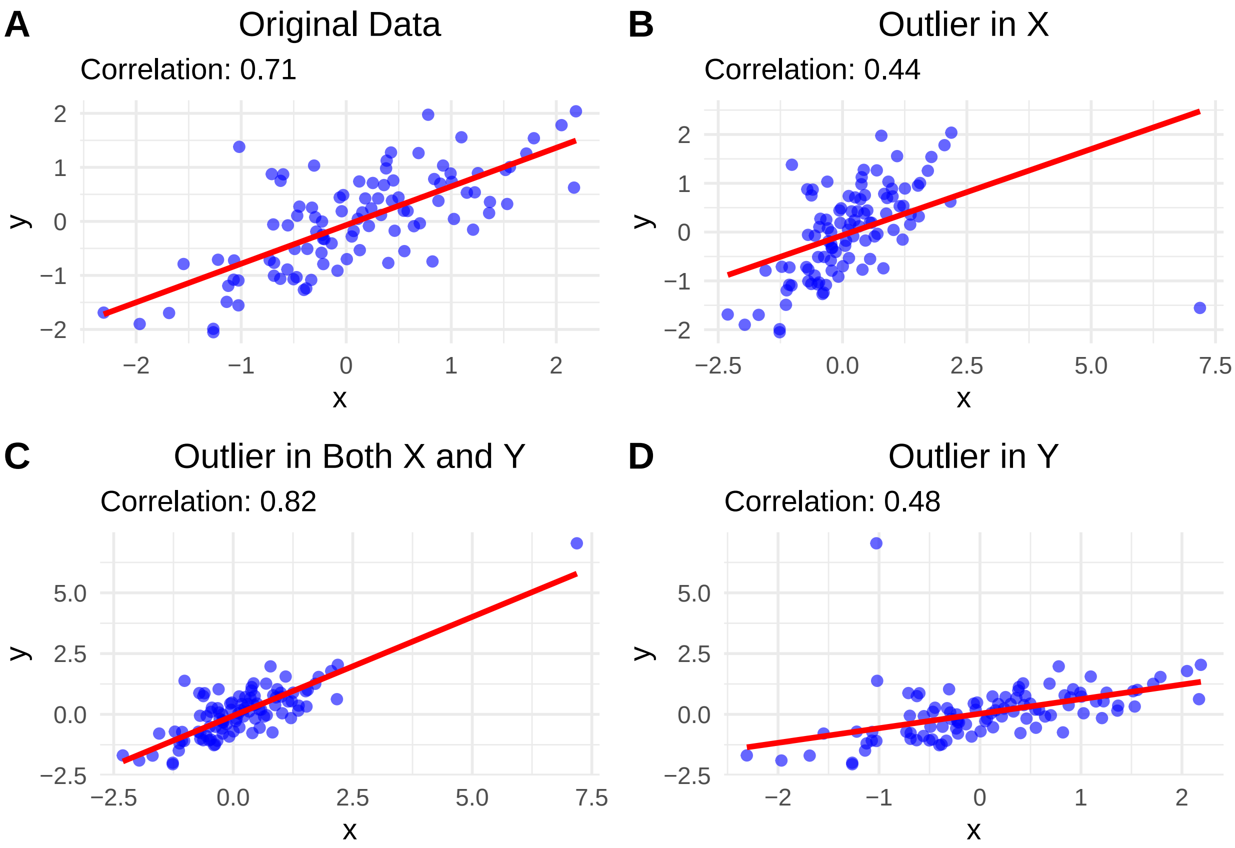
(The sample) Correlation (coefficient) is based (see the formula above) on the artihmetic mean $\bar{X}$ (capital X since we are talking about the random variable which is realised when the X_i have materialized into the x_i) and sample standard deviation σ.

The sample mean is not robust against outliers. A single large value can distort the mean arbitrarily much. Hence, the correlation coefficient is not robust against outliers. Let's see this in action:

```r
# Load necessary libraries
library(ggplot2)
library(ggpubr)

# Set seed for reproducibility
set.seed(123)

# Parameters
n <- 100  # Sample size
rho <- 0.75  # Desired correlation

# Generate bivariate normal data with specified correlation
x <- rnorm(n)
y <- rho * x + sqrt(1 - rho^2) * rnorm(n)

# Store original data
data_original <- data.frame(x = x, y = y)

# Function to create scatterplot with trend line and correlation
create_plot <- function(data, title) {
  ggplot(data, aes(x = x, y = y)) +
    geom_point(alpha = 0.6, color = "blue") +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(title = title, subtitle = paste("Correlation:",
                                         round(cor(data$x, data$y), 2))) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))  # Center the title
}

# Original scatterplot
p_original <- create_plot(data_original, "Original Data")

# Add outlier in x
data_outlier_x <- data_original
data_outlier_x$x[n] <- max(x) + 5  # Extreme value in x
p_outlier_x <- create_plot(data_outlier_x, "Outlier in X")

# Add outlier in both x and y
data_outlier_xy <- data_original
```

```
data_outlier_xy$x[n] <- max(x) + 5  # Extreme value in x
data_outlier_xy$y[n] <- max(y) + 5  # Extreme value in y
p_outlier_xy <- create_plot(data_outlier_xy, "Outlier in Both X and Y")

# Add outlier in y
data_outlier_y <- data_original
data_outlier_y$y[n] <- max(y) + 5  # Extreme value in y
p_outlier_y <- create_plot(data_outlier_y, "Outlier in Y")

# Arrange plots using ggarrange
ggarrange(p_original, p_outlier_x, p_outlier_xy, p_outlier_y,
          labels = c("A", "B", "C", "D"),
          ncol = 2, nrow = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



- A: $r = 0.71$ for the original data.
- B: When we add an outlier in x, r decreases to 0.44 and the trendline is notably shifted.

  The reason for the smaller correlation is that the outlier is above the mean of the x values and below the mean of the y values and therefore adds negatively to the covariance.

- C: When we add an outlier in both x and y, r stays relatively stable. The reason is that the outlier is above the mean of the x values and above the mean of the y values, so it adds positively to the covariance. Also, the outlier is in line with the trend of the data. We could ask, what happend in this data point? Was another scale used for instance?
- D: When we add an outlier in y, the r decreases to 0.48. The reason for the smaller correlation is that the outlier is above the mean of the y values and below the mean of the x values and therefore adds negatively to the covariance.
- We could also ask why the trend lines (created with simple linear regression) change the way they do. We will come to this later in our courses.

3.5 Exercises

3.5.1 Exercise 1 - Recreate table with fake data

- Create fake data for the study mentioned above in R.
- Recreate Table 2 of the paper mentioned above with fake data in R (using GPT, the R package *gtsummary* and other useful packages). This is rather helpful later on in your master thesis.
- Try to export the table to Excel and Word.

3.5.2 Exercise 2 - Outliers and estimates

Let's assume we know that the *Cognitive flexibility* is normally distributed with a mean of 60 and a standard deviation of 7.4 (Table 1): $CognFlex \sim N(60, 7.4)$.

- Draw a sample of 165 persons from this distribution and calculate the mean and standard deviation of the sample. How good is the estimate of the true (and in this case: known) mean and standard deviation?
- Let's replace some of the data points with outliers. Change the score of 5 persons with to impossible CognFlex score of 100. Calculate the mean and standard deviation of the sample. How do the estimates change?
- When we try to estimate the location of our *Cognitive flexibility* distribution with the median, how many outliers of what magnitude are necessary to disturb the estimate by 5 points?

3.5.3 Exercise 3 - Recreating data in Table 2

We assume that *age* in both groups is normally distributed with a mean of 48 (47) years and a standard deviation of 16 (15) years:

$Age_{UsualCare} \sim Normal(\mu = 48, \sigma = 16)$ and

$Age_{CFT} \sim Normal(\mu = 47, \sigma = 15)$.

- Under these assumptions, what is the probability, that we would see a person of age 60 or older in a new sample (in either group)?
- What is the probability, that we would see a person of age 18 or younger in a new sample?
- Give a 99% interval for the age in CFT, where we would expect a new person drawn from the same population.

Let's assume *Sex* is binomially distributed with a probability of $p = 0.59$ for both groups.

- What is the probability, that we would see a woman as the next recruited person in the *Usual care* group?
- What is the probability, that we would see no man in a sample of 10 persons in the *CFT* group?
- Sometimes you want balanced samples. How many patients would we need to recruit to get at least 45 women with a probability of at least 90%. (We could solve this with simulation.)

3.5.4 Exercise 4 - Z-scores

- Show with a simulation that the z-scores are standard normally distributed if the original variable is normally distributed.
- Try different parameter values for μ and σ and plot the histogram of the z-scores.
- Optional: Try to prove this mathematically.

3.5.5 Exercise 5 - Correlation

The following R-code creates correlated samples with a (true) correlation ρ:

```r
set.seed(1234)
n <- 1000
rho <- 0.5
x <- rnorm(n)
y <- rho * x + sqrt(1 - rho^2) * rnorm(n)
cor(x, y)
```

```
## [1] 0.541743
```

- Try different values for ρ and n and plot the scatterplot of x and y.
- Execute the code above 1000 times and save the correlation coefficient in a vector.
- Calculate the sample quantiles and interquantile range of the simulated correlation coefficients. What do you observe with respect to variability?
- Plot the histogram of the simulated correlation coefficients.

3.6 Solutions

Solutions for this chapter can be found here.

Bibliography

Peter Westfall. *Understanding Regression Analysis: An Conditional Distribution Approach*. Wiley, 2020.