

Introduction to Data Analysis

Michael Franke

last rendered at: 2020-02-04 15:58:32

Contents

Preface	7
I. Preliminaries	9
1. General Introduction	11
1.1. Learning goals	11
1.2. Course structure	11
1.3. Tools and topics covered (and not covered) here	12
1.4. Data sets covered	13
1.5. Installation	13
2. Basics of R	19
2.1. First steps	20
2.2. Data types	25
2.3. Functions	38
2.4. Loops and maps	41
2.5. Piping	43
2.6. Rmarkdown	44
II. Data	45
3. Data, variables & experimental designs	47
3.1. Different kinds of data	47
3.2. On the notion of “variables”	49
3.3. Basics of experimental design	52
4. Data Wrangling	57
4.1. Data in, data out	57
4.2. Tidy data	58
4.3. Data manipulation: the basics	61
4.4. Grouped operations	70
4.5. Case study: the King of France	72
5. Summary statistics	77
5.1. Counts and proportions	78
5.2. Central tendency and dispersion	81
5.3. Co-variance & correlation	95

6. Data Visualization	101
6.1. Motivating example: Anscombe's quartet	101
6.2. Visualization: the good, the bad and the info-graphic	103
6.3. Basics of ggplot	110
6.4. A rendezvous with popular geoms	129
6.5. Faceting	150
6.6. Customization etc.	153
III. Models and inferences	165
7. Basics of Probability Theory	167
7.1. Probability	167
7.2. Structured events & marginal distributions	172
7.3. Conditional probability	174
7.4. Random variables	177
7.5. Probability distributions in R	184
8. Models	185
8.1. Probabilistic models in statistics	185
8.2. Parameters, priors, probability and predictions	190
8.3. Three pillars of data analysis	195
8.4. Notation & graphical representation	197
8.5. Strolling the zoo of models	200
8.6. Expressing hypotheses with models	206
9. Parameter estimation	209
9.1. Bayes rule of parameter estimation	210
9.2. A frequentist approach to parameter estimation	218
9.3. Addressing point-valued hypotheses with parameter estimation	224
9.4. Comparing Bayesian and frequentist estimates	225
9.5. Algorithms for parameter estimation	230
9.6. Probabilistic modeling with <i>greta</i>	233
10. Hypothesis Testing	243
10.1. <i>p</i> -values	244
10.2. Central Limit Theorem	255
10.3. Selected tests	258
10.4. Three approaches	283
10.5. Relation to model checking Section	284
11. Model Comparison	285
11.1. Case study: recall models	286
11.2. Akaike Information Criterion	287
11.3. Likelihood-Ratio Test	292
11.4. Bayes factors	295

11.5. Outlook	301
12. Bayesian hypothesis testing	303
12.1. Data and models for this chapter	303
12.2. Testing as posterior estimation	305
12.3. The Savage-Dickey method	308
12.4. Bayes factors for ROPE-d hypotheses through encompassing models	311
IV. Applied (generalized) linear modeling	315
13. Simple linear regression	317
13.1. Data set: murder data	317
13.2. What is a (simple) linear regression?	320
13.3. Ordinary least-squares regression	325
13.4. A maximum-likelihood approach	328
13.5. A Bayesian approach	331
13.6. Testing coefficients	334
14. Beyond simple linear regression	339
14.1. Two categorical predictors	339
14.2. More than two categorical predictors	343
14.3. Interaction terms in factorial designs	345
A. Further useful material	349
A.1. Material on <i>Introduction to Probability</i> :	349
A.2. Material on <i>Bayesian Data Analysis</i> :	349
A.3. Material on frequentist statistics:	349
A.4. Material on <i>R, tidyverse, etc.</i> :	349
A.5. Further information for RStudio	350
A.6. Resources on WebPPL	350
B. Common probability distributions	351
B.1. Selected continuous distributions of random variables	351
B.2. Selected discrete distributions of random variables	376
B.3. Understanding distributions as random variables	393
C. Exponential Family and Maximum Entropy	401
C.1. An important family: The Exponential Family	401
C.2. Excursions: “Information Entropy” and “Maximum Entropy Principal”	402
D. Data sets used in the book	409
D.1. Mental Chronometry	409
D.2. Simon Task	420
D.3. World Values Survey (wave 6 2010-2014)	428
D.4. King of France	429

Contents

D.5. Bio-Logic Jazz-Metal (and where to consume it)	438
D.6. Avocado prices	443

Preface

This book is the basic reading material for the course “Introduction to Data Analysis”, held at the University of Osnabrück in the winter term of 2019, as part of the BSc *Cognitive Science* program. It introduces key concepts of data analysis from a frequentist and a Bayesian tradition. It uses R to handle, plot and analyze data. It relies on simulation to illustrate selected statistical concepts.

Part I.

Preliminaries

1. General Introduction

This chapter lays out the learning goals of this book (Section 1.1) and describes how these goals are to be achieved (Section 1.2). Section 1.3 details which technical tools and methods are covered here, and which are not. There will be some information on the kinds of data sets we will use during the course in Section 1.4. Finally, Section 1.5 provides information about how to install the necessary tools for this course.

1.1. Learning goals

At the end of this course students should:

- feel confident to pick up any data set to explore in a hypothesis-driven manner
- have gained the competence to
 - understand complex data sets,
 - manipulate a data set, so as to
 - plot aspects of it in ways that are useful for answering a given research question
- understand the general logic of statistical inference in frequentist and Bayesian approach
- be able to independently evaluate statistical analyses based on their adequacy for a given research question and data set
- be able to critically assess the adequacy of analyses commonly found in the literature

Notice that this is, although a lot of hard work already, still rather modest! It doesn't actually say that we necessarily aim at the competence to *do it* or even to *do it flawlessly!* **Our main goal is understanding**, because that is the foundation of practical success *and* the foundation of an ability to learn more in the future. **We do not teach tricks! We do not share recipes!**

1.2. Course structure

The course consists of four parts. After giving a more detailed overview of the course, Part I introduces R the main programming language that we will use. Part II covers what is often called *descriptive statistics*. It also gives us room to learn more about R when we massage data into shape, compute summary statistics and plot various different data types in various different ways.

Part III is the main theoretical part. It covers what is often called *inferential statistics*. Two aspects distinguish this course from the bulk of its cousins out there. First, we use a **dual-pronged approach**, i.e., we are going to introduce both the frequentist and the Bayesian approach to statistical inference side by side. The

1. General Introduction

motivation for this is that seeing the contrast between the two approaches will aid our understanding of either one. Second, we will use a **computational approach**, i.e., we foster understanding of mathematical notions with computer simulations or other variants of helpful code.

Part IV covers applications of what we have learned so far. It focuses on **generalized linear models**, a class of models that have become the new standard for analyses of experimental data in the social and psychological sciences, but are also very useful for data exploration in other domains (such as machine learning).

There are also appendices with additional information:

- Further useful material (textbooks, manuals, etc.) is provided in Appendix A.
- Appendix B covers the most important probability distributions used in this book.
- An excursion providing more information about the important Exponential Family of probability distributions and the Maximum Entropy Principle is given in Appendix C.
- The data sets which reoccur throughout the book as “running examples” are succinctly summarized in Appendix D.

1.3. Tools and topics covered (and not covered) here

The main programming language used in this course is R (R Core Team 2018). We will make heavy use of the *tidyverse* package (Wickham 2017), which provides a unified set of functions and conventions that deviate (sometimes: substantially) from basic R. We will also be using the probabilistic programming language WebPPL (Goodman and Stuhlmüller 2014), but only “passively” in order to quickly obtain results from probabilistic calculations that we can experiment with directly in the browser. We will not learn to write WebPPL code from scratch.

We will rely on the R package `brms` (Bürkner 2017) for running Bayesian generalized regression models, which itself relies on the probabilistic programming language `Stan` (Carpenter et al. 2017). We will, however, not learn about `Stan` in this course. Instead of `Stan` we will use the package `greta` (Golding 2019) to write our models and do inference with them. This is because, for current learning purposes, the language in which `greta` formulates its models is much closer to R and so, let’s hope, easier to learn.

Section 1.5 gives information about how to install these, and other, tools necessary for this course.

The main topics that this course will cover are:

- **data preparation:** how to clean up, and massage a data set into shape for plotting and analysis
- **data visualization:** how to select aspects of data to visualize in informative and useful ways
- **statistical models:** what that is, and why it’s beneficial to think in terms of models, not tests
- **statistical inference:** what that is, and how it’s done in frequentist and Bayesian approaches
- **hypothesis testing:** how Frequentists and Bayesians test scientific hypotheses
- **generalized regression:** how to apply GRMs to different types of data sets

There is, obviously, a lot that we will *not* cover in this course. We will, for instance, not dwell at any length on the specifics of algorithms for computing statistical inferences or model fits. We will also deal with the history and the philosophy of science of statistics only to the extent that it helps understand the theoretical notions and practical habits that are important in the context of this course. We will also not do heavy math.

Data analysis can be quite varied, because data itself can be quite varied. We try to sample some variation, but since this is an introductory course with lots of other ground to cover, we will be slightly conservative in the kind of data that we analyze. There will, for example, not be any pictures, sounds, dates or time points in any of the material covered here.

There are at least two different motivations for data analysis, and it is important to keep them apart. This course focuses on **data analysis for explanation**, i.e., routines that help us understand reality through the inspection and massaging of empirical data. We will only glance at the alternative approach, which is **data analysis for prediction**, i.e., using models to predict future observations, as commonly practiced in machine learning and its applications. In sloppy slogan form, this course treats data science for scientific knowledge gain, not the engineers' applications.

1.4. Data sets covered

We want to learn how to do data analysis. This is impossible without laying hands (keys?) on several data sets. But switching from one data set to another is mentally taxing. It is also difficult to focus and really care about any-old data set. This is why this course relies on a small selection of recurring data sets that are, hopefully, generally interesting for the target audience: students of cognitive science. Appendix D gives an overview of the most important, recurring data sets used in this course.

Most of the data sets that we will use repeatedly in this class come from various psychological experiments. To make this even more emersive, these experiments are implemented as browser-based experiments, using `_magpie`. This makes it possible for students of this course to do the exact experiments whose data we are analyzing (and maybe generate some more intuitions, maybe generate some hypotheses) about the very data at hand. But it also makes it possible that we will analyze ourselves. That's why part of the exercises for this course will run additional analyses on data collected from the aspiring data analysts themselves. If you want to become an analyst, you should also have undergone analysis yourself, so to speak.

1.5. Installation

This course relies on a few different pieces of software. Primarily, we'll be using R, but we'll need installations of Python and C++ in the background.

There are two options for installing. The simplest method, described in Section 1.5.1 is to install VirtualBox and use our provided Ubuntu virtual machine (link provided in class), which has the required software pre-installed and tested. When using this method, you will be working in a virtualized Linux environment. Alternatively, you can go through a manual installation tailored to your own OS, as described in Section 1.5.2. Manual installation is recommended if you do not wish to use a virtualized Linux environment. The

1. General Introduction

VirtualBox method can be a fall-back option if manual installation fails. Both methods of installation are detailed below. Finally, Section 1.5.3 explains how to update the R package that wraps all R packages needed for this course and provides some extra convenience functions.

1.5.1. VirtualBox Setup

1.5.1.1. Step 1. Install VirtualBox

Follow the instructions here to download and install for your platform.

1.5.1.2. Step 2. Download our Ubuntu image

Download the provided VirtualBox Disk Image and move it into a folder such as “VMs”. The link for the VirtualBox Disk Image is provided on StudIP.

1.5.1.3. Step 3. Create a new virtual machine and add the downloaded image

- Open VirtualBox and click New
- Give your new virtual machine a name, e.g. “IDA2019”
- Change the Machine Folder to the folder where you put the disk image
- Change the Type to “Linux” and Version to “Ubuntu (64-bit)” and click Next to proceed to memory allocation
- Allocate about half of your available memory and click Next to proceed to hard disk selection
- Choose “Use an existing virtual hard disk file” and use the file selection icon to add our provided disk image
- Click Create

1.5.1.4. Step 4. Give your virtual machine more processing power

- Select your virtual machine on the right panel and click Settings on the top
- Navigate to System -> Processor
- Increase the Processor(s) to about half of what your computer can provide

1.5.1.5. Step 5. Boot your virtual machine

- Select your virtual machine and click Start
- The username of the system is “user” and the password is “password”

1.5.1.6. Step 6. Install further packages

The virtual machine includes most of the packages required, but not all.

- First, run the following commands in 'Terminal':

```
sudo apt update
```

```
sudo apt install r-cran-devtools r-cran-boot r-cran-extradistr r-cran-ggsignif  
r-cran-naniar
```

- Then, open RStudio and run the following command in the R console:

```
devtools::install_github("n-kall/IDA2019-package")
```

1.5.1.7. Troubleshooting

- If the virtual machine does not boot, you may need to ensure that 'virtualization' is enabled in your computer's BIOS. Talk to a tutor if you're having difficulties doing so.

1.5.2. Manual installation

If you don't want to use the virtual machine, or it doesn't work for you, the following six steps describe how to get the main components installed manually. Depending on your operating system (e.g. macOS, Linux, Windows), you might need to follow slightly different instructions, which are specified. Depending on the exact setup of your computer, the results may vary. The virtual machine has been tested with the required software, so if you can, we recommend using that.

1.5.2.1. Step 1. Install Python

Windows and macOS:

We recommend installing miniconda from here

Linux:

You can install miniconda or you can just use the preinstalled Python (which saves time and space). Make sure you have pip installed, e.g. for Ubuntu `apt install python3-pip`

1.5.2.2. Step 2. Install the required Python packages

We have provided files that list the required Python packages. They can be installed automatically with the following commands in the terminal.

For Anaconda:

Download this environment file

1. General Introduction

```
conda env create -f environment.yml
```

For Linux users who are using pip:

Download this requirements file

```
pip3 install -r requirements.txt
```

1.5.2.3. Step 3. Install R

Windows and macOS:

Download and install R from [here](#)

Linux users:

We need to have at least version 3.5 of R. This may be available in your distribution's repository. e.g. if you are using a recent version of Ubuntu (18.10 or later), you can install R with `apt install r-base`. Otherwise, follow these instructions for your version.

1.5.2.4. Step 4. Install RStudio

All platforms:

Download and install the latest version of RStudio from [here](#)

1.5.2.5. Step 5. Install a C++ toolchain

For the Stan language, which will be interfaced through an R package called `brms`, you'll need a working C++ compiler.

Windows:

Download and install the latest version of RTools from [here](#)

macOS:

Download and install the latest version of RTools for macOS from [here](#)

Note: you may need to register for an Apple Developer account (free of charge).

Linux (e.g. Ubuntu):

You can install a compiler and toolchain with `apt install build-essential`.

1.5.2.6. Step 6. Install R packages

We have created an R package that, when installed, will prompt the installation of the packages we will use in this course. In this step, you'll install this package (and automatically install its dependencies).

For Windows and macOS:

Open RStudio and run the following two commands in the console.

```
install.packages("devtools")
devtools::install_github("n-kall/IDA2019-package")
```

For Linux (e.g. Ubuntu):

It's much faster (and less error-prone) if you install devtools from the app repository via terminal: `apt install r-cran-devtools` and continue to the second line in the R console. But this will only work if you are using a recent version of Ubuntu (18.10 or later).

If you had to manually update your R to version in step 3, you'll need to install the following from terminal:

```
apt install libssl-dev libxml2-dev libcurl4-openssl-dev
```

and then install via the the R console:

```
install.packages("devtools")
devtools::install_github("n-kall/IDA2019-package")
```

1.5.3. Updating the course package

Occasionally, we might have to add packages or functionality as we go through this course. In that case, you will have to update the package `IDA2019-package` that ships all of this for this course. To update, use:

```
devtools::install_github("n-kall/IDA2019-package")
```


2. Basics of R

R is a specialized programming language for data science. Though old, it is heavily supported by an active community. New tools for data handling, visualization, and statistical analysis are provided in the form of **packages**.¹ While other programming languages specialized for scientific computing, like Python or Julia, also lend themselves beautifully for data analysis, the choice of R in this course is motivated because R's *raison d'être* is data analysis. Some of the R packages that this course will use provide cutting-edge methods which are not as conveniently available in other programming languages (yet).

In a manner of speaking, there are two flavors of R. We should distinguish **base R** from the **tidyverse**. Base R is what you have when you do not load any packages. We enter the tidyverse by loading the package `tidyverse` (see below for information on how to do that). The tidyverse consists of several components (which are actually stand-alone packages that can be loaded separately if needed) all of which supply extra functionality for data analysis, based on a unifying philosophy and representation format. While eventually interchangeable, the look-and-feel of base R and the tidyverse is quite different. Figure 2.1 lists a selection of packages from the tidyverse in relation to their role at different stages of the process of data analysis. The image is taken from this introduction to the tidyverse.

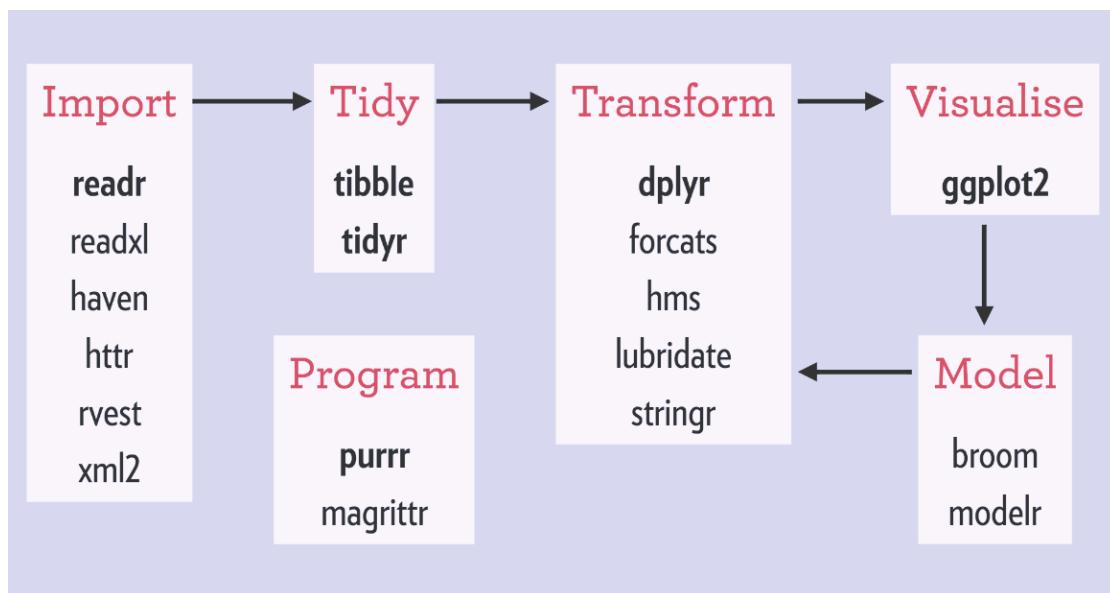


Figure 2.1.: Overview of selected packages from the tidyverse.

¹Packages live in the official package repository CRAN, or are supplied in less standardized forms, e.g., via open repositories, such as GitHub.

2. Basics of R

The course will also introduce Rmarkdown in Section 2.6. Rmarkdown is a nice way of documenting your data analyses in a reproducible form. Participants will use Rmarkdown to prepare their homework assignments.

Make sure to have completely installed everything of relevance for this course, as described in Section 1.5. Unless you have strong opinions or an unassailable favorite, we recommend trying RStudio as an IDE for R.

The official documentation for base R is An Introduction to R. The standard reference for using the tidyverse is R for Data Science (R4DS). There are some very useful cheat sheets which you should definitely check out! There are pointers to further material in Appendix ??.

The learning goals for this chapter are:

- become familiar with R, its syntax and basic notions
- become familiar with the key functionality from the tidyverse
- understand and write simple R scripts
- be able to write documents in Rmarkdown

2.1. First steps

R is an interpreted language. This means that you do not have to compile it. You can just evaluate it line by line, in a so-called **session**. The session stores the current values of all variables. If you do not want to retype, you can store your code in a **script**.²

Try this out by either typing `r` to open an R session in a terminal or load RStudio.³ You can immediately calculate stuff:

```
6 * 7
```

```
## [1] 42
```

2.1.1. Functions

R has many built-in functions. The most common situation is that the function is called by its name using **prefix notation**, followed by round brackets which enclose the function's arguments (separated by commas if multiple). For example, the function `round` takes a number and, per default, returns the closest integer:

²Line-by-line execution of code is useful for quick development and debugging. Make sure to learn about keyboard shortcuts to execute single lines or chunks of code in your favorite editor, e.g., check the RStudio Cheat Sheet for information on its keyboard shortcuts.

³When starting a session in a terminal, you can exit a running R session by typing `quit()` or `q()`.

```
# the function `round` takes a number as argument and
# returns the closest integer (default)
round(0.6)

## [1] 1
```

Actually, `round` allows several arguments. It takes as input the number `x` to be rounded, and another integer number `digits` which gives the number of digits after the comma to which `x` should be rounded. We can then specify these arguments in a function call of `round` by providing the named arguments.

```
# rounds the number `x` to the number `digits` of digits
round(x = 0.138, digits = 2)

## [1] 0.14
```

When providing all arguments with names, the order of arguments does not matter. When providing at least one non-named argument, all non-named arguments have to be presented in the right order (as expected by the function; to find out what that is use `help`, as explained below in 2.1.6) after subtracting the named arguments from the ordered list of arguments.

```
round(x = 0.138, digits = 2) # works as intended
round(digits = 2, x = 0.138) # works as intended
round(0.138, digits = 2) # works as intended
round(0.138, 2) # works as intended
round(x = 0.138, 2) # works as intended
round(digits = 2, 0.138) # works as intended
round(2, x = 0.138) # works as intended
round(2, 0.138) # does not work as intended (returns 2)
```

Functions can have default values for some or all of their arguments. In the case of `round` the default is `digits = 0`. There is obviously no default for `x` in the function `round`.

Some functions can take an arbitrary number of arguments. The function `sum`, which sums up numbers is a point in case.

```
# adds all of its arguments together
sum(1,2,3)

## [1] 6
```

Selected functions can also be called in **infix notation**. This applies to frequently recurring operations, such as mathematical operations or logical comparisons.

2. Basics of R

```
# both of these calls sum 1, 2, and 3 together
sum(1,2,3)      # prefix notation
1 + 2 + 3      # prefix notation
```

Section 2.3 will list some of the most important built-in functions. It will also explain how to define your own functions.

2.1.2. Variables

You can assign values to variables using three assignment operators: `->`, `<-` and `=`, like so:

```
x <- 6          # assigns 6 to variable x
7 -> y          # assigns 7 to variable y
z = 3            # assigns 3 to variable z
x * y / z       # returns 6 * 7 / 3 = 14

## [1] 14
```

Use of `=` is discouraged.⁴

It is good practice to use a consistent naming scheme for variables. This book uses `snake_case_variable_names` and tends towards using `long_and_excessively_informative_names` for important variables, and short variable names, like `i`, `j` or `x`, for local variables, indices etc.

2.1.3. Literate coding

It is good practice to document code with short but informative comments. Comments in R are demarcated with `#`.

```
x <- 4711 # a nice number from Cologne
```

Since everything on a line after an occurrence of `#` is treated as a comment, it is possible to break long function calls across several lines, and to add comments to each line:

```
round(           # call the function `round`
  x = 0.138,     # number to be rounded
  digits = 2      # number of after-comma digits to round to
)
```

⁴You can produce `<-` in RStudio with Option-- (on Mac) and Alt-- (on Windows/Linux). For other useful keyboard shortcuts, see [here](#).

In RStudio, you can use Command+Shift+C (on Mac) and Ctrl+Shift+C (on Windows/Linux) to comment or uncomment code, and you can use comments to structure your scripts. Any comment followed by ---- is treated as a (foldable) section.

```
# SECTION: variable assignments ----
x <- 6
y <- 7
# SECTION: some calculations ----
x * y
```

2.1.4. Objects

Strictly speaking, all entities in R are *objects* but that is not always apparent or important for everyday practical purposes see the manual for more information. R supports an object-oriented programming style, but we will not make (explicit) use of this functionality. In fact, this course heavily uses and encourages a functional programming style (see Section 2.4).

Some functions (e.g., optimizers or fitting functions for statistical models) return objects, however, and we will use this output in various ways. For example, if we run a linear regression model on some data set, the output is an object.

```
# you do not need to understand this code
model_fit = lm(formula = speed~dist, data = cars)
# just notice that the function `lm` returns an object
is.object(model_fit)

## [1] TRUE

# printing an object on the screen usually gives you summary information
print(model_fit)

##
## Call:
## lm(formula = speed ~ dist, data = cars)
##
## Coefficients:
## (Intercept)          dist
##           8.2839        0.1656
```

2.1.5. Packages

Much of R's charm unfolds through the use of packages. CRAN has the official package repository. To install a new package from a CRAN mirror use the `install.packages` function. For example, to install the package `devtools`, you would use:

2. Basics of R

```
install.packages("devtools")
```

Once installed, you need to load your desired packages for each fresh session, using:

```
library(devtools)
```

Once loaded all functions, data etc. that ship with a package are available without additional reference to the package name. If you want to be careful or courteous to an admirer of your code, you can reference the package a function comes from explicitly. For example, the following code calls the function `install_github` from the package `devtools` explicitly (so that you would not need to load the package beforehand, for example):

```
devtools::install_github("SOME-URL")
```

Indeed, the `install_github` function allows you to install bleeding-edge packages from GitHub. You can install all of the relevant packages using (after installing the `devtools` package, as described in Section 1.5):

```
devtools::install_github("n-kall/IDA2019-package")
```

After this installation, you can load all packages for this course simply by using:

```
library(IDA2019)
```

In RStudio, there is a special tab in the pane with information on “files”, “plots” etc. to show all installed packages. This also shows which packages are currently loaded.

2.1.6. Getting help

If you encounter a function like `lm` that you do not know about, you can access its documentation with the `help` function or just typing `?lm`. For example, the following call summons the documentation for `lm`, the first parts are shown in Figure 2.2.

```
help(lm)
```

```
knitr::include_graphics("visuals/R-doc-example.png")
```

If you are looking for help on a more general topic, use the function `help.search`. It takes a regular expression as input and outputs a list of occurrences in the available documentation. A useful shortcut for `help.search` is just to type `??` followed by the (unquoted) string to search for. For example, calling either of the following lines might produce a display like in Figure 2.3.

lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments

<code>formula</code>	an object of class " <code>formula</code> " (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
<code>data</code>	an optional data frame, list or environment (or object coercible by <code>as.data.frame</code> to a data frame) containing the variables in the model. If not found in <code>data</code> , the variables are taken from <code>environment(formula)</code> , typically the environment from which <code>lm</code> is called.

Figure 2.2.: Excerpt from the documentation of the 'lm' function.

```
# two equivalent ways for obtaining help on search term 'linear'
help.search("linear")
??linear
```

```
knitr:::include_graphics("visuals/R-doc-search-example.png")
```

The top entries in Figure 2.3 are **vignettes**. These are compact manuals or tutorial on particular topics or functions, and they are directly available in R. If you want to browse through the vignettes available on your machine (which depend on which packages you have installed), go ahead:

```
browseVignettes()
```

2.2. Data types

Let's briefly go through the data types that are most important for our later purposes. We can assess the type of an object stored in variable `x` with the function `typeof(x)`.

```
typeof(3)          # returns type "double"
typeof(TRUE)       # returns type "logical"
typeof(cars)       # returns 'list' (includes data.frames, tibbles, objects, ...)
typeof("huhu")     # return 'character' (= string)
typeof(mean)        # return 'closure' (= function)
typeof(c)           # return 'builtin' (= deep system internal stuff)
typeof(round)       # returns type "special" (= well, special stuff?)
```

2. Basics of R

Search Results 



Vignettes:

brms::brms_nonlinear	Estimating Non-Linear Models with brms	HTML	source	R code
knitr::docco-linear	R Markdown with the Docco Linear Style	HTML	source	R code
lme4::lmer	Fitting Linear Mixed-Effects Models using lme4	PDF	source	R code
RcppEigen::RcppEigen-Introduction	RcppEigen-intro	PDF	source	R code
SparseM::SparseM	An Introduction to the SparseM Package for Sparse Linear Algebra	PDF	source	R code

Code demonstrations:

quantreg::Frank	Demo of nonlinear in parameters fitting of Frank copula model	(Run demo in console)
SparseM::LeastSquares	Least Squares Linear Regression	(Run demo in console)
SparseM::LinearAlgebra	Basic Linear Algebra for Sparse Matrices	(Run demo in console)
SparseM::Solve	Linear Equation Solving	(Run demo in console)
stats::lm.glm	Some linear and generalized linear modelling examples from 'An Introduction to Statistical Modelling' by Annette Dobson	(Run demo in console)
stats::nlm	Nonlinear least-squares using nlm()	(Run demo in console)

Help pages:

arrayhelpers::colMeans.array-method	Row and column sums and means for numeric arrays.
---	---

Figure 2.3.: Result of calling 'help.search' for the term 'linear'.

To learn more about an object, it can help to just print it out as a string:

```
str(lm)
```

```
## function (formula, data, subset, weights, na.action, method = "qr", model = TRUE,
##   x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL,
##   offset, ...)
```

It is sometimes possible to cast objects of one type into another type XXX using functions `as.XXX` in base R or `as_XXX` in the tidyverse.

```
# casting Boolean value `TRUE` into number format
as.numeric(TRUE) # returns 1

## [1] 1
```

R is essentially an array-based language. Arrays are arbitrary but finite dimensional matrices. We will discuss what is usually referred to as vectors (= one-dimensional arrays), matrices (= two-dimensional arrays) and arrays (= more-than-two-dimensional) in the following section on numeric information. But it is important to keep in mind that arrays can contain objects of other types than numeric information (as long as all objects in the array are of the same type).

2.2.1. Numeric vectors & matrices

2.2.1.1. Numeric information

Standard number format in R is double.

```
typeof(3)
```

```
## [1] "double"
```

We can also represent numbers as integers and complex.

```
typeof(as.integer(3)) # returns 'integer'

## [1] "integer"
```

2. Basics of R

```
typeof(as.complex(3))      # returns 'complex'  
  
## [1] "complex"
```

2.2.1.2. Numeric vectors

As a generally useful heuristic, expect every numerical information to be treated as a vector (or higher-order: matrix, array, ... ; see below), and to expect any (basic, mathematical) operation in R to (most likely) apply to the whole vector, matrix, array, collection.⁵ This makes it possible to ask for the length of a variable to which we assing a single number, for instance:

```
x <- 7  
length(x)  
  
## [1] 1
```

We can even index such a variable:

```
x <- 7  
x[1]      # what is the entry in position 1 of the vector x?  
  
## [1] 7
```

Or assign a new value to a hitherto unused index:

```
x[3] <- 6      # assign the value 6 to the 3rd entry of vector x  
x          # notice that the 2nd entry is undefined, or "NA", not available  
  
## [1] 7 NA 6
```

Vectors in general can be declared with the built-in function `c()`. To memorize this, think of *concatenation* or *combination*.

```
x <- c(4, 7, 1, 1)    # this is now a 4-place vector  
x  
  
## [1] 4 7 1 1
```

There are also helpful functions to generate sequences of numbers:

⁵If you are familiar with Python's `scipy` and `numpy` packages, this is R's default mode of treating numerical information.

```
1:10                                # returns 1, 2, 3, ..., 10
seq(from = 1, to = 10, by = 1)       # returns 1, 2, 3, ..., 10
seq(from = 1, to = 10, by = 0.5)     # returns 1, 1.5, 2, ..., 9.5, 10
seq(from = 0, to = 1, length.out = 11) # returns 0, 0.1, ..., 0.9, 1
```

Indexing in R starts with 1, not 0!

```
x <- c(4, 7, 1, 1)    # this is now a 4-place vector
x[2]
```

```
## [1] 7
```

And now we see what is meant above when we said that (almost) every mathematical operation can be expected to apply to a vector:

```
x <- c(4, 7, 1, 1)    # 4-placed vector as before
x + 1

## [1] 5 8 2 2
```

2.2.1.3. Numeric matrices

Matrices are declared with the function `matrix`. This function takes, for instance, a vector as an argument.

```
x <- c(4, 7, 1, 1)      # 4-placed vector as before
(m <- matrix(x))       # cast x into matrix format

##      [,1]
## [1,]    4
## [2,]    7
## [3,]    1
## [4,]    1
```

Notice that the result is a matrix with a single column. This is important. R uses so-called *column-major mode*.⁶ This means that it will fill columns first. For example, a matrix with three columns based on a six-placed vector 1, 2, ..., 6 will be built by filling the first column from top to bottom, then the second column top to bottom, and so on.⁷

⁶Python, on the other hand, uses the reverse *row-major mode*.

⁷It is in this sense that the “first index moves fastest” in column-major mode, which is another frequently given explanation of column-major mode.

2. Basics of R

```
m <- matrix(1:6, ncol = 3)
m

##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

In line with column-major mode, vectors are treated as column vectors in matrix operations:

```
x = c(1,0,1)    # 3-place vector
m %*% x          # dot product with previous matrix 'm'

##      [,1]
## [1,]    6
## [2,]    8
```

As usual, and independently of column- or row-major mode, matrix indexing starts with the row index:

```
m[1,]    # produces first row of matrix 'm'

## [1] 1 3 5
```

2.2.1.4. Arrays

Arrays are simply higher-dimensional matrices. We will not make use of arrays in this course.

2.2.1.5. Names for vectors, matrices and arrays

The positions in a vector can be given names. This is extremely useful for good “literate coding” and therefore highly recommended. The names of vector x’s positions are retrieved and set by the `names` function:

```
students <- c("Jax", "Jamie", "Jason")    # names of students
grades <- c(1.3, 2.7, 2.0)                  # a vector of grades
names(grades)                                # retrieve names: with no names so far

## NULL
```

```

names(grades) <- students          # assign names
names(grades)
# retrieve names again: names assigned

## [1] "Jax"   "Jamie" "Jason"

grades                         # output shows names

##   Jax Jamie Jason
##   1.3   2.7   2.0

```

We can also set the names of a vector directly during construction:⁸

```

# names of students (this is a character vector, see below)
students <- c("Jax", "Jamie", "Jason")
# constructing a vector with names directly assigned
grades <- c(1.3, 2.7, 2.0, names = students)

```

Names for matrices are retrieved or set with functions `rownames` and `colnames`.

```

# declare matrix
m <- matrix(1:6, ncol = 3)
# assign row and column names, using function
# `str_c` which is described below
rownames(m) <- str_c("row", 1:nrow(m), sep = "_")
colnames(m) <- str_c("col", 1:ncol(m), sep = "_")
m

##      col_1 col_2 col_3
## row_1    1    3    5
## row_2    2    4    6

```

2.2.2. Booleans

There are built-in names for Boolean values “true” and “false”, predictably named `TRUE` and `FALSE`. Equivalent shortcuts are `T` and `F`. If we attempt to do math with Boolean vectors, the outcome is what any reasonable logician would expect:

⁸Notice that we can create strings (actually called ‘characters’ in R) with double quotes

2. Basics of R

```
x <- c(T,F,T)
1 - x

## [1] 0 1 0

x + 3

## [1] 4 3 4
```

Boolean vectors can be used as index sets to extract elements from other vectors.

```
# vector 1, 2, ..., 5
number_vector <- 1:5
# index of odd numbers set to `TRUE`
boolean_vector <- c(T,F,T,F,T)
# returns the elemnts from number vector, for which
# the corresponding element in the Boolean vector is true
number_vector[boolean_vector]

## [1] 1 3 5
```

2.2.3. Special values

There are a couple of keywords reserved in R for special kinds of objects:

- NA: “not availables”; represents missing values in data
- NaN: “not a number”; e.g., division zero by zero
- Inf or -Inf: infinity and negative infinity; returned when number is too big or devision by zero
- NULL: the NULL object; often returned when function is undefined for input

2.2.4. Characters (= strings)

Strings are called characters in R. We will be stubborn and call them strings for most of the time here. We can assign a string value to a variable by putting the string in double quotes:

```
x <- "huhu"
typeof(x)

## [1] "character"
```

We can create vectors of characters in the obvious way:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
chr_vector

## [1] "huhu"  "hello"  "huhu"  "ciao"
```

The package `stringr` from the tidyverse also provides very useful and, in comparison to base R, more uniform functions for string manipulation. The cheat sheet for the `stringr` package is highly recommended for a quick overview. Below are some examples.

Function `str_c` concatenates strings:

```
str_c("Hello", "Hi", "Hey", sep = "! ")

## [1] "Hello! Hi! Hey!"
```

We can find the indeces of matches in a character vector with `str_which`:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
str_which(chr_vector, "hu")

## [1] 1 3
```

Similarly, `str_detect` gives a Boolean vector of matching:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
str_detect(chr_vector, "hu")

## [1] TRUE FALSE TRUE FALSE
```

If we want to get the strings matching a pattern, we can use `str_subset`:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
str_subset(chr_vector, "hu")

## [1] "huhu" "huhu"
```

Replacing all matches with another string works with `str_replace_all`:

2. Basics of R

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
str_replace_all(chr_vector, "h", "B")

## [1] "BuBu"  "Bello" "BuBu"  "ciao"
```

For data preparation we often need to split strings by a particular character. For instance, a set of reaction times could be separated by a character line "|". We can split this string representation to get individual measurements like so:

```
# three measures of reaction time in a single string
reaction_times <- "123|234|345"
# notice that we need to doubly (!) escape character |
# notice also that the results is a list (see below)
str_split(reaction_times, "\\|", n = 3)

## [[1]]
## [1] "123" "234" "345"
```

2.2.5. Factors

Factors are special vectors, which treat its elements as ordered or unorderd categories. This is useful for representing data from experiments, e.g., of categorical or ordinal variables (see Chapter 3). To create a factor, we can use the function `factor`. The following code creates an *unorderd factor*:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
factor(chr_vector)

## [1] huhu hello huhu ciao
## Levels: ciao hello huhu
```

Ordered factors also register the order of the categories:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
factor(
  chr_vector,      # the vector to treat as factor
  ordered = T,    # make sure its treated as ordered factor
  levels = c("huhu", "ciao", "hello") # specify order of levels
)

## [1] huhu hello huhu ciao
## Levels: huhu < ciao < hello
```

We will see that ordered factors are important, for example, in plotting when they determine the order in which different parts of data are arranged on the screen. They are also important for statistical analysis, because they help determine how categories are compared to one another.

Factors are trickier to work with than mere lists, because they are rigid about the represented factor levels. Adding an item that does not belong to any of a factor's levels, leads to trouble:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
my_factor <- factor(
  chr_vector,      # the vector to treat as factor
  ordered = T,     # make sure its treated as ordered factor
  levels = c("huhu", "ciao", "hello") # specify order of levels
)
my_factor[5] <- "huhu" # adding a "known category" is okay
my_factor[6] <- "moin" # adding an "unknown category" does not work
my_factor

## [1] huhu hello huhu ciao huhu <NA>
## Levels: huhu < ciao < hello
```

The `forcats` package from the `tidyverse` helps dealing with factors. You should check the Cheat Sheet for more helpful functionality. Here is an example of how to expand the levels of a factor:

```
chr_vector <- c("huhu", "hello", "huhu", "ciao")
my_factor <- factor(
  chr_vector,      # the vector to treat as factor
  ordered = T,     # make sure its treated as ordered factor
  levels = c("huhu", "ciao", "hello") # specify order of levels
)
my_factor[5] <- "huhu" # adding a "known category" is okay
my_factor <- fct_expand(my_factor, "moin") # add new category
my_factor[6] <- "moin" # adding new item now works
my_factor

## [1] huhu hello huhu ciao huhu moin
## Levels: huhu < ciao < hello < moin
```

It is sometimes useful (especially for plotting) to flexibly reorder the levels of an ordered factor. Here are some useful functions from the `forcats` package:

```
my_factor          # original factor
```

2. Basics of R

```
## [1] huhu hello huhu ciao huhu moin
## Levels: huhu < ciao < hello < moin

fct_rev(my_factor)      # reverse level order

## [1] huhu hello huhu ciao huhu moin
## Levels: moin < hello < ciao < huhu

fct_relevel(            # manually supply new level order
  my_factor,
  c("hello", "ciao", "huhu")
)

## [1] huhu hello huhu ciao huhu moin
## Levels: hello < ciao < huhu < moin
```

2.2.6. Lists, data frames & tibbles

Lists are key-value pairs. They are created with the built-in function `list`. The difference between a list and a named vector is that in the latter all elements must be of the same type. In a list, the elements can be of arbitrary type. They can also be vectors or even lists themselves. For example:

```
my_list <- list(
  single_number = 42,
  chr_vector   = c("huhu", "ciao"),
  nested_list   = list(x = 1, y = 2, z = 3)
)
my_list

## $single_number
## [1] 42
##
## $chr_vector
## [1] "huhu" "ciao"
##
## $nested_list
## $nested_list$x
## [1] 1
##
## $nested_list$y
## [1] 2
```

```
##  
## $nested_list$z  
## [1] 3
```

To access a list element by its name (=key), we can use the \$ sign followed by the unquoted name, double square brackets [["name"]] with the quoted name inside, or indices in double brackets, like so:

```
# all of these return the same list element  
my_list$chr_vector
```

```
## [1] "huhu" "ciao"
```

```
my_list[["chr_vector"]]
```

```
## [1] "huhu" "ciao"
```

```
my_list[[2]]
```

```
## [1] "huhu" "ciao"
```

Lists are very important in R because almost all structured data that belongs together is stored as lists. Objects are special kinds of lists. Data is stored in special kinds of lists, so-called *data frames* or so-called *tibbles*.

A data frame is base R's standard format to store data in. A data frame is a list of vectors of equal length. Data sets are instantiated with the function `data.frame`:

```
# fake experimental data  
exp_data <- data.frame(  
  trial = 1:5,  
  condition = factor(  
    c("C1", "C2", "C1", "C3", "C2"),  
    ordered = T  
)  
  response = c(121, 133, 119, 102, 156)  
)  
exp_data
```

```
##   trial condition response  
## 1      1         C1      121  
## 2      2         C2      133  
## 3      3         C1      119  
## 4      4         C3      102  
## 5      5         C2      156
```

2. Basics of R

We can access columns of a data frame, just like we access elements in a list. Additionally, we can also use index notation, like in a matrix:

```
# gives the value of the cell in row 2, column 3  
exp_data[2,3] # return 133  
  
## [1] 133
```

In RStudio, you can inspect data in data frames (and tibbles (see below)) with the function `View`.

Tibbles are the tidyverse counterpart of data frames. We can cast a data frame into a tibble, using `as_tibble`.

```
as_tibble(exp_data)  
  
## # A tibble: 5 x 3  
##   trial condition response  
##   <int> <ord>     <dbl>  
## 1      1    C1       121  
## 2      2    C2       133  
## 3      3    C1       119  
## 4      4    C3       102  
## 5      5    C2       156
```

But we can also create a tibble directly with the keyword `tibble`. Indeed, creation of tibbles is conveniently more flexible than the creation of data frames: the former allow dynamic look-up of previously defined elements.

```
my_tibble    <- tibble(x = 1:10, y = x^2)      # dynamic construction possible  
my_dataframe <- data.frame(x = 1:10, y = x^2)  # ERROR :/
```

Another important difference between data frames and tibbles concerns default treatment of character (=string) vectors. When reading in data from a CSV file as a data frame (using function `read.csv`) each character vector is treated as a factor per default. But when using `read_csv` to read CSV data into a tibble character vector are not treated as factors.

2.3. Functions

2.3.1. Some important built-in functions

Many helpful functions are defined in base R or supplied by packages. We recommend browsing the Cheat Sheets every now and then to pick up more useful stuff for your inventory. Here are some functions that are very basic and generally useful.

2.3.1.1. Standard logic

- `&`: "and"
- `|`: "or"
- `!`: "not"
- `negate()`: a pipe-friendly ! (see Section 2.5 for more on piping)
- `all()`: returns true of a vector if all elements are T
- `any()`: returns true of a vector if at least one element is T

2.3.1.2. Comparisons

- `<`: smaller
- `>`: greater
- `==`: equal (you can also use `near()` instead of `==` e.g. `near(3/3, 1)` returns TRUE)
- `>=`: greater or equal
- `<=`: less or equal
- `!=`: not equal

2.3.1.3. Set theory

- `%in%`: whether an element is in a vector
- `union(x, y)`: union of x and y
- `intersect(x, y)`: intersection of x and y
- `setdiff(x, y)`: all elements in x that are not in y

2.3.1.4. Sampling and combinatorics

- `rnorm()`: random number from unit interval [0;1]
- `sample(x, size, replace)`: take size samples from x (with replacement if `replace` is T)
- `choose(n, k)`: number of subsets of size n out of a set of size k (binomial coefficient)

2.3.2. Defining your own functions

If you find yourself in a situation in which you would like to copy-paste some code, possibly with minor amendments, this usually means that you should wrap some recurring operations into a custom-defined function.

There are two ways of defining your own functions: as a named function, or an anonymous function.

2. Basics of R

2.3.2.1. Named functions

The special operator supplied by base R to create new functions is the keyword `function`. Here is an example of defining a new function with two input variables `x` and `y` that returns a computation based on these numbers. We assign this newly created function to the variable `cool_function`, so that we can use this name to call the function later. Notice that the use of the `return` keyword is optional here. If it is left out, the evaluation of the last line is returned.

```
# define a new function
# takes two numbers x & y as argument
# return x * y + 1
cool_function <- function(x, y) {
  return(x * y + 1)
}

# apply `cool_function` to some numbers:
cool_function(3,3)      # return 10
cool_function(1,1)      # return 2
cool_function(1:2,1)    # returns vector [2,3]
cool_function(1)        # throws error: 'argument "y" is missing, with no default'
cool_function()         # throws error: 'argument "x" is missing, with no default'
```

We can give default values for the parameters passed to a function:

```
# same function as before but with
# default values for each argument
cool_function_2 <- function(x = 2, y = 3) {
  return(x * y + 1)
}

# apply `cool_function_2` to some numbers:
cool_function_2(3,3)      # return 10
cool_function_2(1,1)      # return 2
cool_function_2(1:2,1)    # returns vector [2,3]
cool_function_2(1)        # returns 4 (= 1 * 3 + 1)
cool_function_2()          # returns 7 (= 2 * 3 + 1)
```

2.3.2.2. Anonymous functions

Notice that we can feed functions as parameters to other functions. This is an important ingredient of a functional-style of programming, and something that we will rely on heavily in this course (see Section 2.4). When supplying a function as an argument to another function, we might not want to name the function that is passed. Here's a (stupid, but hopefully illustrating) example:

```
# define a function that takes a function as argument
new_applier_function <- function(input, function_to_apply) {
  return(function_to_apply(input))
}

# sum vector with built-in & named function
new_applier_function(
  input = 1:2,                      # input vector
  function_to_apply = sum            # built-in & named function to apply
)    # returns 3

# sum vector with anonymous function
new_applier_function(
  input = 1:2,                      # input vector
  function_to_apply = function(input) {
    return(input[1] + input[2])
}
)    # returns 3 as well
```

2.4. Loops and maps

For iteratively performing computation steps, R has a special syntax for `for` loops. Here is an example of a (stupid, but illustrative) example of a `for` loop in R:

```
# fix a vector to transform
input_vector <- 1:6

# create output vector for memory allocation
output_vector <- integer(length(input_vector))

# iterate over length of input
for (i in 1:length(input_vector)) {
  # multiply by 10 if even
  if (input_vector[i] %% 2 == 0) {
    output_vector[i] = input_vector[i] * 10
  }
  else {
    output_vector[i] = input_vector[i]
  }
}

output_vector
```

2. Basics of R

```
## [1] 1 20 3 40 5 60
```

R also provides functional iterators (e.g., `apply`), but we will use the functional iterators from the `purrr` package. The main functional operator from `purrr` is `map` which takes a vector and a function, applies the function to each element in the vector and returns a list with the outcome. There are also versions of `map`, written as `map_dbl` (double), `map_lgl` (logical) or `map_df` (data frame), which return a vector of doubles, Booleans or a data frame. Here is a first example of how this code looks in a functional style using the functional iterator `map_dbl`:

```
map_dbl(  
  input_vector,  
  function(i) {  
    if (input_vector[i] %% 2 == 0) {  
      return (input_vector[i] * 10)  
    }  
    else {  
      return (input_vector[i])  
    }  
  })  
  
## [1] 1 20 3 40 5 60
```

We can write this even shorter, using `purrr`'s short-hand notation for functions:

```
map_dbl(  
  input_vector,  
  ~ ifelse( .x %% 2 == 0, .x * 10, .x)  
)  
  
## [1] 1 20 3 40 5 60
```

The trailing `~` indicates that we define an anonymous function. It therefore replaces the usual `function(...)` call which indicates which arguments the anonymous function expects. To make up for this, after the `~` we can use `.x` for the first (and only) argument of our anonymous function.

To apply a function to more than one input vector, element per element, we can use `pmap` and its derivatives, like `pmap_dbl` etc. `pmap` takes a list of vectors and a function. In short-hand notation we can define an anonymous function with `~` and integers like `..1`, `..2` etc, for the first, second ... argument. For example:

```

x <- 1:3
y <- 4:6
z <- 7:9

pmap_dbl(
  list(x, y, z),
  ~ ..1 - ..2 + ..3
)

## [1] 4 5 6

```

2.5. Piping

When we use a functional style of programming, piping is your best friend. Consider the standard example of applying functions in what linguists would call “center-embedding”. We start with the input (written inside the inner-most bracketing), then apply the first function `round`, then the second `mean`, writing each next function call “around” the previous.

```

# define input
input_vector <- c(0.4, 0.5, 0.6)

# first round, then take mean
mean(round(input_vector))

## [1] 0.3333333

```

Things quickly get out of hand when more commands are nested. A common practice is to store intermediate results of computations in new variables which are only used to pass the result into the next step.

```

# define input
input_vector <- c(0.4, 0.5, 0.6)

# rounded input
rounded_input <- round(input_vector)

# mean of rounded input
mean(rounded_input)

## [1] 0.3333333

```

Piping let’s you pass the result of a previous function call into the next. The `magrittr` package supplies a special infix operator `%>%` for piping.⁹ The pipe `%>%` essentially takes what results from evaluating the

⁹The pipe symbol `%>%` can be inserted in RStudio with Ctrl+Shift+M (Win/Linux) or Cmd+Shift+M (Mac).

2. Basics of R

expression on its left-hand side and inputs it as the first argument in the function on its right-hand side. So `x %>% f` is equivalent to `f(x)`. Or, to continue the example from above, we can now write:

```
input_vector %>% round %>% mean  
  
## [1] 0.3333333
```

The functions defined as part of the tidyverse are all constructed in such a way that the first argument is the most likely input you would like to pipe into them. But if you want to pipe the left-hand side into another argument slot than the first, you can do that by using the `.` notation to mark the slot where the left-hand side should be piped into: `y %>% f(x, .)` is equivalent to `f(x, y)`.

2.6. Rmarkdown

Homework assignments will be issued, filled and submitted in Rmarkdown. To get familiar with Rmarkdown, please follow this tutorial.

Part II.

Data

3. Data, variables & experimental designs

The focus of this course is on data from behavioral psychology experiments.¹ In a sense, this is perhaps the most “well-behaved” data to analyze and therefore an excellent starting point into data analysis. However, we should not lose sight of the rich and diverse guises of data that are relevant for scientific purposes. The current chapter therefore starts, in Section 3.1, with sketching some of that richness and diversity. But it then hones in on some basic distinctions of the kinds of data we will frequently deal with in the cognitive sciences in Section 3.2. We also pick up a few relevant concepts from standard experimental design in Section 3.3.

The learning goals for this chapter are:

- appreciate the diversity of data
- distinguish different kinds of variables
 - dependent vs independent
 - nominal vs ordinal vs metric
- get familiar with basic aspects of experimental design
 - factorial designs
 - within- vs between subjects design
 - repeated measures
 - randomization, fillers and controls
 - sample size
- understand the notion of “tidy data”

3.1. Different kinds of data

Some say we live in the **data age**. But what is data actually? Purist pedants say: “The plural of datum” and add that a datum is just an observation. But when we say “data” we usually mean a bit more than a bunch of observations. The Merriam-Webster offers the following:

¹A *behavioral experiment* is an experiment that records participants’ behavioral choices, such as button clicks or linguistic responses in the form of text or speech. This contrasts with, say, *neurological experiments* in which participants’ brain activity is recorded, such as fMRI or EEG, or, e.g., in a psycholinguistic context, *processing-related experiments* in which secondary measures of cognitive activity are measured, such as eye-movements, pupil dilation or galvanic skin responses.

3. Data, variables & experimental designs

Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.

This is a teleological definition in the sense that it refers to the purpose of the thing: “used as basis for reasoning, discussion, or calculation”. So, what we mean by “data” is in large part defined by what we intend to do with it. Another important aspect of this definition is that we usually consider data to be systematically structured in some way or another. Even when we speak of “raw data”, we expect there to be some structure (maybe labels, categories etc.) that distinguishes data from uninterpretable noise (e.g., the notion of a “variable”, discussed in Section 3.2). In sum, we can say that **data is a representation of information stored in a systematic way for the purpose of inference, argument or decision making**.

There are different kinds of data. Figure 3.1 shows some basic distinctions, represented in a conceptual hierarchy.

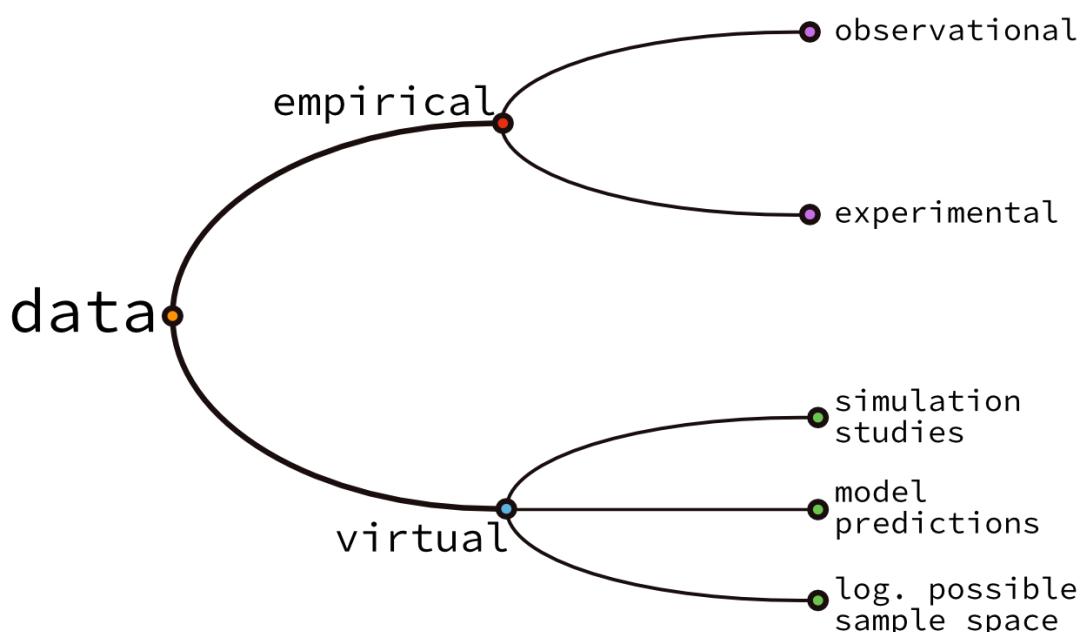


Figure 3.1.: Hierarchy of different kinds of data relevant for 'data science'.

It is easy but wrong to think that data always has to be information based on observations of the world. It is easy to think this because **empirical data**, i.e., data obtained from empirical observation, is the most common form of data (given that it is, arguably, most relevant for decision making and argument). But it is wrong to think this because we can just as well look at **virtual data**. For example, virtual data which is of interest to a data analyst could be **data obtained from computer simulation studies**, e.g., from, say, a 1 billion runs of a multi-agent simulation intended to shed light on the nature of cooperative interaction. It makes sense to analyse such data with the same tools as data from an experiment. We might find out that some parameter constellations in the simulation run are (statistically) most conducive of producing cooperative behavior among our agents, for instance. Another example of virtual data is **data generated as**

Table 3.1.: Comparison of pro's and cons's of observational data and experimental data.

observational	experimental
ecological valid	possibly artificial
easy/easier to obtain	hard/harder to obtain
correlation & causation hard to tease apart	may yield information on causation vs. correlation

predictions of a model, which we can use to test whether that model is any good, in so-called model criticism (see Section ??).² Finally, we should also include **logically possible sample data** in this list, because of its importance to central ideas of statistical inference (especially p -values, see Section 10). Logically possible sample data is that was neither observed, nor predicted by a model, but something that could have been observed hypothetically, something that it is merely logically possible to observe, even if it would almost never happen in reality or would not be predicted by any serious model.

The most frequent form of data, **empirical data** about the actual world, comes in two major variants. **Observational data** is data gathered by (passively) observing and recording what would have happened even if we had not been interested in it, so to speak. Examples of observational data are collections of socio-economic variables, like gender, education, income, number of children etc. In contrast, **experimental data** is data recorded in a strict regime of manipulation-and-observation, i.e., a scientific experiment. Some pieces of information can only be recorded in an observational study (annual income) and others can only be obtained through experimentation (memory span). Both methods of data acquisition have their own pros and cons. Here are some of the more salient ones:

No matter what kind of data we have at hand, there are at least two prominent purposes for which data can be useful: **explanation** and **prediction**. Though related, it is useful to keep these purposes cleanly apart. Data analysis for explanation uses the data to better understand the source of the data (the world, a computer simulation, a model, etc.). Data analysis for prediction tries to extract regularities from the data gathered so far to make predictions (as accurately as possible) about future or hitherto unobserved data.

3.2. On the notion of “variables”

Data used for data analysis, even if it is “raw data”, i.e., data before preprocessing and cleaning, is usually structured or labelled in some way or other. Even if the whole data we have is a vector of numbers, we would usually know what these numbers represent. For instance, we might just have a quintuple of numbers, but we would (usually, ideally) know that these represent the results of an IQ test.

```
# a simple data vector of IQ-scores
IQ_scores <- c(102, 115, 97, 126, 87)
```

²We will later speak of **prior/posterior predictions** for this kind of data. Other applicable terms are **repeat data** or sometimes **fake data**.

3. Data, variables & experimental designs

Or we might have a Boolean vector with the information of whether each of five students passed an exam. But even then we would (usually/ideally) know the association between names and test results, as in a table like this:

```
# who passed the exam
exam_results <-
  tribble(
    ~student,    ~pass,
    "Jax",      TRUE,
    "Jason",    FALSE,
    "Jamie",    TRUE
  )
```

Association of information, as between different columns in a table like the one above, is crucial. Most often we have more than one kind of observation that we care about. Most often, we care about systematic relationships between different observables in the world. For instance, we might want to look at pass/fail results from an exam in co-occurrence with information about the proportion of attendance of the course's tutorial sessions:

```
# proportion of tutorials attended and exam pass/fail
exam_results <-
  tribble(
    ~student,    ~tutorial_proportion,    ~pass,
    "Jax",        0.0,                  TRUE,
    "Jason",     0.78,                 FALSE,
    "Jamie",     0.39,                 TRUE
  )
exam_results

## # A tibble: 3 x 3
##   student tutorial_proportion pass
##   <chr>              <dbl> <lgl>
## 1 Jax                0     TRUE
## 2 Jason              0.78 FALSE
## 3 Jamie              0.39 TRUE
```

Data of this kind is also called **rectangular data**, i.e., data that fits into a rectangular table (More on the structure of rectangular data in Section 4.2.). In the example above, every column represents a **variable** of interest. A (*data*) **variable** stores the observations that are of the same kind.³

Common kinds of variables are distinguished based on the structural properties of the kinds of observations that they contain. Common types of empirical data collected are:

³This sense of “data variable” is not to be confused with the notion of a “random variable”, a concept we will introduce later in Section 7.4. The term “data variable” is not commonly used; the common term is merely “variable”.

- **nominal variable:** each observation is an instance of a (finite) set of clearly distinct categories, lacking a natural ordering;
- **binary variable:** special case of nominal variable when there are only two categories;
- **Boolean variable:** special case of a binary variable when the two categories are Boolean values "true" and "false";
- **ordinal variable:** each observation is an instance of a (finite) set of clearly distinct and naturally ordered categories, but there is no natural meaning of distance between categories (i.e., it makes sense to say that A is "more" than B but not that A is three times "more" than B);
- **metric variable:** each observation is isomorphic to a subset of the reals, and interval-scaled (i.e., it makes sense to say that A is three times "more" than B);

Examples of some different kinds of variables is shown in Figure 3.2 and Table 3.2 lists common and/or natural ways of representing different kinds of (data) variables in R.

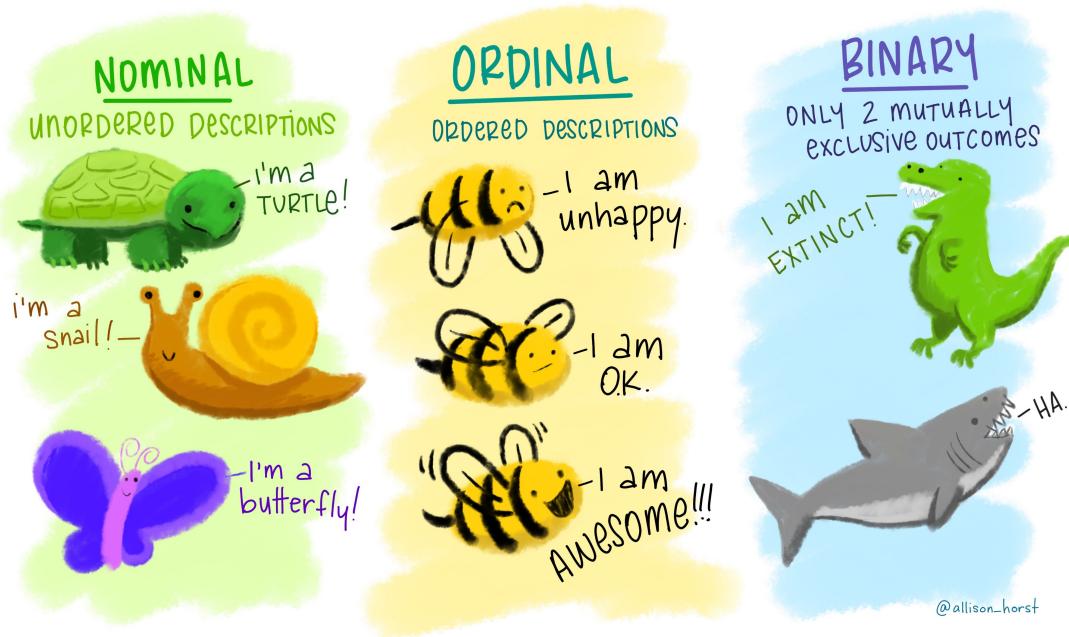


Figure 3.2.: Examples of different kinds of (data) variables.

In experimental data we also distinguish the **dependent variable(s)** from the **independent variables**. The dependent variables are the variables that we do not control or manipulate in the experiment, but the ones that we are curious to record (e.g., whether a patient recovered from an illness within a week). Dependent variables are also called **to-be-explained variables**. The independent variables are the variables in the experiment that we manipulate (e.g., which drug to administer), usually with the intention of seeing a particular effect on the dependent variables. Independent variables are also called **explanatory variables**.

3. Data, variables & experimental designs

Table 3.2.: Common / natural formats for representing data of different kinds in R.

variable type	representation in R
nominal / binary	unordered factor
Boolean	logical vector
ordinal	ordered factor
metric	numeric vector

3.3. Basics of experimental design

The most basic template for an experiment, inspired clearly by the natural sciences, is to just measure a quantity of interest (the dependent variable), without taking into account any kind of variation in any kind of independent variables. For instance, we measure the time it takes for an object, with specific shape and weight, to hit the ground when dropped from exactly 2 meters height. To filter out **measurement noise** we do not just record one observation, but take a fair amount. We use the observed times, for instance, to test a theory about acceleration and gravity. Data from such a simple measurement experiment would be just a single vector of numbers.

A more elaborate kind of experiment would allow for at least one independent variable. Another archetypical example of an empirical experiment would be a medical study, e.g., one in which we are interested in the effect of a particular drug on the blood pressure of patients. We would then randomly allocate each participant to one of two groups. One group, the **treatment group**, receives the drug in question, the other group, **the control group**, receives a placebo (and nobody, not even the experimenter, knows who receives what). After a pre-defined exposure to either drug or placebo, blood pressure (for simplicity just systolic blood pressure) is measured. The interesting question is whether there is a difference between the measurements across groups. This is a simple example of a **one-factor design**. The factor in question is which group any particular measurement belongs to. Data from such an experiment could look like this:

```
tribble(
  ~subj_id,      ~group,        ~systolic,
  1,             "treatment",   118,
  2,             "control",     132,
  3,             "control",     116,
  4,             "treatment",   127,
  5,             "treatment",   122
)

## # A tibble: 5 x 3
##   subj_id group    systolic
##       <dbl> <chr>    <dbl>
## 1 1       treatment 118
## 2 2       control   132
```

```
## 3 control 116
## 4 treatment 127
## 5 treatment 122
```

For the purposes of this course, which is not a course on experimental design, just a few key concepts of experimental design are important to be aware of. We will go through some of these issues in the following.

3.3.1. What to analyze? – Dependent variables

To begin with, it is important to realize that there is quite some variation in what counts as a dependent variable. Not only can there be more than one dependent variable, each dependent variable can also be of quite a different type (nominal, ordinal, metric, ...), as discussed in the previous section. Moreover, we need to carefully distinguish between the actual measurement/observation and the dependent variable itself. The dependent variable is (usually) what we plot, analyze and discuss, but very often we measure much more or something else. The dependent variable (of analysis) could well just be one part of the measurement. For example, a standard measure of blood pressure has a number for systolic and another for diastolic pressure. Focussing on just one of these numbers is a (hopefully: theoretically motivated; possibly: arbitrary; in the worst case: result-oriented) decision of the analyst. More interesting examples of such **data preprocessing** arise frequently in the cognitive sciences, for example:

- **eye-tracking**: the measured data are triples consisting of a time-point and two spatial coordinates, but what might be analyzed is just the relative proportion of looks at a particular spatial region of interest (some object on the screen) in a particular temporal region of interest (up to 200 ms after the image appeared);
- **EEG**: individual measurements obtained by EEG are very noisy, so that the dependent measure in many analyses is an aggregation over the mean voltage recorded by selected electrodes, where averages are taken for a particular subject over many trials of the same condition (repeated measures) that this subject has seen;

But we do not need to go fancy in our experimental methods, to see how issues of data processing affect data analysis at its earliest stages, namely by selecting the dependent variable (that which is to be analyzed). Just take the distinction between **closed questions** and **open questions** in text-based surveys. In closed questions, participants select an answer from a finite (usually) small number of choices. In open questions, however, they can write text freely, or they can draw, sing, pronounce, gesture etc. Open response formats are great and naturalistic, but they, too, often require the analyst to carve out a particular aspect of the (rich, natural) observed reality to enter analysis.

3.3.2. Conditions, trials, items

A **factorial design** is an experiment with at least two independent variables all of which are (ordered or unordered) factors.⁴ Factorial designs are often described in terms of short abbreviations. For example, an

⁴The archetypical medical experiment discussed above is a *one-factor design*. In contrast, the term 'factorial design' is usually used to refer to what is also often called a **full factorial design**. These are designs with at least two independent variables.

3. Data, variables & experimental designs

Table 3.3.: Comparison of pro's and cons's of between- and within-subjects designs.

between-subjects	within-subjects
no confound between conditions	possible cross-contamination between conditions
more participants needed	fewer participants needed
less associated information for analysis	more associated data for analysis

experiment described as a “ 2×3 factorial design” would have two factors of interest, the first of which has two levels, the second of which has three levels (such as a distinction between control and treatment group, and an orthogonal distinction of gender in categories ‘male’, ‘female’ and ‘non-binary’). Many psychological studies are factorial designs. Whole batteries of analyses techniques have been developed specifically tuned to these kinds of experiments.

For a $2 \times 2 \times 3$ factorial design, there are $2 * 2 * 3 = 12$ different **experimental conditions** (also sometimes called **design cells**). An important distinction in experimental design is whether all participants contribute data to all of the experimental conditions, or whether each only contributes to a part of it. If participants only contribute data to a part of all experimental conditions, this is called a **between-subjects design**. If all participants contribute data to all experimental conditions, we speak of a **within-subjects design**. Clearly, sometimes the nature of a design factor determines whether the study can be within-subjects. For example, switching gender for the purpose of a medical study on blood pressure drugs is perhaps a tad much to ask of a participant (though certainly a very enlightening experience). If there is room for the experimenter’s choice of study type, it pays to be aware of some of the clear advantages and draw-backs of either method, as listed in Table 3.3.

No matter whether we are dealing with a between- or within-subjects design, another important question is whether each participant gives us only one or more than one observation per cell. If participants contribute more than one observation to a design cell, we speak of a *repeated measures* design. Such designs are useful as they help separate the signal from the noise (recall the initial example of time measurement from physics). They are also economic because getting several observations worth of relevant data from a single participant for each condition means that we have to get fewer people to do the experiment (normally).

However, exposing a participant to the same experimental condition repeatedly can be detrimental to an experiment’s purpose. Participants might recognize the repetition and develop quick coping strategies to deal with the boredom, for example. For this reason repeated measures designs usually include different kinds of trials:

- **critical trials** belong to, roughly put, the actual experiment, e.g., one of the experiment’s design cells;
- **filler trials** are packaged around the critical trials to prevent blatant repetition, predictability or recognition of the experiment’s purpose
- **control trials** are trials whose data is used not for statistics inference but for checking the quality of the data (e.g., attention checks or tests of whether a participant understood the task correctly)

When participants are exposed to several different kinds of trials and even several instances of the same experimental condition, it is also often important to introduce some variability between the instances of the

same types of trials. Often psychological experiments therefore use different **items**, i.e., different (theoretically exchangeable) instantiations of the same (theoretically important) pattern. For example, if a careful psycholinguist designs a study on the processing of garden-path sentences, she will include not just one example ("The horse raced past the barn fell") but several (e.g., "Since Jones frequently jogs a mile is a short distance to her"). Item-variability is important also for statistical analyses, as we will see when we talk about hierarchical modeling in Section ??.

In longer experiments, especially within-subjects repeated measures designs in which participants encounter a lot of different items for each experimental condition, clever regimes of **randomization** are important to minimize the possible effect of carry-over artifacts, for example. A frequent method is **pseudo-randomization** where the trial sequence is not completely arbitrary but arbitrary within certain constraints, such as a particular **block design**, where each block presents an identical number of trials of each type, but each block shuffles the sequence of its types completely at random.

The complete opposite of a within-participants repeated measures design is a so-called **single-shot experiment** in which any participant gives exactly one data point for one experimental condition.

3.3.3. Sample size

A very important question for experimental design is that of the **sample size**: how many data points do we need (per experimental condition)? We will come back to this issue only much later in this course, when we talk about statistical inference. This is because the decision of how many, say, participants to invite for a study, should ideally be influenced, not by the available time and money, but also by statistical considerations of the kind: how many data points do I need in order to obtain a reasonable level of confidence in the resulting statistical inferences I care about?

4. Data Wrangling

The information relevant for our analysis goals is not always directly accessible. Sometimes we must first uncover it effortfully from an inconvenient representation. Also, sometimes data must be cleaned (ideally: by *a priori* specified criteria) by removing data points that are deemed of insufficient quality for a particular goal. All of this, and more, is the domain of **data wrangling**: preprocessing, cleaning, reshaping, renaming etc. Section 4.1 describes how to read data from and write data to files. Section 4.2 introduces the concept of **tidy data**. We then look at a few common tricks of data manipulation in Section 4.3. We will learn about grouping operations in Section 4.4 Finally, we look at a concrete application in Section 4.5.

The learning goals for this chapter are:

- be able to read from and write data to files
- understand notion of *tidy data*
- be able to solve common problems of data preprocessing

4.1. Data in, data out

The `readr` package handles the reading and writing of data stored in text files.¹ Here is a cheat sheet on the topic: data I/O cheat sheet. In this course will mostly deal with data stored in CSV files.

Reading a data set from a CSV file works with the `read_csv` function:

```
fresh_raw_data <- read_csv("PATH/FILENAME_RAW_DATA.csv")
```

Writing to a csv file can be done with the `write_csv` function:

```
write_csv(processed_data, "PATH/FILENAME_PROCESSED_DATA.csv")
```

If you want to use a different delimiter (between cells) than a comma, you can use `read_delim` and `write_delim` for example, which take an additional argument `delim` to be set to the delimiter in question.

¹Other packages help with reading data from and writing data to other file types, such as excel sheets. Look at the data I/O cheat sheet for more information.

4. Data Wrangling

```
# reading data from a file where cells are (unconventionally) delimited by string
data_from_weird_file <- read_delim("WEIRD_DATA_FILE.TXT", delim = "|")
```

4.2. Tidy data

The same data can be represented in multiple ways. There is even room for variance in the class of rectangular representations of data. Some manners of representations are more useful for certain purposes than for others. For data analysis (plotting, statistical analyses) we prefer to represent our data as (rectangular) **tidy data**.

4.2.1. Running example

Consider the example of student grades for two exams in a course. A compact way of representing the data for visual digestion is the following representation:

```
exam_results_visual <- tribble(
  ~exam,           ~"Rozz",    ~"Andrew",    ~"Siouxsie",
  "midterm",      "1.3",      "2.0",        "1.7",
  "final" ,       "2.3",      "1.7",        "1.0"
)
exam_results_visual

## # A tibble: 2 x 4
##   exam     Rozz   Andrew Siouxsie
##   <chr>    <chr>  <chr>   <chr>
## 1 midterm 1.3    2.0     1.7
## 2 final   2.3    1.7     1.0
```

This is how such data would frequently be represented, e.g., in tables in a journal. Indeed, Rmarkdown helps us present this data in an appetizing manner, e.g., in Table 4.1, which is produced by the code below:

```
knitr::kable(
  exam_results_visual,
  caption = "Fictitious exam results of fictitious students.",
  booktabs = TRUE
)
```

Though highly perspicuous, this representation of the data is not tidy, in the special technical sense we endorse here. A tidy representation of the course results could be this:

Table 4.1.: Fictitious exam results of fictitious students.

exam	Rozz	Andrew	Siouxsie
midterm	1.3	2.0	1.7
final	2.3	1.7	1.0

```

exam_results_tidy <- tribble(
  ~student,      ~exam,      ~grade,
  "Rozz",       "midterm",   1.3,
  "Andrew",     "midterm",   2.0,
  "Siouxsie",   "midterm",   1.7,
  "Rozz",       "final",     2.3,
  "Andrew",     "final",     1.7,
  "Siouxsie",   "final",     1.0
)
exam_results_tidy

## # A tibble: 6 x 3
##   student  exam    grade
##   <chr>    <chr>   <dbl>
## 1 Rozz     midterm  1.3
## 2 Andrew   midterm  2
## 3 Siouxsie midterm  1.7
## 4 Rozz     final   2.3
## 5 Andrew   final   1.7
## 6 Siouxsie final   1

```

4.2.2. Definition of *tidy* data

Following Wickham (2014), a tidy representation of (rectangular) data is defined as one where:

1. each variable forms a column,
2. each observation forms a row, and
3. each type of observational unit forms a table.

Any data set that is not tidy is **messy data**. Messy data that satisfies the first two constraints, but not the third will be called **almost tidy data** in this course. We will work, wherever possible, with data that is at least almost tidy. Figure 4.1 shows a graphical representation of the concept of tidy data.

4. Data Wrangling



Figure 4.1.: Organization of tidy data (taken from @wickham2016).

4.2.3. Excursion: non-redundant data

The final condition in the definition of tidy data is not particularly important for us here (since we will make do with ‘almost tidy data’), but to understand it nonetheless consider the following data set:

```
exam_results_overloaded <- tribble(
  ~student,      ~stu_number,      ~exam,        ~grade,
  "Rozz",        "666",           "midterm",    1.3,
  "Andrew",      "1969",          "midterm",    2.0,
  "Siouxsie",    "3.14",          "midterm",    1.7,
  "Rozz",        "666",           "final",      2.3,
  "Andrew",      "1969",          "final",      1.7,
  "Siouxsie",    "3.14",          "final",      1.0
)
exam_results_overloaded

## # A tibble: 6 x 4
##   student  stu_number exam    grade
##   <chr>     <chr>     <chr>    <dbl>
## 1 Rozz      666       midterm  1.3
## 2 Andrew    1969      midterm  2
## 3 Siouxsie 3.14      midterm  1.7
## 4 Rozz      666       final   2.3
## 5 Andrew    1969      final   1.7
## 6 Siouxsie 3.14      final   1
```

This table is not tidy in an intuitive sense because it includes redundancy. Why list the student numbers twice, once with each observation of exam score? The table is not tidy in the technical sense that not every observational unit forms a table, i.e., the observation of student numbers and the observation of exam scores should be stored independently in different tables, like so:

```
# same as before
exam_results_tidy <- tribble(
  ~student,      ~exam,        ~grade,
  "Rozz",       "midterm",   1.3,
  "Andrew",     "midterm",   2.0,
  "Siouxsie",   "midterm",   1.7,
  "Rozz",       "final",     2.3,
  "Andrew",     "final",     1.7,
  "Siouxsie",   "final",     1.0
)
# additional table with student numbers
student_numbers <- tribble(
  ~student,      ~student_number,
  "Rozz",        "666",
  "Andrew",      "1969",
  "Siouxsie",    "3.14"
)
```

Notice that, although the information is distributed over two tibbles, it is linked by the common column `student`. If we really need to bring all of the information together, the tidyverse has a quick and elegant solution:

```
full_join(exam_results_tidy, student_numbers, by = "student")
```

```
## # A tibble: 6 x 4
##   student   exam   grade student_number
##   <chr>     <chr>  <dbl>   <chr>
## 1 Rozz     midterm  1.3    666
## 2 Andrew    midterm  2      1969
## 3 Siouxsie midterm  1.7    3.14
## 4 Rozz     final   2.3    666
## 5 Andrew    final   1.7    1969
## 6 Siouxsie final   1      3.14
```

4.3. Data manipulation: the basics

4.3.1. Pivoting

The tidyverse strongly encourages the use of tidy data, or at least almost tidy data. If your data is (almost) tidy, you can be reasonably sure that you can plot and analyze the data without additional wrangling. If your data is not (almost) tidy because it is too wide or too long (see below), what is required is a joyful round of pivoting. There are two directions of pivoting: making data longer, and making data wider.

4. Data Wrangling

4.3.1.1. Making too wide data longer with pivot_longer

Consider the previous example of messy data again:

```
exam_results_visual <- tribble(
  ~exam,          ~"Rozz",    ~"Andrew",     ~"Siouxsie",
  "midterm",      "1.3",      "2.0",        "1.7",
  "final" ,       "2.3",      "1.7",        "1.0"
)
exam_results_visual

## # A tibble: 2 x 4
##   exam    Rozz  Andrew Siouxsie
##   <chr>  <chr> <chr>  <chr>
## 1 midterm 1.3   2.0    1.7
## 2 final   2.3   1.7    1.0
```

This data is “too wide”. We can make it longer with the function `pivot_longer` from the `tidyverse` package. Check out the example below before we plunge into a description of `pivot_longer`.

```
exam_results_visual %>%
  pivot_longer(
    # pivot every column except the first
    cols = - 1,
    # name of new column which contains the
    # names of the columns to be "gathered"
    names_to = "student",
    # name of new column which contains the values
    # of the cells which now form a new column
    values_to = "grade"
  ) %>%
  # optional reordering of columns (to make
  # the output exactly like `exam_results_tidy`)
  select(student, exam, grade)

## # A tibble: 6 x 3
##   student  exam    grade
##   <chr>    <chr>  <chr>
## 1 Rozz     midterm 1.3
## 2 Andrew   midterm 2.0
## 3 Siouxsie midterm 1.7
## 4 Rozz     final   2.3
## 5 Andrew   final   1.7
## 6 Siouxsie final   1.0
```

What `pivot_longer` does, in general, is take a bunch of columns and gather the values of all cells in these columns into a single, new column, the so-called *value column*, i.e., the column with the values of the cells to be gathered. If `pivot_longer` stopped here, we would lose information about which cell values belonged to which original column. Therefore, `pivot_longer` also creates a second new column, the so-called *name column*, i.e., the column with the names of the original columns that we gathered together. Consequently, in order to do its job, `pivot_longer` minimally needs three pieces of information:²

1. which columns to spin around (function argument `cols`)
2. the name of the to-be-created new value column (function argument `values_to`)
3. the name of the to-be-created new name column (function argument `names_to`)

For different ways of selecting columns to pivot around, see Section 4.3.3 below.

4.3.1.2. Making too long data wider with `pivot_wider`

Consider the following example of data which is untidy because it is too long:

```
mixed_results_too_long <-  
  tibble(student = rep(c('Rozz', 'Andrew', 'Siouxsie'), times = 2),  
         what = rep(c('grade', 'participation'), each = 3),  
         howmuch = c(2.7, 2.0, 1.0, 75, 93, 33))  
mixed_results_too_long  
  
## # A tibble: 6 x 3  
##   student   what       howmuch  
##   <chr>     <chr>      <dbl>  
## 1 Rozz      grade       2.7  
## 2 Andrew    grade       2  
## 3 Siouxsie grade       1  
## 4 Rozz      participation 75  
## 5 Andrew    participation 93  
## 6 Siouxsie participation 33
```

This data is untidy because it lumps two types of different measurements (a course grade, and the percentage of participation) in a single column. These are different variables, and so should be represented in different columns.

To fix a data representation that is too long, we can make it wider with the help of the `pivot_wider` function from the `tidyverse` package. We look at an example before looking at the general behavior of the `pivot_wider` function.

²There are alternative possibilities for specifying names of the value and name column, which allow for more dynamic construction of strings. We will not cover all of these details here, but we will use some of these alternative specifications in subsequent examples.

4. Data Wrangling

```
mixed_results_too_long %>%
  pivot_wider(
    # column containing the names of the new columns
    names_from = what,
    # column containing the values of the new columns
    values_from = howmuch
  )

## # A tibble: 3 x 3
##   student  grade participation
##   <chr>     <dbl>        <dbl>
## 1 Rozz       2.7          75
## 2 Andrew      2            93
## 3 Siouxsie   1            33
```

In general, `pivot_wider` picks out two columns, one column of values to distribute into new to-be-created columns, and one vector of names or groups which contains the information about the, well, names of the to-be-created new columns. There are more refined options for `pivot_wider` some of which we will encounter in the context of concrete cases of application.

4.3.2. Subsetting row & columns

If a data set contains too much information for your current purposes, you can discard irrelevant (or unhelpful) rows and columns. The function `filter` takes a Boolean expression and returns only those rows of which the Boolean expression is true:

```
exam_results_tidy %>%
  # keep only entries with grades better than 1.7
  filter(grade <= 1.7)

## # A tibble: 4 x 3
##   student  exam    grade
##   <chr>    <chr>   <dbl>
## 1 Rozz     midterm  1.3
## 2 Siouxsie midterm  1.7
## 3 Andrew    final   1.7
## 4 Siouxsie final   1
```

To select rows by an index or a vector of indeces, use the `slice` function:

```
exam_results_tidy %>%
  # keep only entries from rows with an even index
  dplyr::slice(c(2,4,6)) # explicit call to avoid name clash (package 'greta')

## # A tibble: 3 x 3
##   student    exam    grade
##   <chr>     <chr>    <dbl>
## 1 Andrew    midterm     2
## 2 Rozz      final      2.3
## 3 Siouxsie final      1
```

The function `select` allows to pick out a subset of columns. Interestingly, it can also be used to reorder columns, because the order in which column names are specified matches the order in the returned tibble.

```
exam_results_tidy %>%
  # select columns `grade` and `name`
  select(grade, exam)

## # A tibble: 6 x 2
##   grade exam
##   <dbl> <chr>
## 1 1.3   midterm
## 2 2     midterm
## 3 1.7   midterm
## 4 2.3   final
## 5 1.7   final
## 6 1     final
```

4.3.3. Tidy selection of column names

To select the columns in several functions within the tidyverse, such as `pivot_longer` or `select`, there are useful helper functions from the `tidyselect` package. Here are some examples:³

```
# bogus code for illustration of possibilities!
SOME_DATA %>%
  select( ... # could be one of the following
         # all columns indexed 2, 3, ..., 10
         2:10
         # all columns except the one called "COLNAME"
         - COLNAME)
```

³The helpers from the `tidyselect` package also accept regular expressions.

4. Data Wrangling

```
# all columns with names starting with "STRING"
... starts_with("STRING")
# all columns with names ending with "STRING"
... ends_with("STRING")
# all columns with names containing "STRING"
... contains("STRING")
# all columns with names of the form "Col_i" with i = 1, ..., 10
... num_range("Col_", 1:10)
)
```

4.3.4. Adding, changing and renaming columns

To add a new column, or to change an existing one use function `mutate`, like so:

```
exam_results_tidy %>%
  mutate(
    # add a new column called 'passed' depending on grade
    # [NB: severe passing conditions in this class!!]
    passed = grade <= 1.7,
    # change an existing column; here: change
    # character column 'exam' to ordered factor
    exam = factor(exam, ordered = T)
  )

## # A tibble: 6 x 4
##   student  exam    grade passed
##   <chr>    <ord>   <dbl> <lgl>
## 1 Rozz     midterm  1.3  TRUE
## 2 Andrew    midterm  2    FALSE
## 3 Siouxsie midterm  1.7  TRUE
## 4 Rozz     final    2.3  FALSE
## 5 Andrew    final    1.7  TRUE
## 6 Siouxsie final    1    TRUE
```

If you want to rename a column, function `rename` is what you want:

```
exam_results_tidy %>%
  # rename existing column "student" to new name "participant"
  # [NB: rename takes the new name first]
  rename(participant = student)
```

```
## # A tibble: 6 x 3
##   participant exam     grade
##   <chr>        <chr>    <dbl>
## 1 Rozz         midterm  1.3
## 2 Andrew       midterm  2
## 3 Siouxsie    midterm  1.7
## 4 Rozz         final    2.3
## 5 Andrew       final    1.7
## 6 Siouxsie    final    1
```

4.3.5. Splitting and uniting columns

Here is data from course homework:

```
homework_results_untidy <-
  tribble(
    ~student,      ~results,
    "Rozz",        "1.0,2.3,3.0",
    "Andrew",      "2.3,2.7,1.3",
    "Siouxsie",    "1.7,4.0,1.0"
  )
```

This is not a useful representation format. Results of three homework sets are mashed together in a single column. Each value is separated by a comma, but it's all stored as a character vector.

To disentangle information in a single column, use the `separate` function:

```
homework_results_untidy %>%
  separate(
    # which column to split up
    col = results,
    # names of the new column to store results
    into = str_c("HW_", 1:3),
    # separate by which character / reg-exp
    sep = ",",
    # automatically (smart-)convert the type of the new cols
    convert = T
  )

## # A tibble: 3 x 4
##   student   HW_1   HW_2   HW_3
##   <chr>     <dbl>  <dbl>  <dbl>
## 1 Rozz       1     2.3     3
```

4. Data Wrangling

```
## 2 Andrew      2.3  2.7  1.3
## 3 Siouxsie   1.7   4     1
```

If you have reason to perform the reverse operation, i.e., join together several columns, use the `unite` function.

4.3.6. Sorting a data set

If you want to indicate a fixed order of the reoccurring elements in a (character) vector, e.g., for plotting in a particular order, you should make this column an ordered factor. But if you want to order a data set along a column, e.g., for inspection or printing as a table, then you can do that using the `arrange` function. You can specify several columns to sort alpha-numerically in ascending order, and also indicate a descending order using the `desc` function:

```
exam_results_tidy %>%
  arrange(desc(student), grade)
```

```
## # A tibble: 6 x 3
##   student    exam    grade
##   <chr>     <chr>   <dbl>
## 1 Siouxsie final     1
## 2 Siouxsie midterm  1.7
## 3 Rozz       midterm  1.3
## 4 Rozz       final    2.3
## 5 Andrew     final    1.7
## 6 Andrew     midterm  2
```

4.3.7. Combining tibbles

There are frequently occasions on which data from two separate variables needs to be combined. The simplest case is where two entirely disjoint data sets merely need to be glued together, either horizontally (binding columns together with function `cbind`) or vertically (binding rows together with function `rbind`).

```
new_exam_results_tidy <- tribble(
  ~student,     ~exam,      ~grade,
  "Rozz",       "bonus",   1.7,
  "Andrew",     "bonus",   2.3,
  "Siouxsie",   "bonus",  1.0
)
rbind(
  exam_results_tidy,
  new_exam_results_tidy
)
```

```
## # A tibble: 9 x 3
##   student    exam    grade
##   <chr>     <chr>    <dbl>
## 1 Rozz      midterm   1.3
## 2 Andrew    midterm   2
## 3 Siouxsie midterm   1.7
## 4 Rozz      final    2.3
## 5 Andrew    final    1.7
## 6 Siouxsie final    1
## 7 Rozz      bonus    1.7
## 8 Andrew    bonus    2.3
## 9 Siouxsie bonus    1
```

If two data sets have information in common, and the combination should respect that commonality, the `join` family of functions is of great help. Consider the case of distributed information again that we looked at to understand the third constraint of the concept of “tidy data”. There are two tibbles, both of which contain information about the same students. They share the column `student` (this does not necessarily have to be in the same order!) and we might want to join the information from both sources into a single (messy but almost tidy) representation, using `full_join`. We have seen an example already, which is repeated here:

```
# same as before
exam_results_tidy <- tribble(
  ~student,     ~exam,       ~grade,
  "Rozz",       "midterm",   1.3,
  "Andrew",     "midterm",   2.0,
  "Siouxsie",   "midterm",   1.7,
  "Rozz",       "final",    2.3,
  "Andrew",     "final",    1.7,
  "Siouxsie",   "final",    1.0
)
# additional table with student numbers
student_numbers <- tribble(
  ~student,     ~student_number,
  "Rozz",       "666",
  "Andrew",     "1969",
  "Siouxsie",   "3.14"
)
full_join(exam_results_tidy, student_numbers, by = "student")

## # A tibble: 6 x 4
##   student    exam    grade student_number
##   <chr>     <chr>    <dbl>    <chr>
```

4. Data Wrangling

```
## 1 Rozz      midterm   1.3 666
## 2 Andrew    midterm   2   1969
## 3 Siouxsie midterm   1.7 3.14
## 4 Rozz      final     2.3 666
## 5 Andrew    final     1.7 1969
## 6 Siouxsie final     1   3.14
```

If two data sets are to be joined by a column that is not exactly shared by both sets (one contains entries in this columns that the other doesn't) then a `full_join` will retain all information from both. If that is not what you want, check out alternative functions like `right_join`, `semi_join` etc. using the data wrangling cheat sheet.

4.4. Grouped operations

A frequently occurring problem in data analysis is to obtain a summary statistic (see Section 5) for different subsets of data. For example, we might want to calculate the average score for each student in our class. We could do that by filtering like so (notice that `pull` gives you the column vector specified):

```
# extracting mean grade for Rozz
mean_grade_Rozz <- exam_results_tidy %>%
  filter(student == "Rozz") %>% pull(grade) %>% mean
mean_grade_Rozz

## [1] 1.8
```

But then we need to do that two more times, so we shouldn't copy-paste code, so we write a function and use `mutate` to add a mean for each student:

```
get_mean_for_student = function(student_name) {
  exam_results_tidy %>%
    filter(student == student_name) %>% pull(grade) %>% mean
}

map_dbl(
  exam_results_tidy %>% pull(student) %>% unique,
  get_mean_for_student
)

## [1] 1.80 1.85 1.35
```

Also not quite satisfactory, clumsy and error-prone. Enter, grouping in the tidyverse. If we want to apply a particular operation to all combinations of levels of different variables (no matter whether they are encoded as factors or not when we group), we can do this with the function `group_by`, followed by either a call to `mutate` or `summarise`. Check this example:

```
exam_results_tidy %>%
  group_by(student) %>%
  summarise(
    student_mean = mean(grade)
  )

## # A tibble: 3 x 2
##   student  student_mean
##   <chr>        <dbl>
## 1 Andrew      1.85
## 2 Rozz        1.8
## 3 Siouxsie   1.35
```

The function `summarise` returns a single row for each combination of levels of grouping variables. If we use the function `mutate` instead, the summary statistic is added (repeatedly) in each of the original rows:

```
exam_results_tidy %>%
  group_by(student) %>%
  mutate(
    student_mean = mean(grade)
  )

## # A tibble: 6 x 4
## # Groups:   student [3]
##   student  exam     grade student_mean
##   <chr>    <chr>   <dbl>        <dbl>
## 1 Rozz     midterm  1.3        1.8
## 2 Andrew   midterm  2          1.85
## 3 Siouxsie midterm  1.7        1.35
## 4 Rozz     final    2.3        1.8
## 5 Andrew   final    1.7        1.85
## 6 Siouxsie final    1          1.35
```

The latter can sometimes be handy, for example when overlaying a plot of the data with grouped means, for instance.

It may be important to remember that after a call of `group_by`, the resulting tibbles retains the grouping information for *all* subsequent operations. To remove grouping information, use the function `ungroup`.

4. Data Wrangling

4.5. Case study: the King of France

Let's go through one case study of data preprocessing. We look at the example introduced and fully worked out in Appendix D.4. (Please read Section D.4.1 to find out more about where this data set is coming from.)

The raw data set is stored in the GitHub repository that also hosts this web-book. It can be loaded using:

```
data_KoF_raw <- read_csv(url('https://raw.githubusercontent.com/michael-franke/introduction-to-data-science/master/datasets/KoF.csv'))
```

We can then get a glimpse at the data using:

```
glimpse(data_KoF_raw)
```

```
## Observations: 2,813
## Variables: 16
## $ submission_id    <dbl> 192, 192, 192, 192, 192, 192, 192, 192, 19...
## $ RT                <dbl> 8110, 35557, 3647, 16037, 11816, 6024, 4986, 13019, ...
## $ age               <dbl> 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, ...
## $ comments          <chr> NA, ...
## $ item_version      <chr> "none", "none", "none", "none", "none", "none", ...
## $ correct_answer    <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, ...
## $ education         <chr> "Graduated College", "Graduated College", "Graduated...
## $ gender            <chr> "female", "female", "female", "female", "f...
## $ languages         <chr> "English", "English", "English", "English", "English...
## $ question          <chr> "World War II was a global war that lasted from 1914...
## $ response          <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, ...
## $ timeSpent         <dbl> 39.48995, 39.48995, 39.48995, 39.48995, 39.48995, 39...
## $ trial_name        <chr> "practice_trials", "practice_trials", "practice_tria...
## $ trial_number      <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ trial_type        <chr> "practice", "practice", "practice", "practice", "pra...
## $ vignette          <chr> "undefined", "undefined", "undefined", "undefined", ...
```

The variables in this data set are:

- `submission_id`: unique identifier for each participant
- `RT`: the reaction time for each decision
- `age`: the (self-reported) age of the participant
- `comments`: the (optional) comments each participant may have given
- `item_version`: the condition which the test sentence belongs to (only given for trials of type `main` and `special`)
- `correct_answer`: for trials of type `filler` and `special` what the true answer should have been
- `education`: the (self-reported) education level with options `Graduated College`, `Graduated High School`, `Higher Degree`

- gender: (self-reported) gender
- languages: (self-reported) native languages
- question: the sentence to be judged true or false
- response: the answer ("TRUE" or "FALSE") on each trial
- trial_name: whether the trial is a main or practice trials (levels main_trials and practice_trials)
- trial_number: consecutive numbering of each participant's trial
- trial_type: whether the trial was of the category filler, main, practice or special, where the latter encodes the "background checks"
- vignette: the current item's vignette number (applies only to trials of type main and special)

Let's have brief look at the comments (sometimes helpful, usually entertaining) and the self-reported native languages:

```
data_KoF_raw %>% pull(comments) %>% unique
```

```
## [1] NA
## [2] "I hope I was right most of the time!"
## [3] "My level of education is Some Highschool, not finished. So I couldn't input what was
## [4] "It was interesting, and made re-read questions to make sure they weren't tricks. I ha
## [5] "Worked well"
## [6] "A surprisingly tricky study! Thoroughly enjoyed completing it, despite several red he
## [7] "n/a"
## [8] "Thank you for the opportunity."
## [9] "this was challenging"
## [10] "I'm not good at learning history so i might of made couple of mistakes. I hope I did
## [11] "Interesting survey - thanks!"
## [12] "no"
## [13] "Regarding the practice question - I'm aware that Alexander Bell invented the telephon
## [14] "Fun study!"
## [15] "Fun stuff"
```

```
data_KoF_raw %>% pull(languages) %>% unique
```

```
## [1] "English"           "english"          "English, Italian"
## [4] "English/ ASL"      "English and Polish" "Chinese"
## [7] "English, Mandarin" "Polish"            "Turkish"
## [10] NA                 "English, Sarcasm"   "English, Portuguese"
```

We might wish to exclude people who do not include "English" as one of their native languages in some studies. Here, we do not since we also have strong, more specific filters on comprehension (see below). Since we are not going to use this information later on, we might as well discard it now:

4. Data Wrangling

```
data_KoF_raw <- data_KoF_raw %>%
  select(-languages, -comments, -age, -RT, -education, -gender)
```

But even after pruning irrelevant columns, this data set is still not ideal. We need to preprocess it more thoroughly to make it more intuitively manageable. For example, the information in column `trial_name` does not give the trial's name in an intuitive sense, but its type: whether it is a practice or a main trial. But this information, and more, is also represented in the column `trial_type`. The column `item_version` contains information about the experimental condition. To see this (mess) the code below prints the selected information from the main trials of only one participant in an order that makes it easier to see what is what.

```
data_KoF_raw %>%
  # ignore practice trials for the moment
  # focus on one participant only
  filter(trial_type != "practice", submission_id == 192) %>%
  select(trial_type, item_version, question) %>%
  arrange(desc(trial_type), item_version) %>%
  print(n = Inf)

## # A tibble: 24 x 3
##   trial_type item_version question
##   <chr>     <chr>       <chr>
## 1 special    none        The Pope is currently not married.
## 2 special    none        Germany has volcanoes.
## 3 special    none        France has a king.
## 4 special    none        Canada is a democracy.
## 5 special    none        Belgium has rainforests.
## 6 main       0          The volcanoes of Germany dominate the landscape.
## 7 main       1          Canada has an emperor, and he is fond of sushi.
## 8 main       10         Donald Trump, his favorite nature spot is not the Be-
## 9 main       6          The King of France isn't bald.
## 10 main      9          The Pope's wife, she did not invite Angela Merkel fo-
## 11 filler    none        The Solar System includes the planet Earth.
## 12 filler    none        Vatican City is the world's largest country by land ~
## 13 filler    none        Big Ben is a very large building in the middle of Pa-
## 14 filler    none        Harry Potter is a series of fantasy novels written b-
## 15 filler    none        Taj Mahal is a mausoleum on the bank of the river in-
## 16 filler    none        James Bond is a spanish dancer from Madrid.
## 17 filler    none        The Pacific Ocean is a large ocean between Japan and-
## 18 filler    none        Australia has a very large border with Brazil.
## 19 filler    none        Steve Jobs was an American inventor and co-founder o-
## 20 filler    none        Planet Earth is part of the galaxy 'Milky Way'.
```

```

## 21 filler      none      Germany shares borders with France, Belgium and Denm-
## 22 filler      none      Antarctica is a continent covered almost completely ~
## 23 filler      none      The Statue of Liberty is a colossal sculpture on Lib-
## 24 filler      none      English is the main language in Australia, Britain a-

```

We see that the information in `item_version` specifies the critical condition. To make this more intuitively manageable, we would like to have a column called `condition` and it should, ideally, also contain useful information for the cases where `trial_type` is not `main` or `special`. That is why we will therefore remove the column `trial_name` completely, and create an informative column `condition` in which we learn of every row whether it belongs to one of the 5 experimental conditions, and if not whether it is a filler or a “background check” (= `special`) trial.

```

data_KoF_processed <- data_KoF_raw %>%
  # drop redundant information in column `trial_name`
  select(-trial_name) %>%
  # discard practice trials
  filter(trial_type != "practice") %>%
  mutate(
    # add a 'condition' variable
    condition = case_when(
      trial_type == "special" ~ "background check",
      trial_type == "main" ~ str_c("Condition ", item_version),
      TRUE ~ "filler"
    ) %>%
    # make the new 'condition' variable a factor
    factor(
      ordered = T,
      levels = c(
        str_c("Condition ", c(0, 1, 6, 9, 10)),
        "background check", "filler"
      )
    )
  )
# write_csv(data_KoF_processed, "data_sets/king-of-france_data_processed.csv")

```

4.5.1. Cleaning the data

We clean the data in two consecutive steps:

1. Remove all data from any participant who got more than 50% of the answer to filler material wrong.
2. Remove individual main trials if the corresponding “background check” question was answered wrongly.

4. Data Wrangling

4.5.1.1. Cleaning by-participant

```
# look at error rates for filler sentences by subject
# mark every subject as an outlier when they
# have a proportion of correct responses of less than 0.5
subject_error_rate <- data_KoF_processed %>%
  filter(trial_type == "filler") %>%
  group_by(submission_id) %>%
  summarise(
    proportion_correct = mean(correct_answer == response),
    outlier_subject = proportion_correct < 0.5
  ) %>%
  arrange(proportion_correct)
```

Apply the cleaning step:

```
# add info about error rates and exclude outlier subject(s)
d_cleaned <-
  full_join(data_KoF_processed, subject_error_rate, by = "submission_id") %>%
  filter(outlier_subject == FALSE)
```

4.5.1.2. Cleaning by-trial

```
# exclude every critical trial whose 'background' test question was answered wrong
d_cleaned <-
  d_cleaned %>%
  # select only the 'background question' trials
  filter(trial_type == "special") %>%
  # is the background question answered correctly?
  mutate(
    background_correct = correct_answer == response
  ) %>%
  # select only the relevant columns
  select(submission_id, vignette, background_correct) %>%
  # right join lines to original data set
  right_join(d_cleaned, by = c("submission_id", "vignette")) %>%
  # remove all special trials, as well as main trials with incorrect background che
  filter(trial_type == "main" & background_correct == TRUE)

# write_csv(d_cleaned, "data_sets/king-of-france_data_cleaned.csv")
```

5. Summary statistics

A **summary statistics** is a single number which represents one aspect of a possibly much more complex chunk of data. This single number might, for example, indicate the maximum or minimum value of a vector of a billion observations. The large data set (1 billion observations) is reduced to a single number which represents one aspect of that data. Summary statistics are, as a general (but violable) rule, many-to-one (surjections). They compress complex information into a simpler, compressed representation.

Summary statistics are useful for understanding the data at hand, for communication about a data set, but also for subsequent statistical analyses. As we will see later on, many statistical tests look at a summary statistic x , which is a single value derived from data set D , and compare x to an expectation of what x should be like if the process that generated D really had a particular property. For the moment, however, we use summary statistics only to get comfortable with data: understanding it better, and gaining competence to manipulate it.

Chapter 5.1 first uses the Bio-Logic Jazz-Metal data set to look at a very intuitive class of summary statistics for categorical data, namely counts and proportions. Chapter 5.2 introduces summary statistics for simple, one dimensional vectors with numeric information. In doing so we will also learn about *nested tibbles*. Chapter 5.3 looks at measures of relation between two numerical vectors, namely *covariance* and *correlation*. These last two chapters use the avocado data set.

The learning goals for this chapter are:

- become able to compute counts and frequencies for categorical data
- understand and be able to compute summary statistics for one-dimensional metric data:
- measures of central tendency
- mean/mode/median/
- measures of dispersion
- variance, standard deviation, quantiles
- non-parametric estimates of confidence
- bootstrapped CI of mean
- understand and be able to compute for two-dimensional metric data:
- covariance
- Bravais-Pearson correlation

5. Summary statistics

5.1. Counts and proportions

Very familiar instances of summary statistics are counts and frequencies. While there is no conceptual difficulty in understanding these numerical measures, we have yet to see how to obtain counts for categorical data in R. The Bio-Logic Jazz-Metal data set provides nice material for doing so. If unfamiliar with the data and the experiment that generated it, please have a look at Appendix Chapter D.5.

5.1.1. Loading and inspecting the data

We load the preprocessed data immediately (Appendix Chapter D.5 for details how this preprocessing was performed).

```
data_BLJM_processed <- read_csv(url('https://raw.githubusercontent.com/michael-fran...'))
```

The preprocessed data lists, for each participant (in column `submission_id`) the binary choice (in column `response`) given for a particular condition (in column `condition`).

```
head(data_BLJM_processed)

## # A tibble: 6 x 3
##   submission_id condition response
##       <dbl>     <chr>    <chr>
## 1         379     BM      Beach
## 2         379     LB      Logic
## 3         379     JM      Metal
## 4         378     JM      Metal
## 5         378     LB      Logic
## 6         378     BM      Beach
```

5.1.2. Obtaining counts with `n`, `count` and `tally`

To obtain counts, the `dplyr` package offers functions `n`, `count` and `tally`, among others.¹ The function `n` does not take arguments and is useful for counting rows. It works inside of `summary` and `mutate` and is usually applied to grouped data sets. For example, we can get a count of how many observations the data in `data_BLJM_processed` contains for each condition by first grouping by variable `condition` and then calling `n` (without arguments) inside of `summarise`:

```
data_BLJM_processed %>%
  group_by(condition) %>%
  summarise(nr_observation_per_condition = n()) %>%
  ungroup()
```

¹Useful base R functions for obtaining counts are `table` and `prop.table`.

```
## # A tibble: 3 x 2
##   condition nr_observation_per_condition
##   <chr>                <int>
## 1 BM                  102
## 2 JM                  102
## 3 LB                  102
```

Notice that calling `n` without grouping just gives you the number of rows in the data set:

```
data_BLJM_processed %>% summarize(n_rows = n())

## # A tibble: 1 x 1
##   n_rows
##   <int>
## 1     306
```

This can also be obtained simply by (although in a different output format!):

```
nrow(data_BLJM_processed)

## [1] 306
```

Counting can be helpful also when getting acquainted with a data set, or when checking whether the data is complete. For example, we can verify that every participant in the experiment contributed three data points like so:

```
data_BLJM_processed %>%
  group_by(submission_id) %>%
  summarise(nr_data_points = n())

## # A tibble: 102 x 2
##   submission_id nr_data_points
##   <dbl>            <int>
## 1 278                 3
## 2 279                 3
## 3 280                 3
## 4 281                 3
## 5 282                 3
## 6 283                 3
## 7 284                 3
## 8 285                 3
## 9 286                 3
## 10 287                3
## # ... with 92 more rows
```

5. Summary statistics

The functions `tally` and `count` are essentially just convenience wrappers around `n`. While `tally` expects that the data is already grouped in the relevant way, `count` takes a column specification as an argument and does the grouping (and final ungrouping) implicitly.

For instance, the following code blocks produce the same output, one using `n`, the other using `count`, namely the total number of times a particular response has been given in a particular condition:

```
data_BLJM_processed %>%
  group_by(condition, response) %>%
  summarise(n = n())

## # A tibble: 6 x 3
##   condition response     n
##   <chr>     <chr>    <int>
## 1 BM        Beach      44
## 2 BM        Mountains  58
## 3 JM        Jazz       64
## 4 JM        Metal      38
## 5 LB        Biology    58
## 6 LB        Logic      44

data_BLJM_processed %>%
  # function `count` is masked by another package, must call explicitly
  dplyr::count(condition, response)

## # A tibble: 6 x 3
##   condition response     n
##   <chr>     <chr>    <int>
## 1 BM        Beach      44
## 2 BM        Mountains  58
## 3 JM        Jazz       64
## 4 JM        Metal      38
## 5 LB        Biology    58
## 6 LB        Logic      44
```

So, these counts suggest that there is an overall preference for mountains over beaches, Jazz over Metal and Biology over Logic. Who would have known!?

These counts are overall numbers. They do not tell us anything about any potentially interesting relationship between preferences. So, let's have a closer look at the number of people who selected which music-subject pair. We collect these counts in variable `BLJM_associated_counts`. We first need to pivot the data, using `pivot_wider`, to make sure each participant's choices are associated with each other, and then take the counts of interest:

```

BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts

## # A tibble: 4 x 3
##   JM     LB       n
##   <chr>  <chr>   <int>
## 1 Jazz   Biology    38
## 2 Jazz   Logic      26
## 3 Metal  Biology    20
## 4 Metal  Logic      18

```

We can also produce a table of proportions from this, simply by dividing the column called `n` by the total number of observations, i.e., by `sum(n)`. We can also flip the table around into a more convenient (though messy) representation:

```

BLJM_associated_counts %>%
  # look at relative frequency, not total counts
  mutate(n = n / sum(n)) %>%
  pivot_wider(names_from = LB, values_from = n)

## # A tibble: 2 x 3
##   JM     Biology Logic
##   <chr>   <dbl> <dbl>
## 1 Jazz     0.373 0.255
## 2 Metal    0.196 0.176

```

Eye-ball this table of relative frequencies, we might indeed hypothesis that preference for musical style are not independent of preference for academic subject. The impression is corroborated by looking at the plot in Figure 5.1. More on this later!

5.2. Central tendency and dispersion

This section will look at two types of summary statistics: measures of central tendency and measures of dispersion.

Measures of central tendency map a vector of observations onto a single number that represents, roughly put, “the center”. Since what counts as a “center” is ambiguous, there are several measures of central

5. Summary statistics

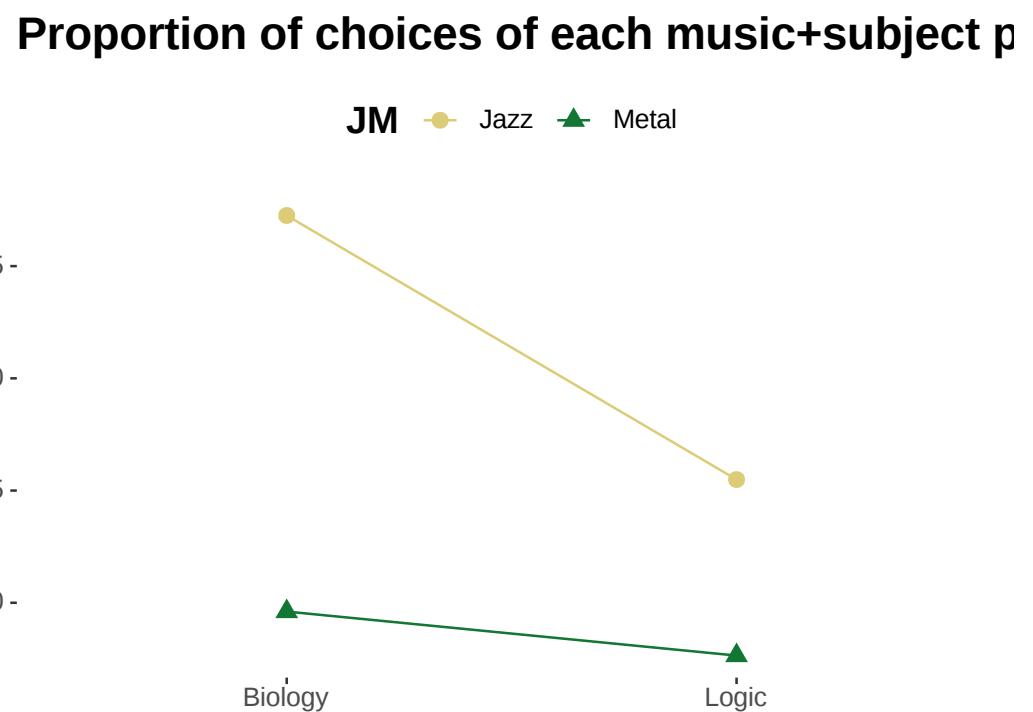


Figure 5.1.: Proportions of jointly choosing a musical style and an academic subfield in the Bio-Logic Jazz-Metal data set.

tendencies. Different measures of central tendencies can be more or less adequate for one purpose or another. The type of variable (nominal, ordinal or metric, for instance) will also influence the choice of measure. We will visit three prominent measures of central tendency here: *(arithmetic) mean, median and mode*.

Measures of dispersion indicate how much the observations are spread out around, let's say, "a center". We will visit three prominent measures of dispersion: the *variance*, the *standard deviation* and *quantiles*.

To illustrate these ideas, consider the case of a numeric vector of observations. Central tendency and dispersion together describe a (numeric) vector by giving indicative information about the point around which the observations spread, and how far away from that middle point they tend to lie. Fictitious examples of observation vectors with higher or lower central tendency and higher or lower dispersion are given in Figure 5.2.

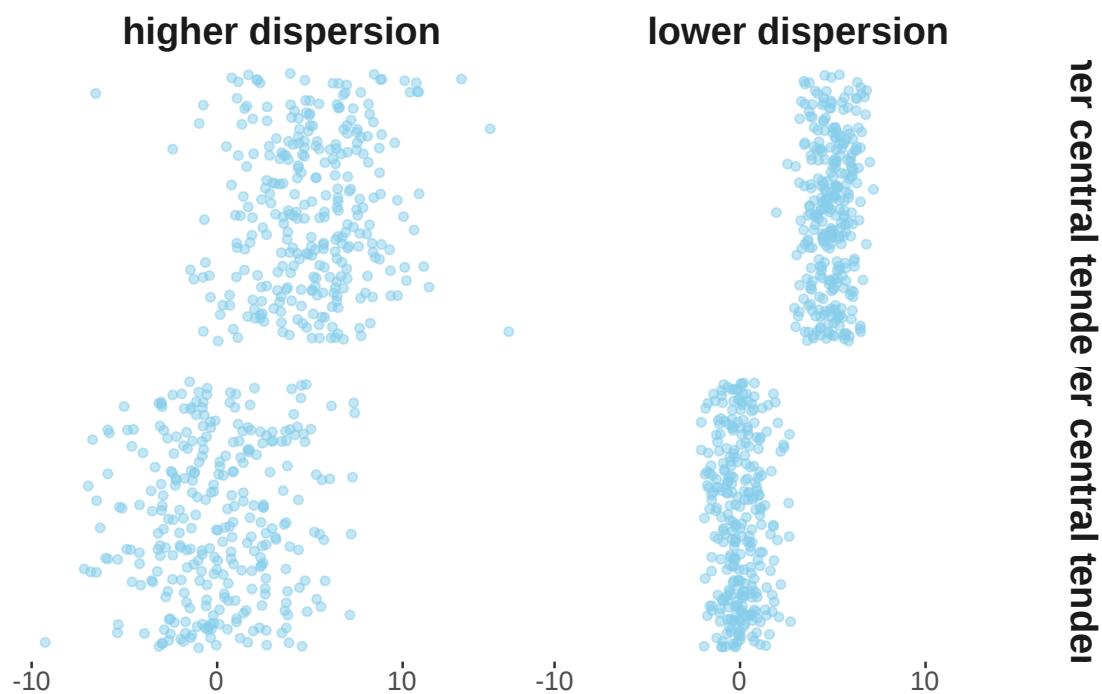


Figure 5.2.: Fictitious data points with higher/lower central tendencies and higher/lower dispersion. NB: points are 'jittered' along the vertical dimension for better visibility; only the horizontal dimension is relevant here.

5.2.1. The data for the remainder of the chapter

In the remainder of this chapter we will use the avocado data set, a very simple and theory-free example in which we can explore two metric variables: the average price at which avocados were sold during specific

5. Summary statistics

intervals of time and the total amount of avocados sold. Please check Appendix Chapter D.6 for more information on this data set.

We load the data into a variable named `avocado_data` but also immediately rename some of the columns to have more convenient handles (see also Appendix Chapter D.6):

```
avocado_data <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-to-r/gh-pages/data/avocado.csv')) %>%  
  # remove currently irrelevant columns  
  select(-X1, -contains("Bags"), -year, -region) %>%  
  # rename variables of interest for convenience  
  rename(  
    total_volume_sold = `Total Volume`,  
    average_price = `AveragePrice`,  
    small = '4046',  
    medium = '4225',  
    large = '4770',  
  )
```

We can then take a glimpse:

```
avocado_data
```

```
## # A tibble: 18,249 x 7  
##   Date      average_price total_volume_sold small  medium large type  
##   <date>        <dbl>            <dbl>     <dbl> <dbl> <dbl> <chr>  
## 1 2015-12-27       1.33           64237.  1037. 54455. 48.2 conventional  
## 2 2015-12-20       1.35           54877.  674.  44639. 58.3 conventional  
## 3 2015-12-13       0.93          118220.  795. 109150. 130. conventional  
## 4 2015-12-06       1.08          78992.  1132. 71976. 72.6 conventional  
## 5 2015-11-29       1.28          51040.  941.  43838. 75.8 conventional  
## 6 2015-11-22       1.26          55980.  1184. 48068. 43.6 conventional  
## 7 2015-11-15       0.99          83454. 1369. 73673. 93.3 conventional  
## 8 2015-11-08       0.98          109428. 704. 101815. 80 conventional  
## 9 2015-11-01       1.02          99811. 1022. 87316. 85.3 conventional  
## 10 2015-10-25      1.07          74339.  842.  64757. 113 conventional  
## # ... with 18,239 more rows
```

The columns that will interest us the most in this chapter are:

- `average_price` - average price of a single avocado
- `total_volume_sold` - Total number of avocados sold
- `type` - whether the price/amount is for a conventional or an organic avocado

In particular, we will look at summary statistics for the `average_price` and `total_volume_sold`, either for the whole data set or independently for each type of avocado. Notice that both of these variables are numeric. They are vectors of numbers, each representing an observation.

5.2.2. Measures of central tendency

5.2.2.1. The (arithmetic) mean

If $\vec{x} = \langle x_1, \dots, x_n \rangle$ is a vector of n observations with $x_i \in \mathbb{R}$ for all $1 \leq i \leq n$, the (arithmetic) **mean** of x , written $\mu_{\vec{x}}$, is defined as

$$\mu_{\vec{x}} = \frac{1}{n} \sum_{i=1}^n x_i.$$

The arithmetic mean can be understood intuitively as **the center of gravity**. If we place a marble on a wooden board for every x_i such that every marble is equally heavy and the differences between all data measurements are identical to the distances between the marbles, the arithmetic mean is where you can balance the board with the tip of your finger.

Example. The mean of the vector $\vec{x} = \langle 0, 3, 6, 7 \rangle$ is $\mu_{\vec{x}} = \frac{0+3+6+7}{4} = \frac{16}{4} = 4$. The black dots in the graph below show the data observations and the red cross indicates the mean. Notice that the mean is clearly *not* the mid-point between the maximum and the minimum (which here would be 3.5).



To calculate the mean of a large vector, R has a built-in function `mean`, which we have in fact used frequently before. Let's use it to calculate the mean of the variable `average_price` for different types of avocados:

```
avocado_data %>%
  group_by(type) %>%
  summarise(
    mean_price = mean(average_price)
  )

## # A tibble: 2 x 2
##   type      mean_price
##   <chr>        <dbl>
## 1 conventional     1.16
## 2 organic          1.65
```

5. Summary statistics

Unsurprisingly, the overall mean of the observed prices is (numerically) higher for organic avocados than for conventional ones.

Excursion. It is also possible to conceptualize the arithmetic mean as the **expected value** when sampling from the observed data. This is useful for linking the mean of a data sample to the expected value of a random variable, a concept we will introduce in Chapter 7. Suppose you have gathered the data $\vec{x} = \langle 0, 3, 6, 7 \rangle$. What is the expected value that you think you will obtain if you sample from this data vector once? – Wait! What does that even mean? Expected value? Sampling once?

Suppose that some joker from around town invites you for a game. The game goes like this. The joker puts a ball in an urn, one for each data observation. The joker writes the observed value on the ball corresponding to that value. You pay the joker a certain amount of money to be allowed to draw one ball from the urn. The balls are indistinguishable and the process of drawing is entirely fair. You receive the number corresponding to the ball you drew paid out in silver coins. (For simplicity, we assume that all numbers are non-negative, but that is not crucial. If a negative number is drawn, you just have to pay the joker that amount.)

How many silver coins would you maximally pay to play one round? Well, of course, no more than four (unless you value gaming on top of silver)! This is because 4 is the expected value of drawing once. This, in turn, is because every ball has a chance of 0.25 of being drawn. So you can expect to earn 0 silver with a 25% chance, 3 with a 25% chance, 6 with a 25% chance and 7 with a 25% chance. In this sense, the mean is the expected value of sampling once from the observed data.

5.2.2.2. The median

If $\vec{x} = \langle x_1, \dots, x_n \rangle$ is a vector of n data observations from an at least ordinal measure and if \vec{x} is ordered such that for all $1 \leq i < n$ we have $x_i \leq x_{i+1}$, the **median** is the value x_i such that the number of data observations that are bigger or equal to x_i and the number of data observations that are smaller or equal to x_i are equal. Notice that this definition may yield no unique median. In that case different alternative strategies are used, depending on the data type at hand (ordinal or metric). (See also the example below.) The median corresponds to the 50% quartile, a concept introduced below.

Example. The median of the vector $\vec{x} = \langle 1 = 0, 3, 6, 7 \rangle$ does not exist by the definition given above. However, for metric measures, where distances between measurements are meaningful, it is customary to take the two values “closest to where the median should be” and average them. In the example at hand, this would be $\frac{3+6}{2} = 4.5$. The plot below shows the data points in black, the mean as a red cross (as before) and the median as a blue circle



The function `median` from base R computes the median of a vector. It also takes an ordered factor as an argument.

```
median(c(0, 3, 6, 7))

## [1] 4.5
```

To please the avocados, let's also calculate the median price of both types of avocados and compare these to the means we calculated earlier already:

```
avocado_data %>%
  group_by(type) %>%
  summarise(
    mean_price = mean(average_price),
    median_price = median(average_price)
  )

## # A tibble: 2 x 3
##   type      mean_price median_price
##   <chr>        <dbl>       <dbl>
## 1 conventional     1.16       1.13
## 2 organic          1.65       1.63
```

5.2.2.3. The mode

The **mode** is the value that occurred most frequently in the data. While the mean is only applicable to metric variables, and the median only to variables that are at least ordinal, the mode is only reasonable for variables that have a finite set of different possible observations (nominal or ordinal).

There is no built-in function in R to return the mode of a (suitable) vector, but it is easily retrieved by obtaining counts.

5.2.3. Measures of dispersion

5.2.3.1. Variance

Variance is a widely used and very useful measure of dispersion for metric data. The variance $\text{Var}(\vec{x})$ of a vector of metric observations \vec{x} of length n is defined as the average of the squared distances from the mean:

$$\text{Var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\vec{x}})^2$$

5. Summary statistics

Example. The variance of the vector $\vec{x} \langle 0, 3, 6, 7 \rangle$ is computed as:

$$\text{Var}(\vec{x}) = \frac{1}{4} ((0 - 4)^2 + (3 - 4)^2 + (6 - 4)^2 + (7 - 4)^2) = \frac{1}{4} (16 + 1 + 4 + 9) = \frac{30}{4} = 7.5$$

Figure 5.3 shows a geometric interpretation of variance for the running example of the vector $\vec{x} \langle 0, 3, 6, 7 \rangle$.

Geometric visualization of variance

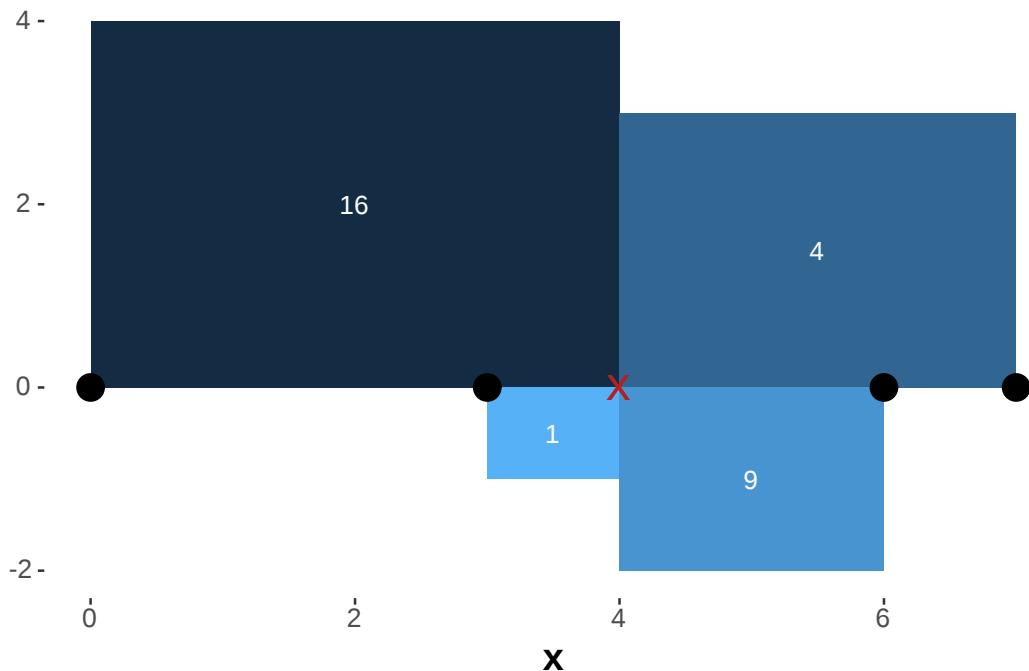


Figure 5.3.: Geometrical interpretation of variance. Four data points (black dots) and their mean (red cross) are shown, together with the squares whose sides are the differences between the observed data points and the mean. The numbers in white give the area of each square, which is also indicated by the coloring of each rectangle.

We can calculate the variance in R explicitly:

```
x <- c(0, 3, 6, 7)
sum((x - mean(x))^2) / length(x)

## [1] 7.5
```

There is also a built-in function `var` from base R. Using this we get a different result though:

```
x <- c(0, 3, 6, 7)
var(x)
```

```
## [1] 10
```

This is because `var` computes the variance by a slightly different formula to obtain an **unbiased estimator** of the variance for the case that the mean is not known but also estimated from the data. The formula for the unbiased estimator that R uses simply replaces the n in the denominator by $n - 1$:²

$$\text{Var}(\vec{x}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_{\vec{x}})^2$$

5.2.3.2. Standard deviation

The standard deviation $\text{SD}(\vec{x})$ or numeric vector \vec{x} is just the square root of the variance:

$$\text{SD}(\vec{x}) = \sqrt{\text{Var}(\vec{x})} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\vec{x}})^2}$$

Let's calculate the (biased) variance and standard deviation for the `average_price` of different types of avocados:

```
avocado_data %>%
  group_by(type) %>%
  summarize(
    variance_price = var(average_price),
    stddev_price   = sd(average_price),
  )

## # A tibble: 2 x 3
##   type      variance_price stddev_price
##   <chr>        <dbl>       <dbl>
## 1 conventional 0.0692      0.263
## 2 organic      0.132       0.364
```

²For current purpose it is not important what biased or unbiased estimator is and why this subtle change in the formula matters. We will come back to the issue of estimation in Chapter 9.

5. Summary statistics

5.2.3.3. Quantile

For a vector \vec{x} of at least ordinal measures, we can generalize the concept of a median to an arbitrary quantile. An $k\%$ quantile is the element x_i in \vec{x} such that $k\%$ of the data in \vec{x} lies below x_i . If this definition does not yield a unique element for some $k\%$ threshold, similar methods to what we saw for the median are applied.

We can use the base R function `quantile` to obtain the 10%, 25%, 50% and 85% quantiles (just arbitrary picks) for the `average_price` in the avocado data set:

```
quantile(  
  # vector of observations  
  x = avocado_data$average_price,  
  # which quantiles  
  probs = c(0.1, 0.25, 0.5, 0.85)  
)  
  
## 10% 25% 50% 85%  
## 0.93 1.10 1.37 1.83
```

This tells us, for instance, that only about ten percent of the data observations had prices lower than \$0.93.

5.2.4. Quantifying confidence with bootstrapping

Bootstrapping is an elegant way to obtain measures of confidence for summary statistics. These measures of confidence can be used for parameter inference, too. We will discuss parameter inference at length in Section 9. In this course, we will not use bootstrapping as an alternative approach to parameter inference. We will, however, follow a common practice (at least in some areas of Cognitive Psychology) to use **bootstrapped 95% confidence intervals of the mean** as part of descriptive statistics, i.e., in summaries and plots of the data.

The bootstrap is a method from a more general class of algorithms, namely so-called **resampling methods**. The general idea is, roughly put, that we treat the data at hand as the true representation of reality. We then imagine that we run an experiment on that (restricted, hypothetical) reality. We then ask ourselves: what would we estimate (e.g., as a mean) in any such hypothetical experiment. The more these hypothetical measures derived from hypothetical experiments based on a hypothetical reality differ, the less confident we are in the estimate. Sounds weird, but is mindblowingly elegant.

An algorithm for constructing a 95% confidence interval of the mean of vector D of numeric data with length k looks as follows:

1. take k samples from D with replacement, call this D^{rep} ³

³ D^{rep} is short for “repeated Data”. We will use this concept more later on. The idea is that we consider “hypothetical data” which we have not perceived, but which we might have. Repeated data is (usually) of the same shape and form as the original, observed data, which is also sometimes noted as D^{obs} for clarity in comparison to D^{rep} .

2. calculate the mean $\mu(D^{\text{rep}})$ of the newly sampled data
3. repeat steps 1 and 2 to gather r means of different resamples of D ; call the result vector μ_{sampled}
4. the boundaries of the 95% inner quantile of μ_{sampled} are the bootstrapped 95% confidence interval of the mean

The higher r , i.e., the more samples we take, the better the estimate. The higher k , i.e., the more observations we have to begin with, the less variable the means $\mu(D^{\text{rep}})$ of the resampled data will usually be. Hence, usually, the higher k the smaller the bootstrapped 95% confidence interval of the mean.

Here is a convenience function that we will use throughout the book to produce bootstrapped 95% confidence intervals of the mean:

```
## takes a vector of numbers and returns bootstrapped 95% ConfInt
## for the mean, based on `n_resamples` re-samples (default: 1000)
bootstrapped_CI <- function(data_vector, n_resamples = 1000) {
  resampled_means <- map_dbl(1:n_resamples, function(i) {
    mean(sample(x = data_vector,
                 size = length(data_vector),
                 replace = T))
  })
}
)
tibble(
  'lower' = quantile(resampled_means, 0.025),
  'mean' = mean(data_vector),
  'upper' = quantile(resampled_means, 0.975)
)
}
```

Applying this method to the vector of average avocado prices we get:

```
bootstrapped_CI(avocado_data$average_price)
```

```
## # A tibble: 1 x 3
##   lower  mean upper
##     <dbl> <dbl> <dbl>
## 1 1.40  1.41  1.41
```

Notice that, since `average_price` has length 18249, i.e., we have $k = 18249$ observations in the data, the bootstrapped 95% confidence interval is rather narrow. Compare this against a case of $k = 300$ obtained by only looking at the first 300 entries in `average_price`:

5. Summary statistics

```
# first 300 observations of `average price` only
smaller_data <- avocado_data$average_price[1:300]
bootstrapped_CI(smaller_data)

## # A tibble: 1 x 3
##   lower  mean upper
##   <dbl> <dbl> <dbl>
## 1 1.14  1.16  1.17
```

The mean is different (because we are looking at earlier time points) but, importantly, the interval is larger because with only 300 observations we have less confidence in the estimate.

5.2.4.1. Excursion: Summary functions with multiple outputs, using nested tibbles

To obtain summary statistics for different groups of a variable, we can use the function `bootstrapped_CI` conveniently in concert with **nested tibbles**, as demonstrated here:

```
avocado_data %>%
  group_by(type) %>%
  # nest all columns except grouping-column 'type' in a tibble
  # the name of the new column is 'price_tibbles'
  nest(.key = "price_tibbles") %>%
  # collect the summary statistics for each nested tibble
  # the outcome is a new column with nested tibbles
  summarise(
    CIs = map(price_tibbles, function(d) bootstrapped_CI(d$average_price))
  ) %>%
  # unnest the newly created nested tibble
  unnest(CIs)

## # A tibble: 2 x 4
##   type      lower  mean upper
##   <chr>     <dbl> <dbl> <dbl>
## 1 conventional 1.15  1.16  1.16
## 2 organic      1.65  1.65  1.66
```

Using nesting in this case is helpful because we only want to run the bootstrap function once, but we need both of the numbers it returns. The following explains nesting based on this example.

To understand what is going on with nested tibbles, notice that the `nest` function in this example creates a nested tibble with just two rows, one for each value of the variable `type`, each of which contains a tibble that contains all the data. The column `price_tibbles` in the first row contains the whole data for all observations for conventional avocado:

```

avocado_data %>%
  group_by(type) %>%
  # nest all columns except grouping-column 'type' in a tibble
  # the name of the new column is 'price_tibbles'
  nest(.key = "price_tibbles") %>%
  # extract new column with tibble
  pull(price_tibbles) %>%
  # peak at the first entry in this vector
  .[1] %>% head()

## <list_of<
##   tbl_df<
##     Date          : date
##     average_price : double
##     total_volume_sold: double
##     small         : double
##     medium        : double
##     large         : double
##   >
## >[1]>
## [[1]]
## # A tibble: 9,126 x 6
##       Date     average_price total_volume_sold small  medium large
##     <date>        <dbl>            <dbl>      <dbl>    <dbl>   <dbl>
## 1 2015-12-27      1.33        64237.   1037.  54455.   48.2
## 2 2015-12-20      1.35        54877.   674.   44639.   58.3
## 3 2015-12-13      0.93       118220.   795.  109150.  130.
## 4 2015-12-06      1.08       78992.   1132.  71976.   72.6
## 5 2015-11-29      1.28       51040.   941.   43838.   75.8
## 6 2015-11-22      1.26       55980.   1184.  48068.   43.6
## 7 2015-11-15      0.99       83454.   1369.  73673.   93.3
## 8 2015-11-08      0.98       109428.  704.   101815.   80
## 9 2015-11-01      1.02       99811.  1022.  87316.   85.3
## 10 2015-10-25     1.07       74339.   842.   64757.  113
## # ... with 9,116 more rows

```

After nesting, we call the custom function `bootstrapped_CI` on the variable `average_price` inside of every nested tibble, so first for the conventional, then the organic avocados. The results is a nested tibble. If we now look inside the new column `CI`, we see that it's cells contain tibbles with the output of each call of `bootstrapped_CI`:

5. Summary statistics

```
avocado_data %>%
  group_by(type) %>%
  # nest all columns except grouping-column 'type' in a tibble
  # the name of the new column is 'price_tibbles'
  nest(.key = "price_tibbles") %>%
  # collect the summary statistics for each nested tibble
  # the outcome is a new column with nested tibbles
  summarise(
    CIs = map(price_tibbles, function(d) bootstrapped_CI(d$average_price))
  ) %>%
  # extract new column vector with nested tibbles
  pull(CIs) %>%
  # peak at the first entry
  .[1] %>% head()

## [[1]]
## # A tibble: 1 x 3
##   lower  mean upper
##   <dbl> <dbl> <dbl>
## 1 1.15  1.16  1.16
```

Finally, we unnest the new column CIs to obtain the final result (code repeated from above):

```
avocado_data %>%
  group_by(type) %>%
  # nest all columns except grouping-column 'type' in a tibble
  # the name of the new column is 'price_tibbles'
  nest(.key = "price_tibbles") %>%
  # collect the summary statistics for each nested tibble
  # the outcome is a new column with nested tibbles
  summarise(
    CIs = map(price_tibbles, function(d) bootstrapped_CI(d$average_price))
  ) %>%
  # unnest the newly created nested tibble
  unnest(CIs)

## # A tibble: 2 x 4
##   type      lower  mean upper
##   <chr>     <dbl> <dbl> <dbl>
## 1 conventional 1.15  1.16  1.16
## 2 organic      1.65  1.65  1.66
```

5.3. Co-variance & correlation

5.3.1. Covariance

Let \vec{x} and \vec{y} are two vectors of numeric data of the same length, such that all pairs of x_i and y_i are associated observation. For example: the vectors `avocado_data$total_volume_sold` and `avocado_data$average_price` would be such vectors. The covariance between \vec{x} and \vec{y} measures, intuitively put, the degree to which changes in one vector correspond with changes in the other. Formally, covariance is defined as follows (notice that we use $n - 1$ in the denominator, to obtain an unbiased estimator if the means are unknown):

$$\text{Cov}(\vec{x}, \vec{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_{\vec{x}})(y_i - \mu_{\vec{y}})$$

There is a visually intuitive geometric interpretation of covariance. To see this, let's look at a short contrived example.

```
contrived_example <-
  tribble(
    ~x,      ~y,
    2,       2,
    2.5,    4,
    3.5,    2.5,
    4,       3.5
  )
```

First, notice that the mean of x and y is 3:

```
means_contr_example <- map_df(contrived_example, mean)
means_contr_example

## # A tibble: 1 x 2
##       x     y
##   <dbl> <dbl>
## 1     3     3
```

We can then compute the covariance as follows:

```
contrived_example <-
  contrived_example %>%
  mutate(
    area_rectangle =
      (x-mean(x)) * (y - mean(y)),
```

5. Summary statistics

```

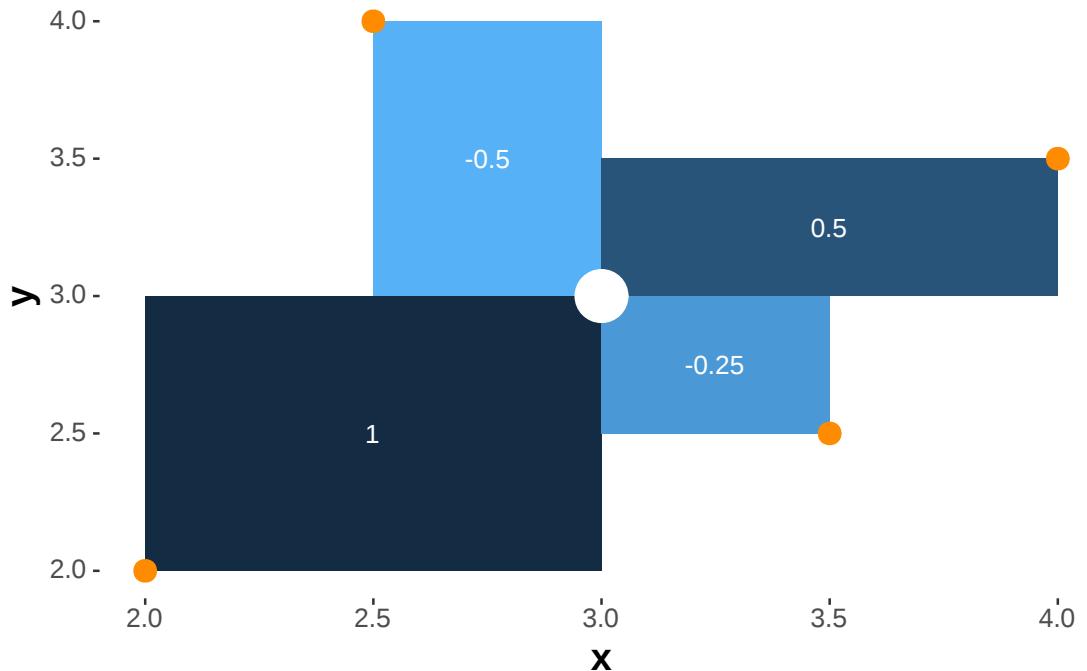
covariance      =  1/ (n()-1) * sum((x-mean(x)) * (y - mean(y)))
)
contrived_example

## # A tibble: 4 x 4
##       x     y area_rectangle covariance
##   <dbl> <dbl>     <dbl>        <dbl>
## 1     2     2         1        0.25
## 2     2.5    4        -0.5       0.25
## 3     3.5    2.5      -0.25      0.25
## 4     4     3.5         0.5       0.25

```

Similar to what we did with variance, we can give a geometrical interpretation of covariance. Figure ?? shows the four summands contributing to the covariance of the `contrived_example`. What this graph clearly shows is that summands can have different signs. If x_i and y_i are both bigger than the mean, or if both are smaller than the mean, then the corresponding summand is positive. Otherwise, the corresponding summand is negative. This means that covariance captures the degree to which pairs x_i and y_i tend to deviate from the mean in the same general direction. A positive covariance is indicative of a positive general association between \vec{x} and \vec{y} , while a negative covariance suggests that as you increase x_i the associated y_i becomes smaller.

Rectangular areas contributing to the computational covariance



We can, of course, also calculate covariance just with a built-in base R function, `cov`:

```
with(contrived_example, cov(x,y))

## [1] 0.25
```

And, using this function we can calculate the covariance between `total_volume_sold` and `average_price` in the avocado data:

```
with(avocado_data, cov(log(total_volume_sold), average_price))

## [1] -0.5388084
```

Interestingly, the negative covariance in this example suggests that that across all associated data pairs, the larger `total_volume_sold`, the lower `average_price`. It is important that this is a descriptive statistics, and that this is not to be interpreted as evidence or a causal relation between the two measures of interest. Not in this example, not in any other. Covariance describes associated data points; it is not evidence for causal relationships.

5.3.2. Correlation

The problem with covariance is that it is not invariant under linear transformation. Consider the `contrived_example` from above once more. The original data had the following covariance:

```
with(contrived_example, cov(x,y))

## [1] 0.25
```

But if we just linearly transform, say, vector `y` to `1000 * y + 500` (e.g., because we switch to an equivalent, but numerically different measuring scale), we obtain:

```
with(contrived_example, cov(x,1000 * y + 500))

## [1] 250
```

To compensate for this problem, we can look at **Bravais-Pearson correlation**, which is covariance standardized by standard deviations:

$$r_{\vec{x}\vec{y}} = \frac{\text{Cov}(\vec{x}, \vec{y})}{\text{SD}(\vec{x}) \text{SD}(\vec{y})}$$

Let's check invariance under linear transformation, using the built-in function `cor`. The correlation coefficient for the original data is:

5. Summary statistics

```
with(contrived_example, cor(x,y))
```

```
## [1] 0.3
```

The correlation coefficient for the data with linearly transformed y is:

```
with(contrived_example, cor(x,1000 * y + 500))
```

```
## [1] 0.3
```

Indeed, the correlation coefficient is nicely bounded to lie between -1 and 1. A correlation coefficient of 0 is to be interpreted as the absence of any correlation. A correlation coefficient of 1 is a perfect positive correlation (the higher x_i , the higher y_i) and -1 indicates a perfect negative correlation (the higher x_i the lower y_i). Again, pronounced positive or negative correlations are *not* to be confused with strong evidence for a causal relation. It is just a descriptive statistic on properties of associated measurements.

In the avocado data, the logarithm of `total_volume_sold` shows a noteworthy correlation with `average_price`. This is also visible in Figure 5.4.

```
with(avocado_data, cor(log(total_volume_sold), average_price))
```

```
## [1] -0.5834087
```

Avocado prices plotted against the (log) amount

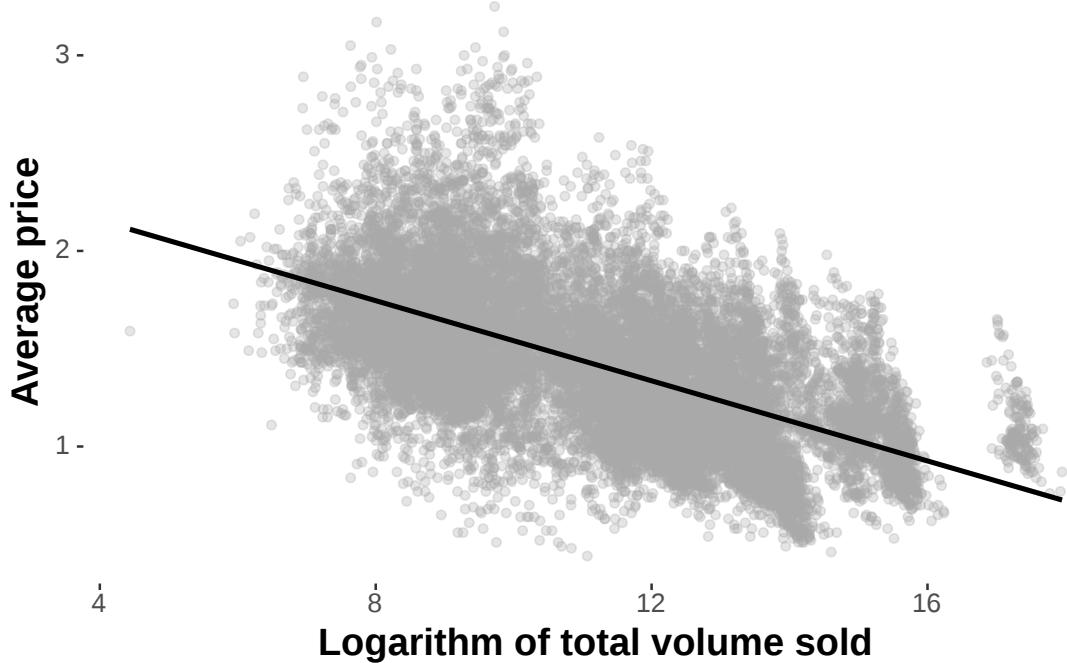


Figure 5.4.: Scatter plot of avocado prices, plotted agains (logarithms of) the total amount sold. The black line is a linear regression line indicating the (negative) correlation between these measures (more on this later).

6. Data Visualization

Numerical summaries of complex data always incur information loss. Still lossy, but less so (if done well), is visualization. Any serious data analysis should start with a process in which the analyst becomes intimate with the data at hand. Visualization is an integral part of data-intimacy.

Section 6.1 demonstrates how summary statistics can be misleading and how a simple visualization can be much more revealing. Section 6.2 offers some reflection on what makes a data visualization successful. Section 6.3 introduces the basics of data visualization with the `ggplot` package, an integral part of the `tidyverse`, based on a scatter plot for the avocado price data. Going beyond scatter plots, section 6.4 looks at some common types of plots and how to realize them using the `geom_` family of functions in `ggplot`.

The learning goals for this chapter are:

- obtain a basic understanding of better/worse plotting
 - understand the idea of *hypothesis-driven visualization*
- develop a basic understanding of the 'grammar of graphs'
- get familiar with frequent visualization strategies
 - barplots, densities, violins, error bars etc.
- be able to fine-tune graphs for better visualization

6.1. Motivating example: Anscombe's quartet

To see how summary statistics can be highly misleading, and how a simple plot can reveal a lot more, consider a famous dataset available in R (Anscombe 1973):

```
glimpse(anscombe %>% as_tibble)

## Observations: 11
## Variables: 8
## $ x1 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x2 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x3 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x4 <dbl> 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
```

6. Data Visualization

```
## $ y1 <dbl> 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
## $ y2 <dbl> 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
## $ y3 <dbl> 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
## $ y4 <dbl> 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
```

There are four pairs of x and y coordinates. Unfortunately, these are stored in long format with two pieces of information buried inside of the column name: for instance, the name `x3` contains the information that this column contains the x coordinates for the 3rd pair. This is rather untidy. But, using tools from the `dplyr` package, we can tidy up quickly:

```
tidy_anscombe <- anscombe %>% as_tibble %>%
  pivot_longer(
    ## we want to pivot every column
    everything(),
    ## use reg-exps to capture 1st and 2nd character
    names_pattern = "(.)(.)",
    ## assign names to new cols, using 1st part of
    ## what reg-exp captures as new column names
    names_to = c(".value", "grp")
  ) %>%
  mutate(grp = paste0("Group ", grp))
tidy_anscombe

## # A tibble: 44 x 3
##       grp      x     y
##   <chr>  <dbl> <dbl>
## 1 Group 1     10  8.04
## 2 Group 2     10  9.14
## 3 Group 3     10  7.46
## 4 Group 4      8  6.58
## 5 Group 1      8  6.95
## 6 Group 2      8  8.14
## 7 Group 3      8  6.77
## 8 Group 4      8  5.76
## 9 Group 1     13  7.58
## 10 Group 2    13  8.74
## # ... with 34 more rows
```

Here are some summary statistics for each of the four pairs:

```
tidy_anscombe %>%
  group_by(grp) %>%
```

```

summarise(
  mean_x    = mean(x),
  mean_y    = mean(y),
  min_x     = min(x),
  min_y     = min(y),
  max_x     = max(x),
  max_y     = max(y),
  crrltn    = cor(x,y)
)

## # A tibble: 4 x 8
##   grp      mean_x mean_y min_x min_y max_x max_y crrltn
##   <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Group 1     9    7.50     4    4.26    14   10.8   0.816
## 2 Group 2     9    7.50     4     3.1     14    9.26   0.816
## 3 Group 3     9    7.5      4    5.39    14   12.7   0.816
## 4 Group 4     9    7.50     8    5.25    19   12.5   0.817

```

These numeric indicators suggest that each pair of x and y values is very similar. Only the ranges seem to differ. A brilliant example of how misleading numeric statistics can be, as compared to a plot of the data:¹

```

tidy_anscombe %>%
  ggplot(aes(x, y)) +
  geom_smooth(method = lm, se = F, color = "darkorange") +
  geom_point(color = project_colors[3], size = 2) +
  scale_y_continuous(breaks = scales::pretty_breaks()) +
  scale_x_continuous(breaks = scales::pretty_breaks()) +
  labs(
    title = "Anscombe's Quartet", x = NULL, y = NULL,
    subtitle = bquote(y == 0.5 * x + 3 ~ (r %~~% .82) ~ "for all groups")
  ) +
  facet_wrap(~grp, ncol = 2, scales = "free_x") +
  theme(strip.background = element_rect(fill = "#f2f2f2", colour = "white"))

```

6.2. Visualization: the good, the bad and the info-graphic

Producing good data visualization is very difficult. There are no uncontroversial criteria for what a good visualization should be. There are, unfortunately, quite clear examples of really bad visualizations. We will look at some of these examples in the following.

¹It is not important to understand this code when you first read this chapter. But at the end of the chapter you should be able to understand (passively) what is going on here.

Anscombe's Quartet

$y = 0.5x + 3$ ($r \approx 0.82$) for all groups

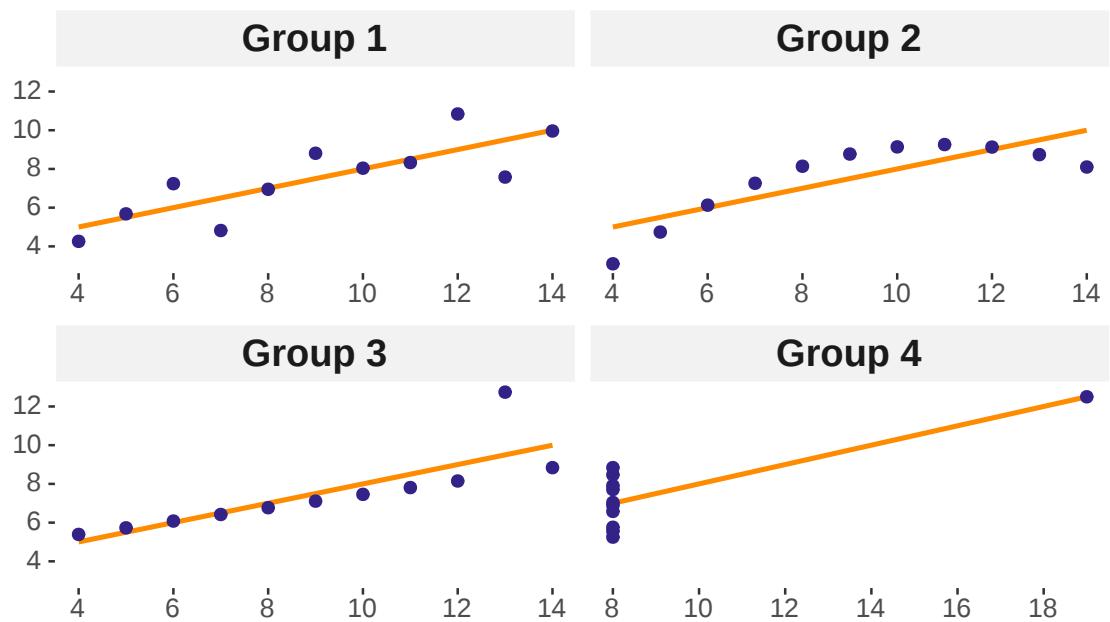


Figure 6.1.: Anscombe's Quartet: four different data sets all of which receive the same correlation score.

An absolute classic on data visualization is an early book by Edward Tufte (1983) entitled "The Visual Display of Quantitative Information". A distilled and over-simplified summary of Tufte's proposal is that we should eliminate *chart junk* and increase the *data-ink ratio*, a concept which Tufte defines formally. The more information (=data) a plot conveys, the higher the data-ink ratio. The more ink it requires, the lower it is.

However, not all information in the data is equally relevant. Also, spending extra ink to reduce the recipients mental effort of retrieving the relevant information can be justified. Essentially, I would here propose to consider a special case of data visualization, common to scientific presentations. I want to speak of **hypothesis-driven visualization** as a way of communicating a clear message, the message we care most about at the current moment of (scientific) exchange. Though merely a special instances of all the goals one could pursue with data visualization, focusing on this special case is helpful because it allows us to formulate a (defeasible) rule of thumb for good visualization in analogy to how natural language ought to be used in order to achieve optimal cooperative information flow (at least as conceived by authors such as ...)

The vague & defeasible rule of thumb of good data visualization (according to the author).

"Communicate a maximal degree of relevant true information in a way that minimizes the recipient's effort of retrieving this information."

Interestingly, just like natural language also needs to rely on a conventional medium for expressing ideas which might put additional constraints on what counts as optimal communication (e.g., we might not be allowed to drop a pronoun in English even though it is clearly recoverable from the context, and Italian speakers would happily omit it), so do certain unarticulated conventions in each specific scientific field.²

Here are a few examples of bad plotting.³ To begin with, check out this fictitious data set:

```
large_contrast_data <- tribble(
  ~group, ~treatment, ~measurement,
  "A",      "on",       1000,
  "A",      "off",      1002,
  "B",      "on",       992,
  "B",      "off",      990
)
```

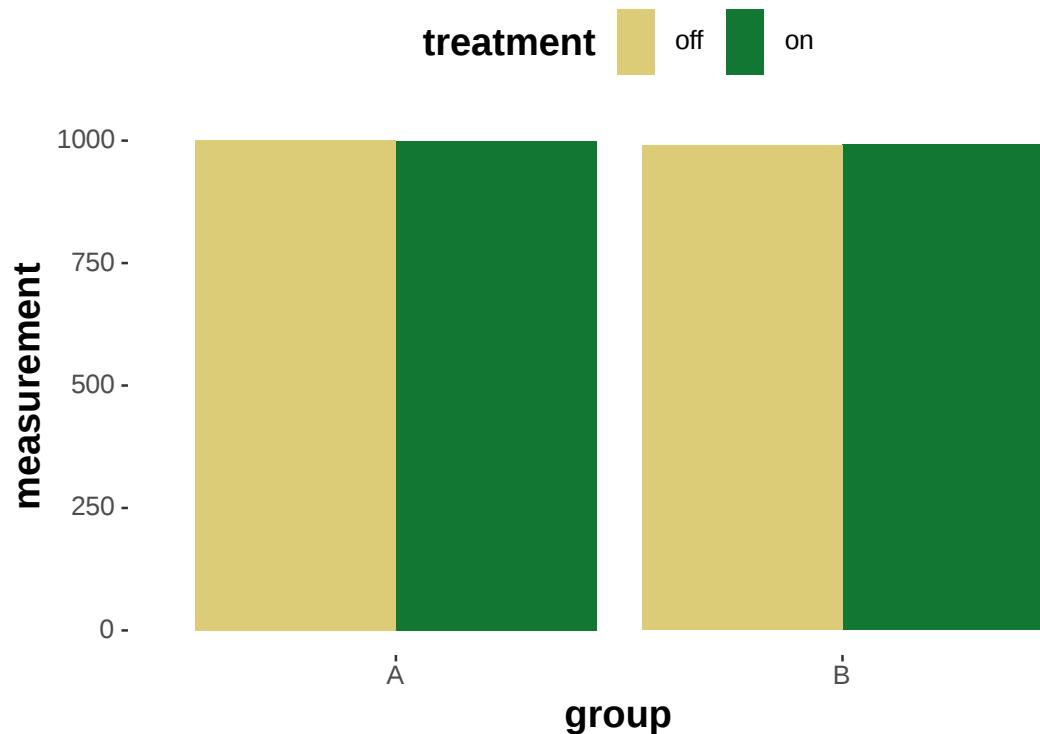
If we are interested in any potential influence of variables `group` and `treatment` on the measurement in question, the following graph is ruinously unhelpful because the large size of the bars renders the relatively small differences between them almost entirely unspottable.

²If your community only understands scatter plots and bar plots, it will not help communication but only mark you as a pompous show-off if you communicate in any other way, no matter how much better you think this is.

³For more disinspiration, see for example this curated list of delightfully bad visualizations from actual publications.

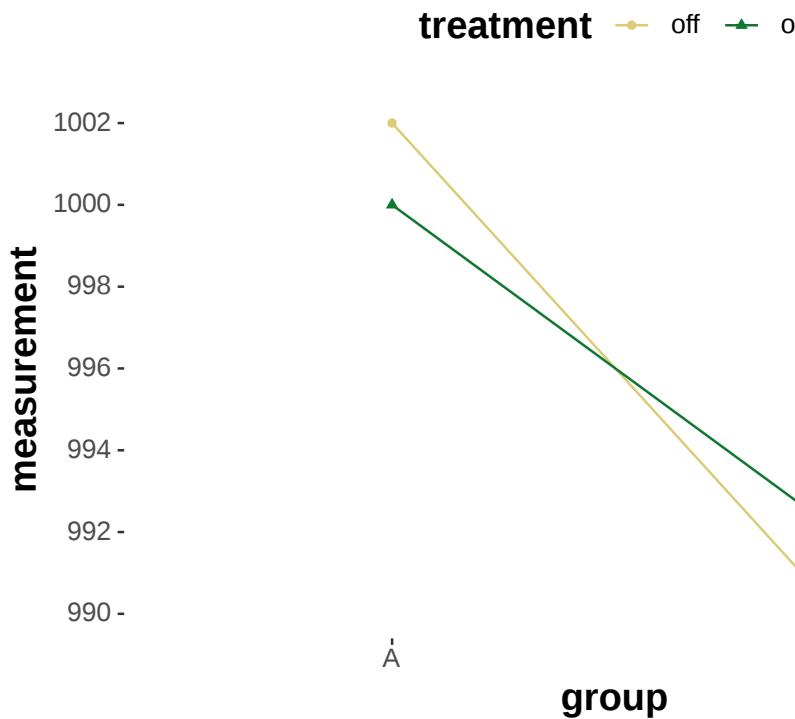
6. Data Visualization

```
large_contrast_data %>%
  ggplot(aes(x = group, y = measurement, fill = treatment)) +
  geom_bar(stat = "identity", position = "dodge")
```



A better visualization would be this:

```
large_contrast_data %>%
  ggplot(aes(
    x = group,
    y = measurement,
    shape = treatment,
    color = treatment,
    group = treatment
  ))
) +
  geom_point() +
  geom_line() +
  scale_y_continuous(breaks = scales::pretty_breaks())
```



The following examples use the Bio-Logic Jazz-Metal data set, in particular the following derived table of counts or the derived table of proportions:

`BLJM_associated_counts`

```
## # A tibble: 4 x 3
##   JM     LB       n
##   <chr> <chr>   <int>
## 1 Jazz   Biology  38
## 2 Jazz   Logic    26
## 3 Metal  Biology  20
## 4 Metal  Logic    18
```

It is probably hard to believe but Figure 6.2 was obtained without further intentional uglification just by choosing a default 3D bar plot display in Microsoft's Excel. It does actually show the relevant information but it is entirely useless for a human observer without a magnifying glass, professional measuring tools and a calculator.

It gets slightly better with the following pie chart of the same numerical information, also generated with Microsoft's Excel. Subjectively, Figure 6.3 is pretty much anything but pretty. Objectively, it is better than the previous visualization in terms of 3D bar plots shown in Figure 6.2 but the pie chart is still not useful for

Counts of music-subject choice pairs

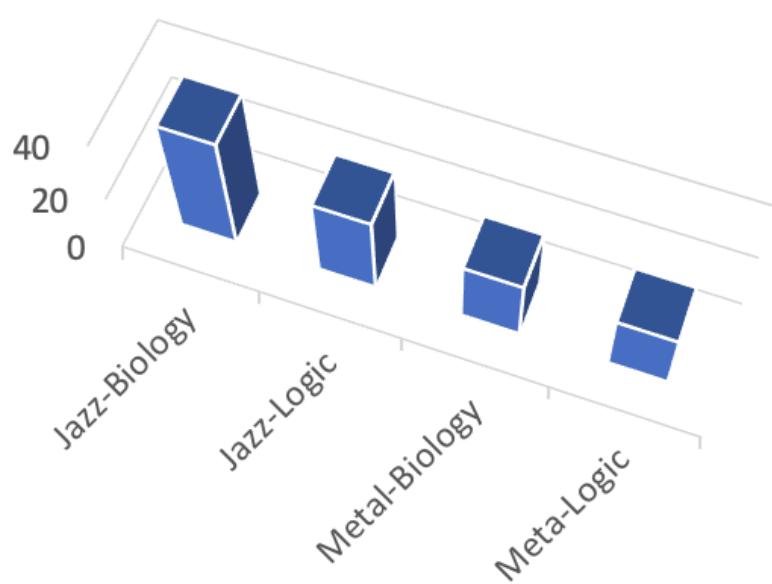


Figure 6.2.: Example of a frontrunner for the prize of today's most complete disaster in the visual communication of information.

answering the question which we care about, namely whether logicians are more likely to prefer Jazz over Metal than biologists.

Proportions of music-subject choice pairs

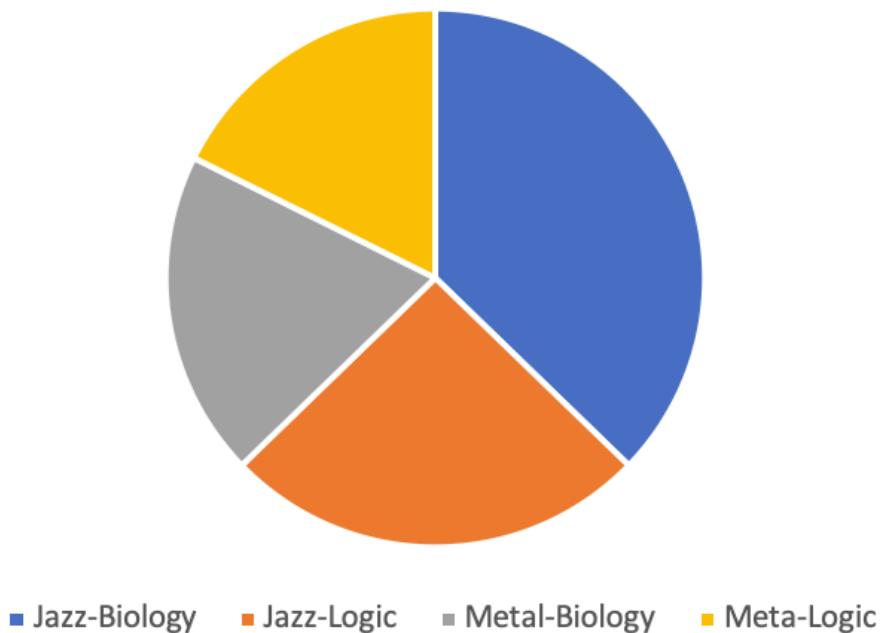


Figure 6.3.: Example of a rather unhelpful visual representation of the BLJM data (when the research question is whether logicians are more likely to prefer Jazz over Metal than biologists).

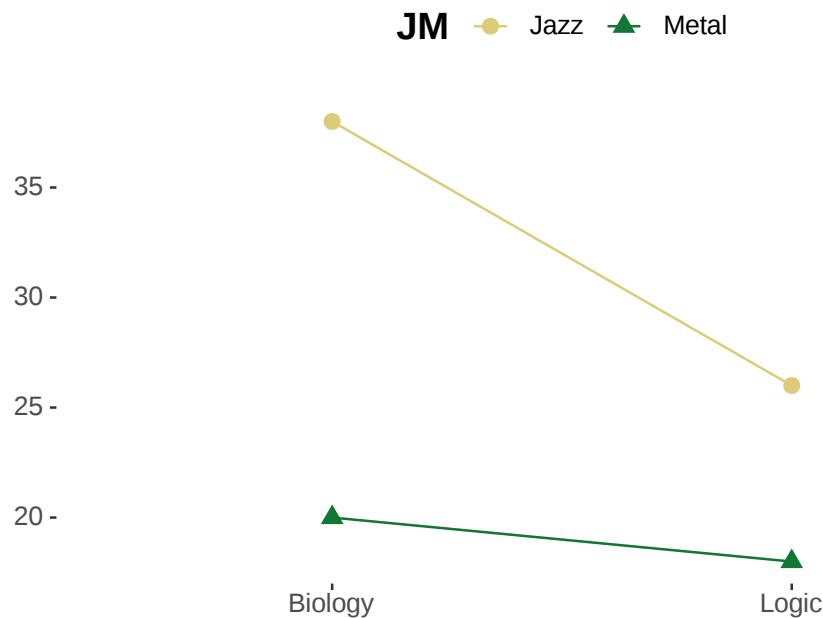
We can produce a much more useful representation with the code below. (A similar visualization also appeared as Figure 5.1 in the previous chapter.)

```
BLJM_associated_counts %>%
  ggplot(
    aes(
      x = LB,
      y = n,
      color = JM,
      shape = JM,
      group = JM
    )
  ) +
  geom_point(size = 3) + geom_line() +
  labs(
    title = "Counts of choices of each music+subject pair",
    x = "",
```

6. Data Visualization

```
y = ""  
)
```

Counts of choices of each music+subject pair



Info-Graphics. Scientific communication with visualized data is different from other modes of communication with visualized data. These other contexts come with different requirements for good data visualization. Good examples of highly successful *info-graphics* are produced by the famous illustrator Nigel Holmes, for instance. Figure 6.4 is an example from Holmes' website showing different amounts of energy consumption for different household appliances. The purpose of this visualization is not (only) to communicate information about which of the listed household appliances is most energy intensive. It's main purpose is to raise awareness for the unexpectedly large energy consumption of household appliances in general (in stand-by mode).⁴

6.3. Basics of ggplot

In this section we will work towards a first plot with `ggplot`. It will be a scatter plot (more on different kinds of plots in Section 6.4) for the avocado price data. Check out the `ggplot` cheat sheet for a quick overview of the nuts and bolts of `ggplot`.

The following introduces the following key concepts of `ggplot`:

⁴Image retrieved from Nigel Holmes' website website on November 25 2019.

Vampire Energy

Even when household appliances are turned off, most are still using some electricity. Appliances are either in passive standby mode (the clock on the microwave is still ticking) or active standby mode (the VCR is off, but programmed to record something).

These numbers are for average standby modes, showing how much electricity is sucked out annually, in kilowatt hours, and what it costs you—assuming 11 cents per kilowatt hour. **Red lines** show passive standby mode; **blue lines** show active standby mode.

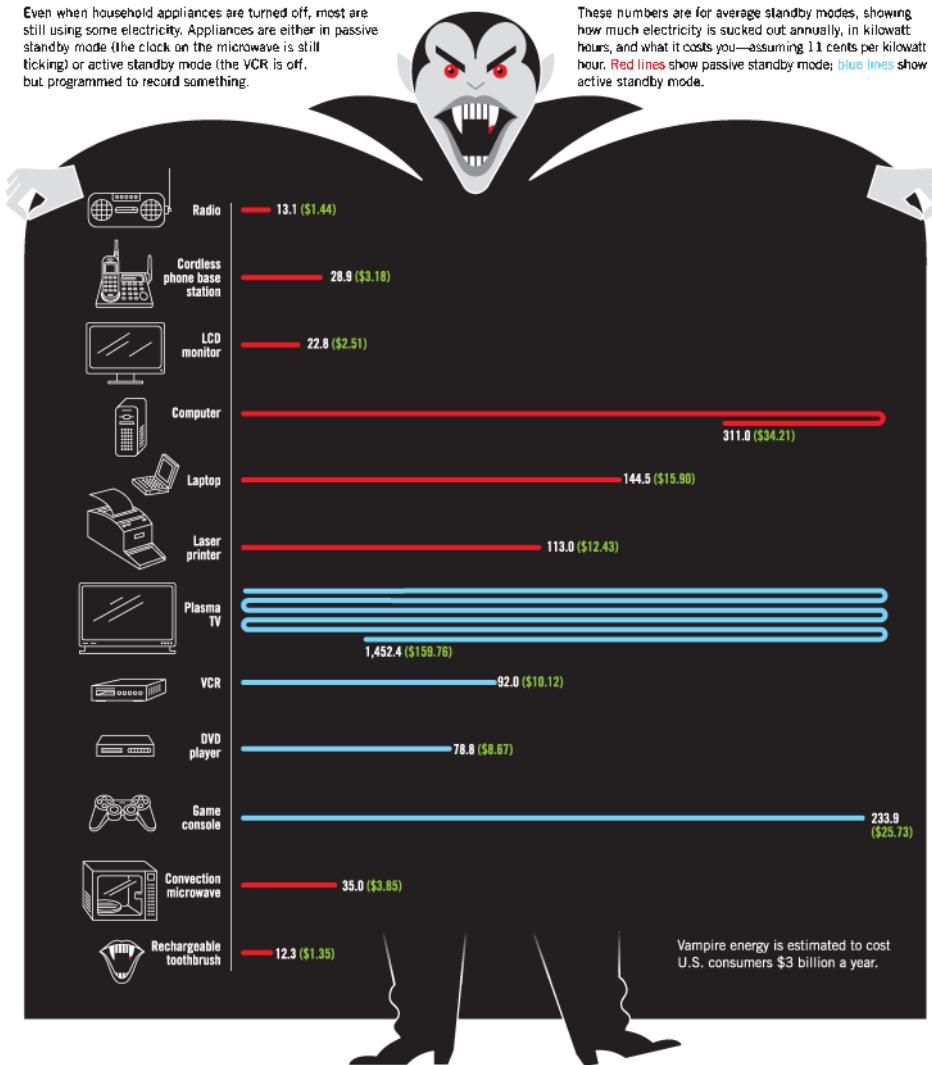


Figure 6.4.: Example of an infographic. While possibly considered 'chart junk' in a scientific context, the eye-catching and highly memorable (and pretty!) artwork serve a strong secondary purpose in contexts other than scientific ones where hypothesis-driven precise communication with visually presented data is key.

6. Data Visualization

- **incremental composition:** adding elements or changing attributes of a plot incrementally
- **convenience functions & defaults:** a closer look at high-level convenience functions (like `geom_point`) and what they actually do
- **layers:** seeing how layers are stacked when we call, e.g. different `geom_` functions in sequence
- **grouping:** what happens when we use grouping information (e.g., for color, shape or in facets)

The section finishes with a first full example of plot that has different layers, uses grouping, and customizes a few other things.

6.3.1. Incrementally composition of a plot

The “gg” in the package name `ggplot` is short for “grammar of graphs”. It provides functions for describing scientific data plots in a compositional manner, i.e., for dealing with different recurrent elements in a plot in an additive way. As a result of this approach, we will use the symbol `+` to add more and more elements (or to override the implicit defaults in previously evoked elements) to build a plot. For example, we can obtain a scatter plot for the avocado price data simply by first calling the function `ggplot`, which just creates an empty plot:

```
incrementally_built_plot <- ggplot()
```

The plot stored in variable `incrementally_built_plot` is very boring. Take a look:

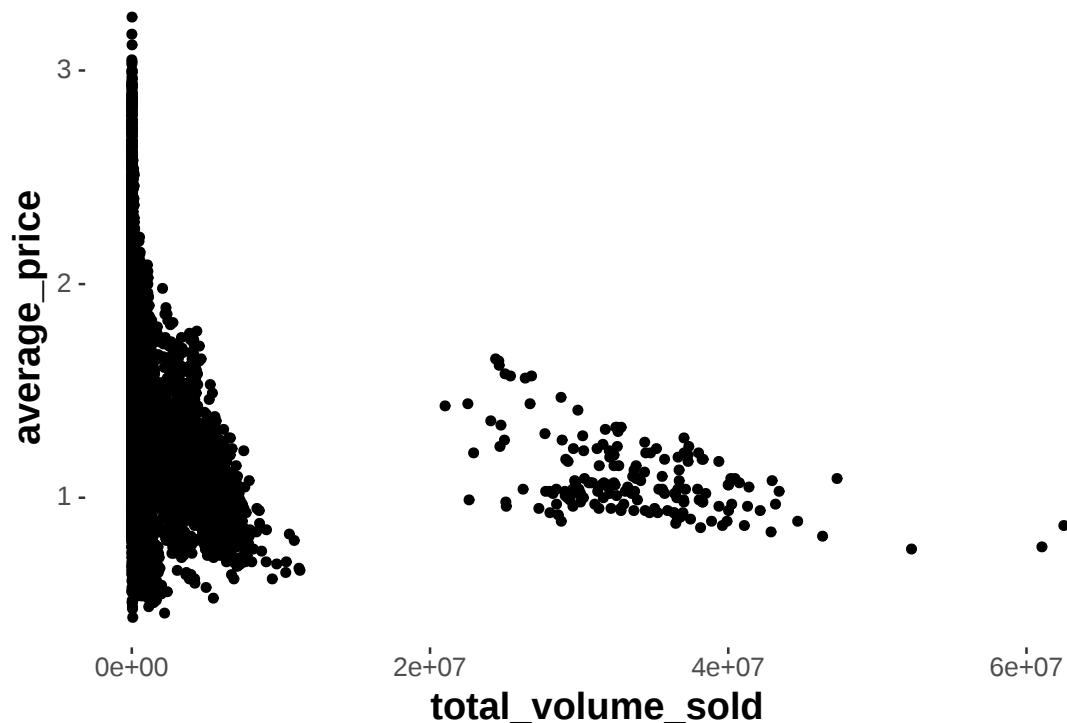
```
incrementally_built_plot
```

As you can see, you do not see anything except a (white) canvas. But we can add some stuff. Don't get hung up on the details right now, just notice that we use `+` to add stuff to our plot.⁵

```
incrementally_built_plot +
  # add a geom of type `point` (=> scatter plot)
  geom_point(
    # what data to use
    data = avocado_data,
    # supply a mapping (in the form of an 'aesthetic' (see below))
    mapping = aes(
      # which variable to map onto the x-axis
      x = total_volume_sold,
      # which variable to map onto the y-axis
      y = average_price
    )
  )
```

⁵If you run this code for yourself, the output is likely to look different from what is shown here. This is because this web-book uses a default theme for all of its plots. We will come back to customization with themes later.

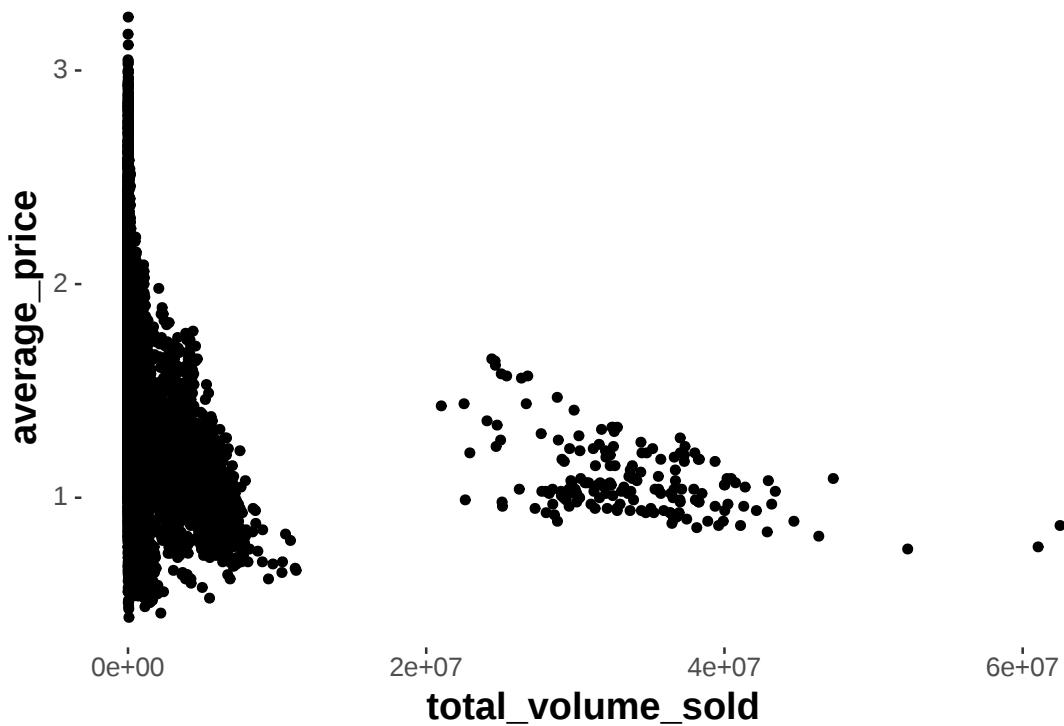
6. Data Visualization



You see that the function `geom_point` is what makes the points appear. You tell it which data to use and which mapping of variables from the data set to elements in the plot you like. That's it, at least to begin with.

We can also supply the information about the data to use and the aesthetic mapping in the call to function `ggplot`. Doing so will make this information the default for any subsequently added layer. Notice also that the `data` argument in function `ggplot` is the first argument, so we will frequently make use of piping, like in the following code which is equivalent to the previous in terms of output:

```
avocado_data %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point()
```



6.3.2. Elements in the layered grammar of graphs

Let's take a step back. Actually, the function `geom_point` is a convenience function that does a lot of things automatically for us. It helps understanding subsequent code if we peek under the hood at least for a brief moment initially, if only to just realize where some of the terminology in and round the "grammar of graphs" comes from.

The `ggplot` package defines a **layered grammar of graphs** (Wickham 2010). This is a structured description language for plots (relevant for data science). It uses a smart system of defaults so that it suffices to often just call a convenience wrapper like `geom_point`. But underneath there is the possibility of tinkering with (almost?) all of the (layered) elements and to change the defaults if need be.

The process of mapping data onto a visualization essentially follows this route:

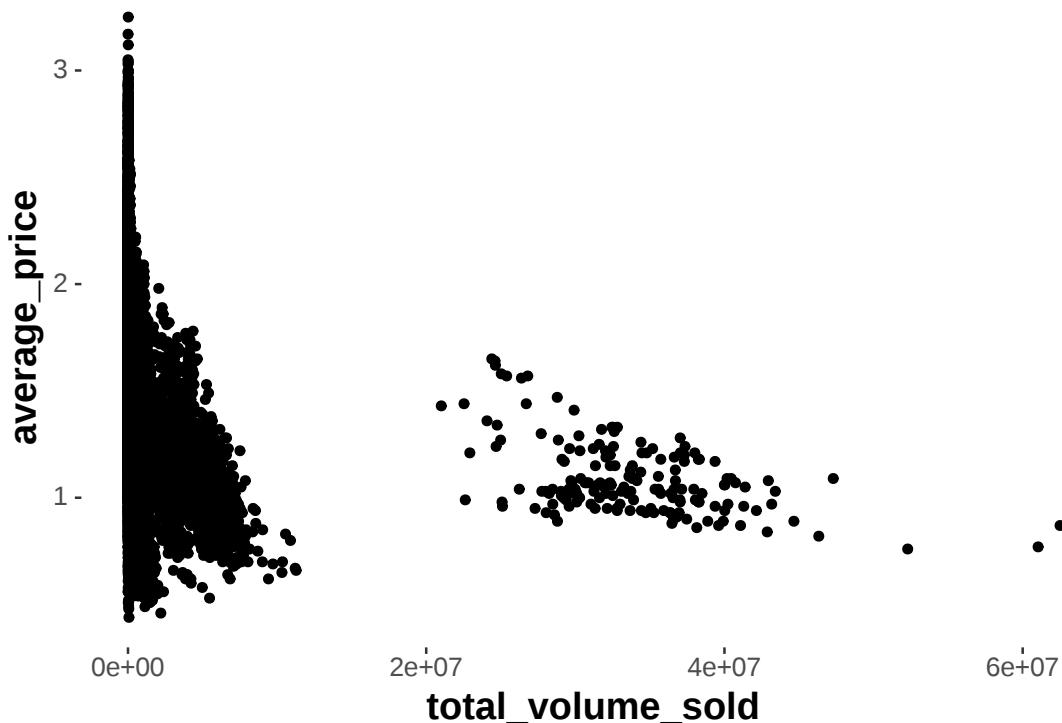
data -> statistical transformation -> geom. object -> aesthetics

You supply (tidy) data. The data is then transformed (e.g., by computing a summary statistic) in some way or another. This could just be an "identity map" in which case you will visualize the data exactly as it is. The resulting data representation is mapped onto some spatial (geometric) appearance, like a line, a dot, or a geometric shape. Finally, there is room to alter the specific asthetics of this mapping from data to visual object, like adjusting the size or the color of a geometric object, possibly depending on some other properties it has (e.g., whether it is an observation for a conventional or an organically grown avocado).

6. Data Visualization

To make explicit the steps which are implicitly carried out by `geom_points` in the example above, here is a fully verbose but output-equivalent sequence of commands that builds the same plot by defining all the basic components manually:

```
avocado_data %>%
  ggplot() +
    # plot consists of layers (more on this soon)
    layer(
      # how to map columns onto ingredients in the plot
      mapping = aes(x = total_volume_sold, y = average_price),
      # what statistical transformation should be used? - here: none
      stat = "identity",
      # how should the transformed data be visually represented? - here: as points
      geom = "point",
      # should we tinker in any other way with the positioning of each element?
      # - here: no, thank you!
      position = "identity"
    ) +
    # x & y axes are non-transformed continuous
    scale_x_continuous() +
    scale_y_continuous() +
    # we use a cartesian coordinate system (not a polar or a geographical map)
    coord_cartesian()
```



In this explicit call, we still need to specify the data and the mapping (which variable to map onto which axis). But we need to specify much more. We tell `ggplot` that we want standard (e.g., not log-transformed axis). We also tell it that our axis are continuous, that the data should not be transformed and that the visual shape (=geom) to which the data is to be mapped is a point (hence the name `geom_point`).

It is not important to understand all of these components right now. It is important to have seen them once, and to understand that `geom_point` is a wrapper around this call which assumes reasonable defaults (such as non-transformed axis, points for representation etc.).

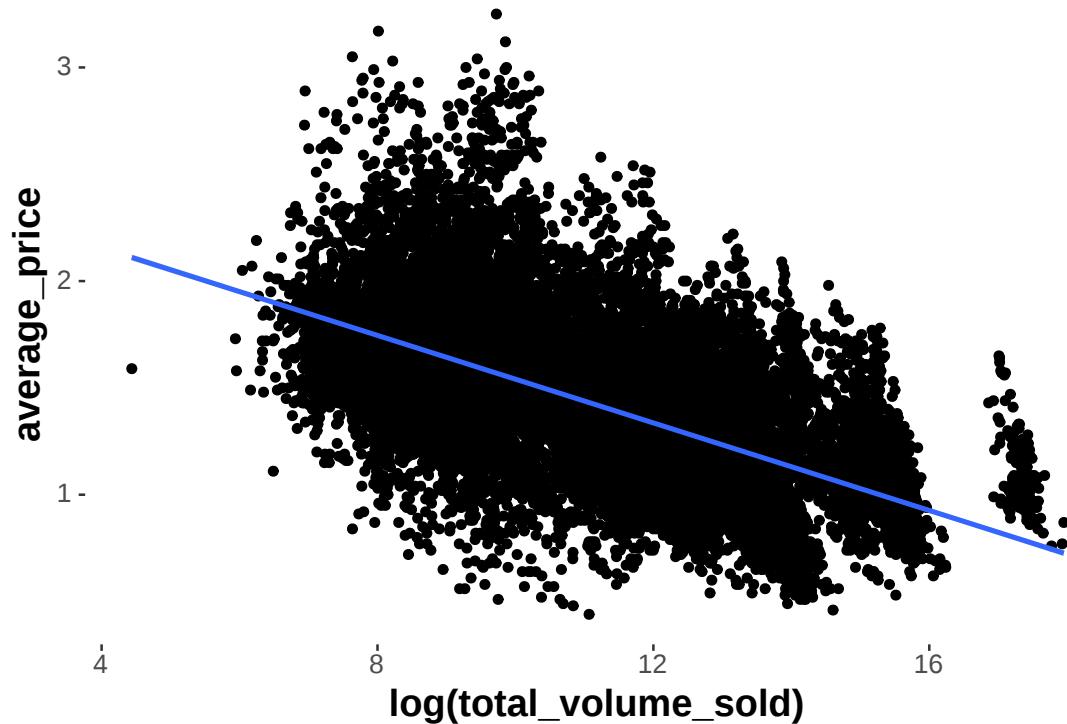
6.3.3. Layers and groups

`ggplot` is the “grammar of layered graphs”. Plots are compositionally built by combining different layers, if need be. For example, we can use another function from the `geom_` family of functions to display a different visualization derived from the same data on top of our previous scatter plot.⁶

⁶Notice that, as soon as we add the linear regression line, it makes sense to use the logarithm of the `total_volume_sold` because otherwise the fit is quite ridiculous. The logarithm helps to spread out the large number of data points where `total_volume_sold` is very low, and to “bring back to the flock” the data points where the `total_volume_sold` is outlierly high. It can be quite useful to use such transformations, if they are well understood. It is controversial whether such transformations should precede statistical analyses, but that is not important right now.

6. Data Visualization

```
avocado_data %>%
  ggplot(
    mapping = aes(
      # notice that we use the log (try without it to understand why)
      x = log(total_volume_sold),
      y = average_price
    )
  ) +
  # add a scatter plot
  geom_point() +
  # add a linear regression line
  geom_smooth(method = "lm")
```



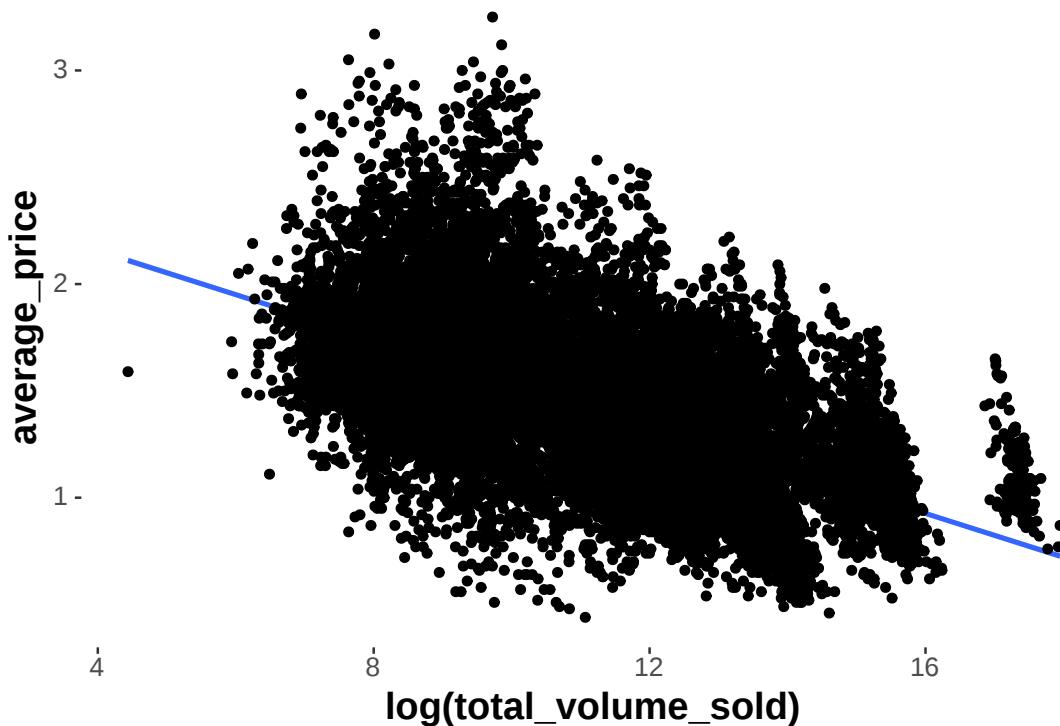
Notice that layering is really sequential. To see this, just check what happens when we reverse the calls of the `geom_` functions in the previous example:

```
avocado_data %>%
  ggplot(
    mapping = aes(
      # notice that we use the log (try without it to understand why)
```

```

x = log(total_volume_sold),
y = average_price
)
) +
# FIRST: add a linear regression line
geom_smooth(method = "lm") +
# THEN: add a scatter plot
geom_point()

```



If you want lower layers to be visible behind layers added later, one possibility is to tinker with opacity, via the `alpha` parameter. Notice that the example below also changes the colors. The result is quite toxic, but at least you see the line underneath the semi-transparent points.

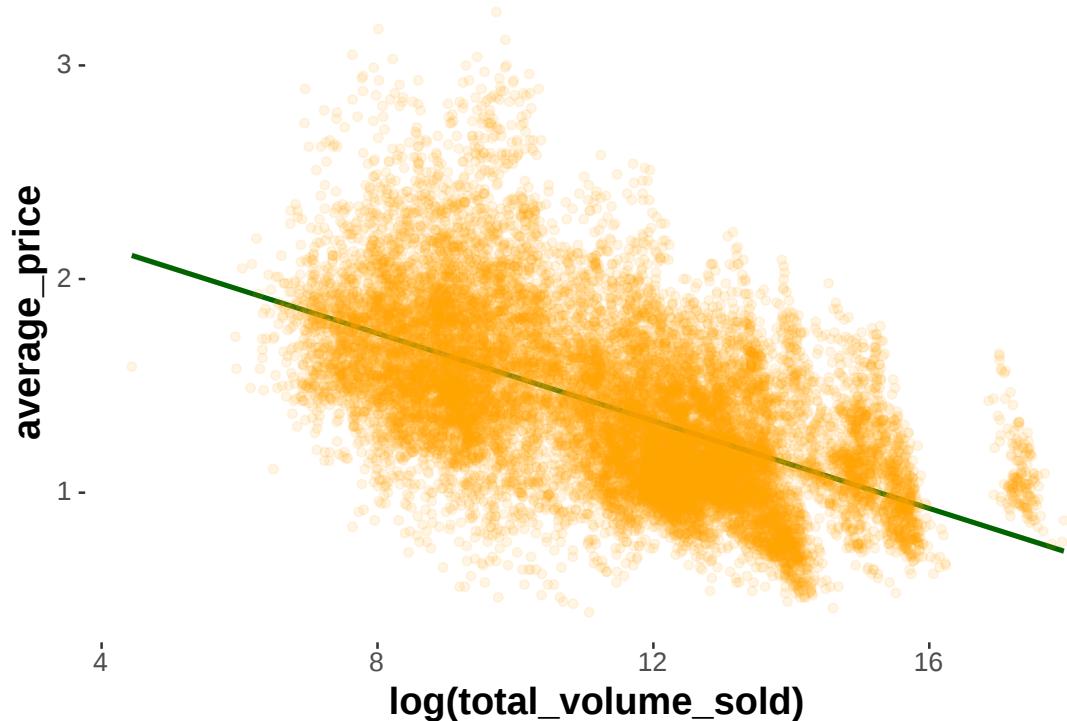
```

avocado_data %>%
  ggplot(
    mapping = aes(
      # notice that we use the log (try without it to understand why)
      x = log(total_volume_sold),
      y = average_price
    )
  )

```

6. Data Visualization

```
) +  
# FIRST: add a linear regression line  
geom_smooth(method = "lm", color = "darkgreen") +  
# THEN: add a scatter plot  
geom_point(alpha = 0.1, color = "orange")
```



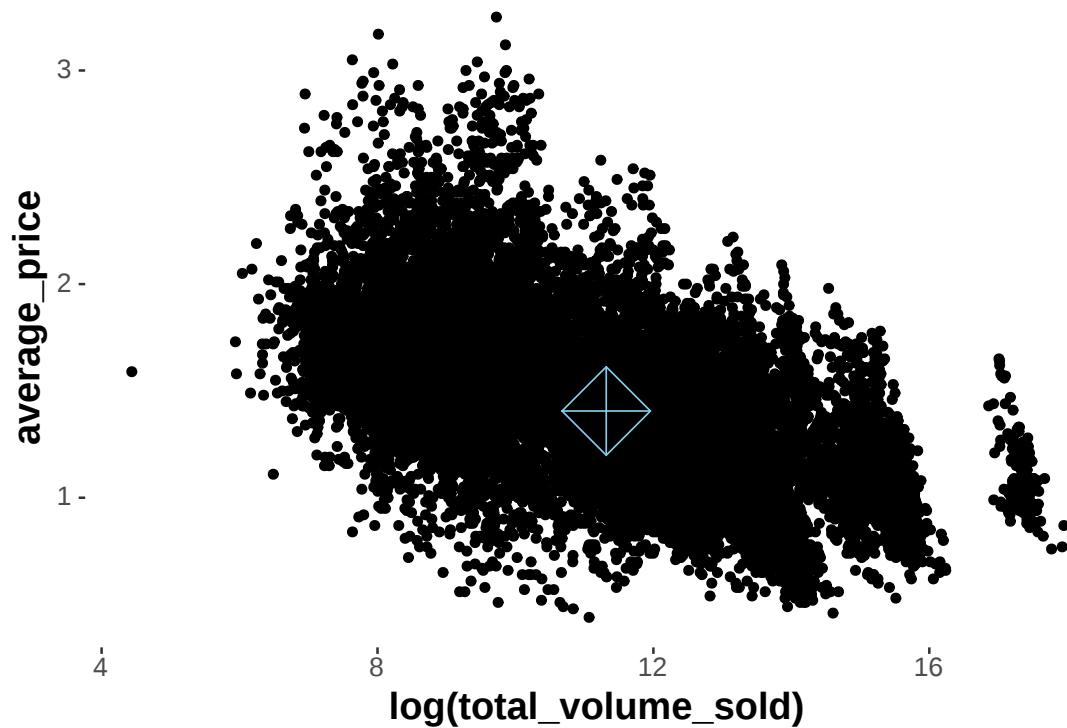
The aesthetics defined in the initial call to `ggplot` are global defaults for all layers to follow, unless they are overwritten. This also holds for the data supplied to `ggplot`. For example, we can create a second layer using another call to `geom_point` from a second data set (e.g., with a summary statistic), like so:

```
# create a small tibble with the means of both  
#   variables of interest  
avocado_data_means <-  
  avocado_data %>%  
  summarize(  
    mean_volume = mean(log(total_volume_sold)),  
    mean_price = mean(average_price)  
  )  
avocado_data_means
```

```
## # A tibble: 1 x 2
##   mean_volume mean_price
##       <dbl>      <dbl>
## 1        11.3      1.41

avocado_data %>%
  ggplot(
    aes(x = log(total_volume_sold),
        y = average_price)
  ) +
  # first layer uses globally declared data & mapping
  geom_point() +
  # second layer uses different data set & mapping
  geom_point(
    data = avocado_data_means,
    mapping = aes(
      x = mean_volume,
      y = mean_price
    ),
    # change shape of element to display (see below)
    shape = 9,
    # change size of element to display
    size = 12,
    color = "skyblue"
  )
```

6. Data Visualization

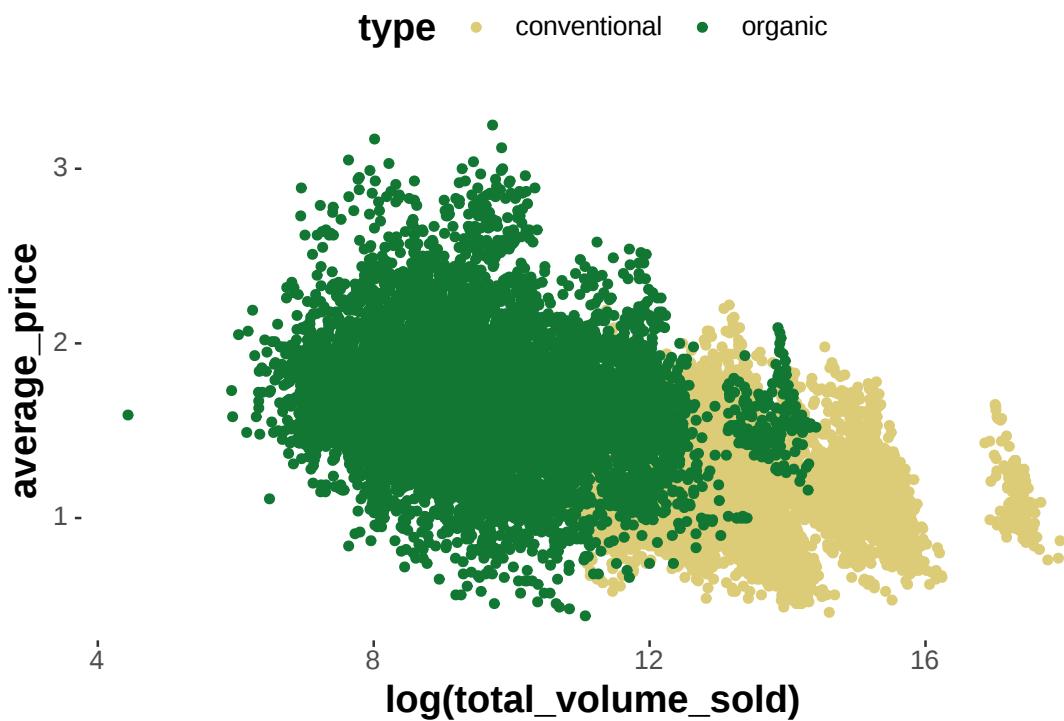


6.3.4. Grouping

Categorical distinction are frequently important in data analysis. Just think of the different combinations of factor levels in a factorial design, or the difference between conventionally grown and organically grown avocados. `ggplot` understands grouping very well and acts on appropriately, if you tell it to in the right way.

Grouping can be relevant for different aspects of a plot: the color of points or lines, their shape, or even whether to plot everything together or separately. For instance, we might want to display different types of avocados in different color. We can do this like so:

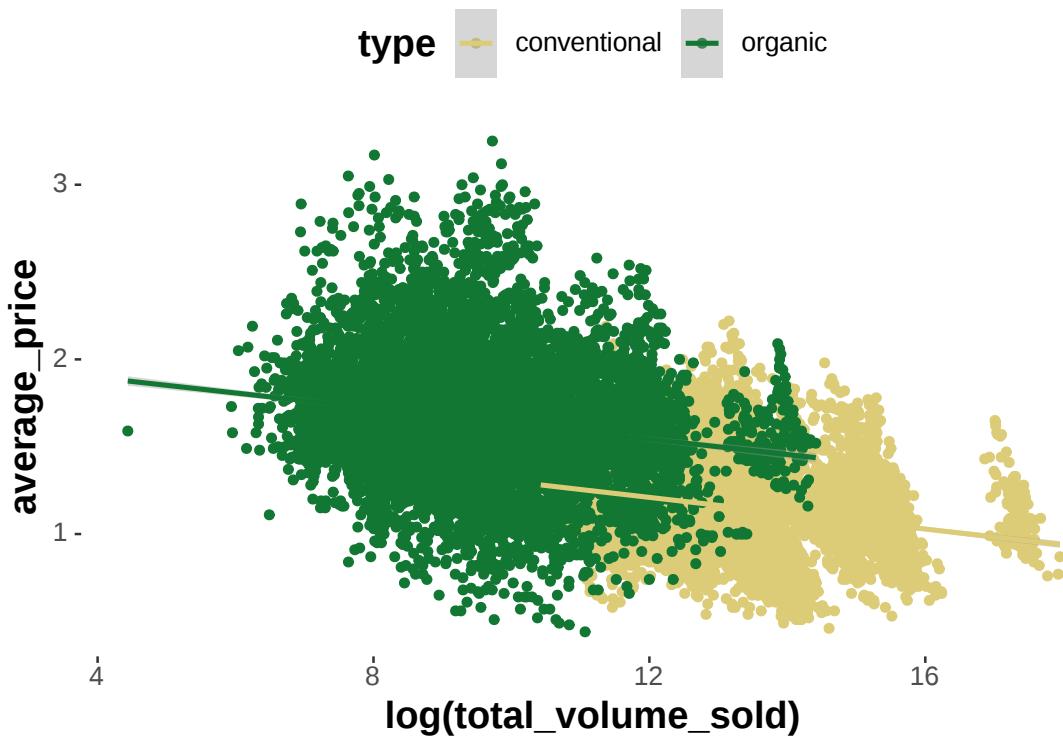
```
avocado_data %>%
  ggplot(
    aes(
      x = log(total_volume_sold),
      y = average_price,
      # use a different color for each type of avocado
      color = type
    )
  ) +
  geom_point(aes(color = type))
```



Notice that we added the grouping information inside of `aes` to the call of `ggplot`. This way the grouping is the global default for the whole plot. Check what happens when we then add another layer, like `geom_smooth`:

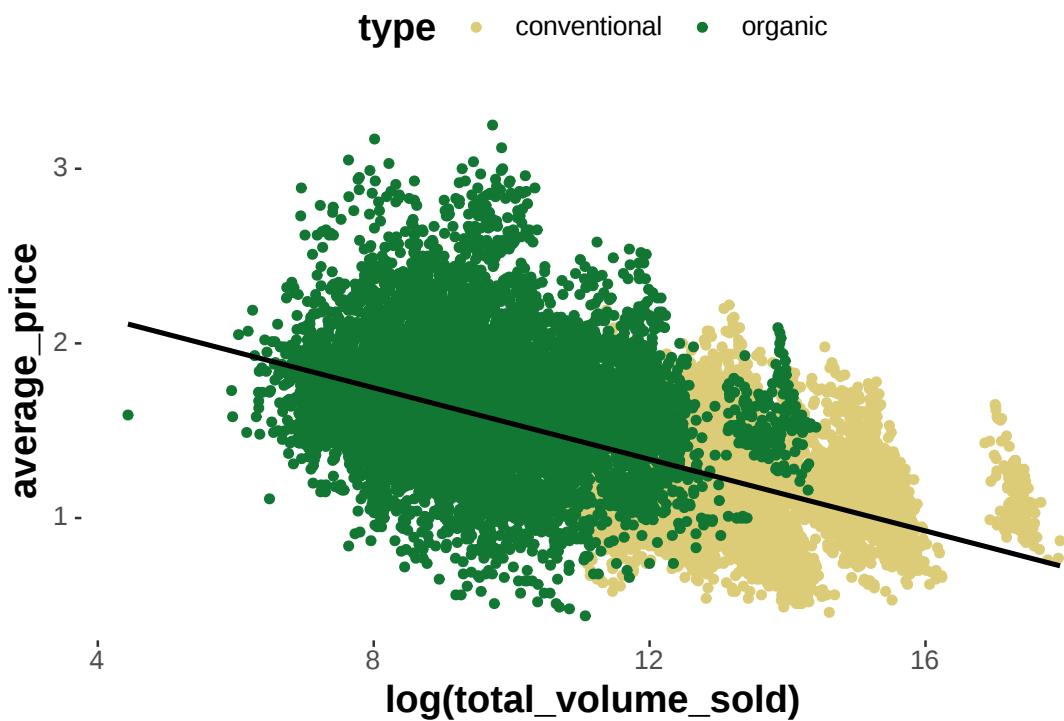
```
avocado_data %>%
  ggplot(
    aes(
      x = log(total_volume_sold),
      y = average_price,
      # use a different color for each type of avocado
      color = type
    )
  ) +
  geom_point(aes(color = type)) +
  geom_smooth(method = "lm")
```

6. Data Visualization



The regression lines will also be shown in the colors of the underlying scatter plot. We can change this by overwriting the `color` attribute locally, but then we loose the grouping information:

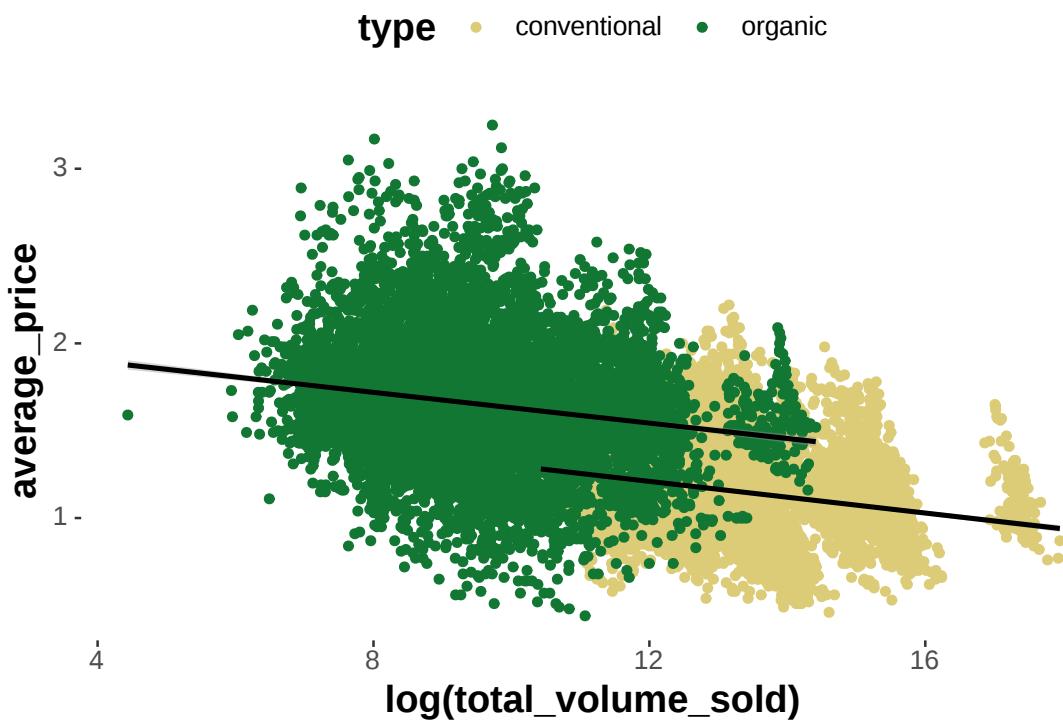
```
avocado_data %>%
  ggplot(
    aes(
      x = log(total_volume_sold),
      y = average_price,
      # use a different color for each type of avocado
      color = type
    )
  ) +
  geom_point(aes(color = type)) +
  geom_smooth(method = "lm", color = "black")
```



To retrieve the grouping information, we can change the explicit keyword `group` (which just treats data from the relevant factor levels differently without directly changing their appearance):

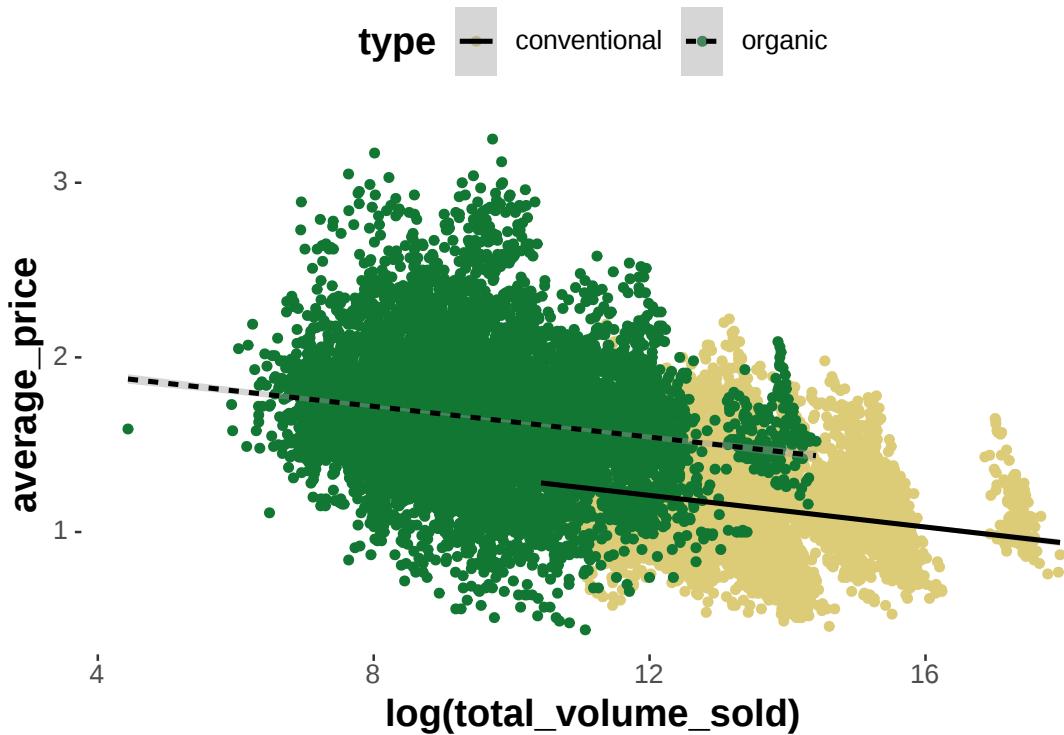
```
avocado_data %>%
  ggplot(
    aes(
      x = log(total_volume_sold),
      y = average_price,
      # use a different color for each type of avocado
      color = type
    )
  ) +
  geom_point(aes(color = type)) +
  geom_smooth(
    # tell the smoother to deal with avocados types separately
    aes(group = type),
    method = "lm",
    color = "black"
  )
```

6. Data Visualization



Finally, we see that the lines are not uniquely associative with the avocado type, so we can also change the regression line's shape attribute conditional on avocado type:

```
avocado_data %>%
  ggplot(
    aes(
      x = log(total_volume_sold),
      y = average_price,
      # use a different color for each type of avocado
      color = type
    )
  ) +
  geom_point(aes(color = type)) +
  geom_smooth(
    # tell the smoother to deal with avocados types separately
    aes(group = type, linetype = type),
    method = "lm",
    color = "black"
  )
```



6.3.5. Example of a customized plot

If done with the proper mind and heart, plots intended to share (and to communicate a point, following the idea of hypothesis-driven visualization) will usually require a lot of tweaking. We will cover some of the most frequently relevant tweaks in Section 6.6.

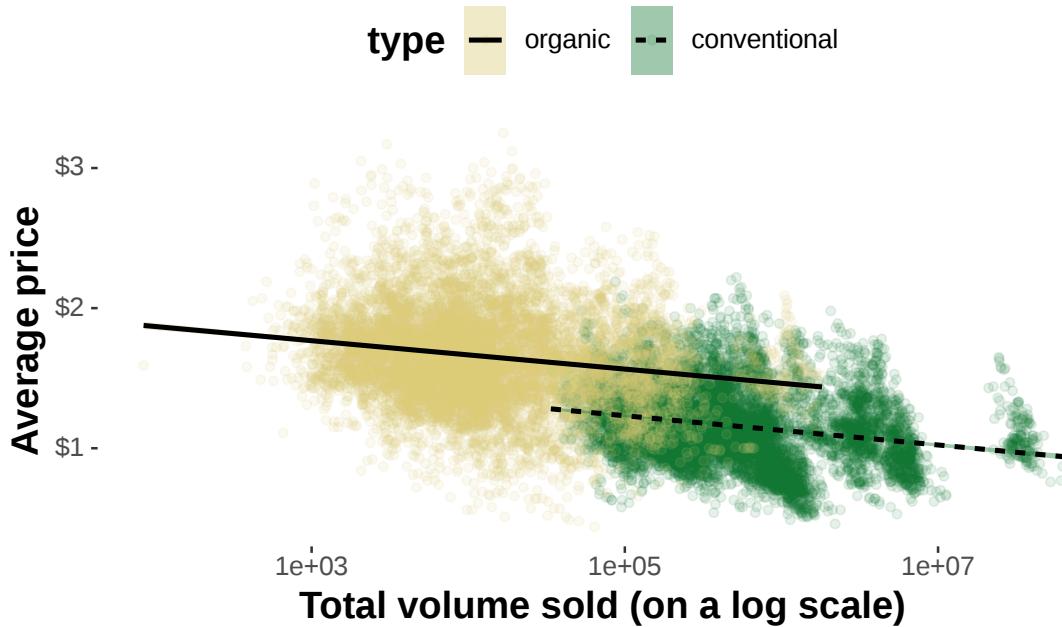
To nevertheless get a feeling of where the journey is going, at least roughly, here is an example of a plot of the avocado data which is much more tweaked and honed. No claim is intended regarding the false idea that this plot is in any sense optimal. There is not even a clear hypothesis or point to communicate. This just showcases some functionality. Notice, for instance, that this plot uses two layers, invoked by `geom_point` which shows the scatter plot of points and `geom_smooth` which layers on top the point cloud regression lines (one for each level in the grouping variable).

```
# pipe data set into function `ggplot`  
avocado_data %>%  
  # reverse factor level so that horizontal legend entries align with  
  # the majority of observations of each group in the plot  
  mutate(  
    type = fct_rev(type)  
  ) %>%  
  # initialize the plot
```

6. Data Visualization

```
ggplot(  
  # defined mapping  
  mapping = aes(  
    # which variable goes on the x-axis  
    x = total_volume_sold,  
    # which variable goes on the y-axis  
    y = average_price,  
    # which groups of variables to distinguish  
    group = type,  
    # color and fill to change by grouping variable  
    fill = type,  
    linetype = type,  
    color = type  
  )  
) +  
# declare that we want a scatter plot  
geom_point(  
  # set low opacity for each point  
  alpha = 0.1  
) +  
# add a linear model fit (for each group)  
geom_smooth(  
  color = "black",  
  method = "lm"  
) +  
# change the default (normal) of x-axis to log-scale  
scale_x_log10() +  
# add dollar signs to y-axis labels  
scale_y_continuous(labels = scales::dollar) +  
# change axis labels and plot title & subtitle  
labs(  
  x = 'Total volume sold (on a log scale)',  
  y = 'Average price',  
  title = "Avocado prices against amount sold",  
  subtitle = "With linear regression lines"  
)
```

Avocado prices against amount sold With linear regression lines



6.4. A rendezvous with popular geoms

In the following we will cover some of the more basic `geom_` functions relevant for our present purposes. It might be useful to read this section top-to-bottom at least once, not to think of it as a mere reference list. More information is provided by the ggplot cheat sheet.

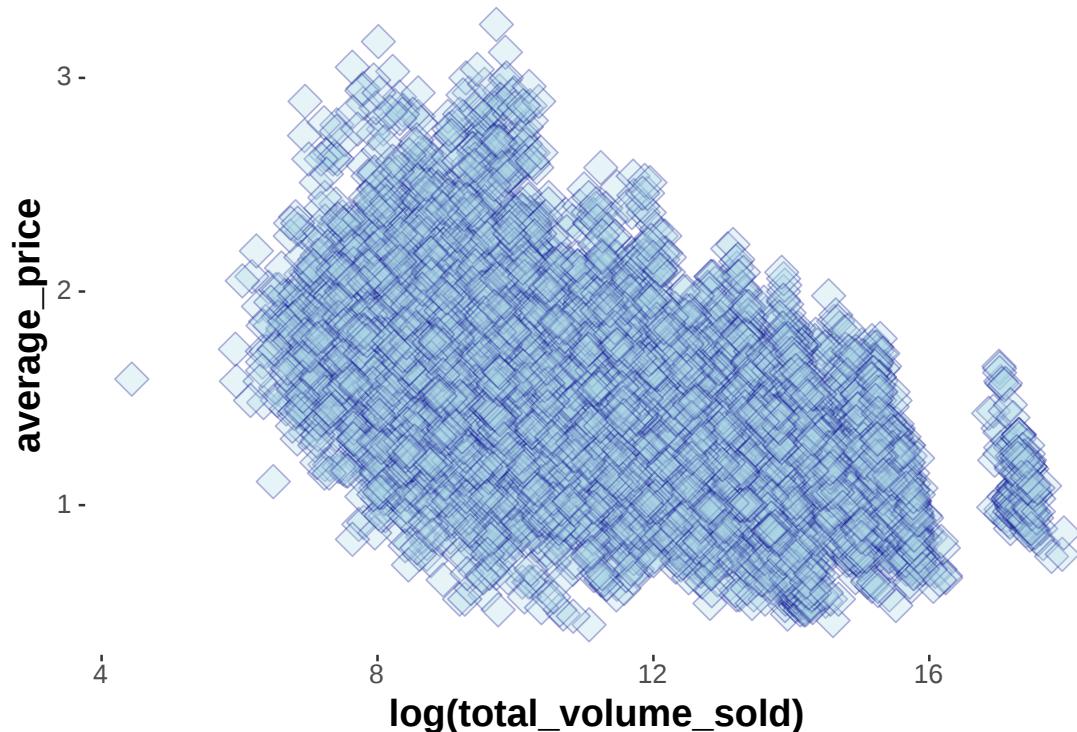
6.4.1. Scatter plots with `geom_point`

Scatter plots visualize pairs of associated observations as points in space. We have seen this for the avocado prize data above. Let's look at some of the further arguments we can use to tweak the presentation by `geom_point`. The following example changes the shape of the objects displayed to tilted rectangles (sometimes called diamonds, e.g., in LaTeX \diamond) away from the default circles, the color of the shapes, their size and opacity.

```
avocado_data %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(
    # shape to display is number 23 (tilted rectangle, see below)
    shape = 23,
```

6. Data Visualization

```
# color of the surrounding line of the shape (for shapes 21-24)
color = "darkblue",
# color of the interior of each shape
fill = "lightblue",
# size of each shape (default is 1)
size = 5,
# level of opacity for each shape
alpha = 0.3
)
```



How do you know which shape is which number? - By looking at the picture in Figure 6.5, for instance.

6.4.2. Smooth

The `geom_smooth` function operates on two-dimensional metric data and outputs a smoothed line, using different kinds of fitting functions. It is possible to show an indicator of certainty for the fit. We will deal with model fits in later parts of the book. For illustration just enjoy a few examples here:

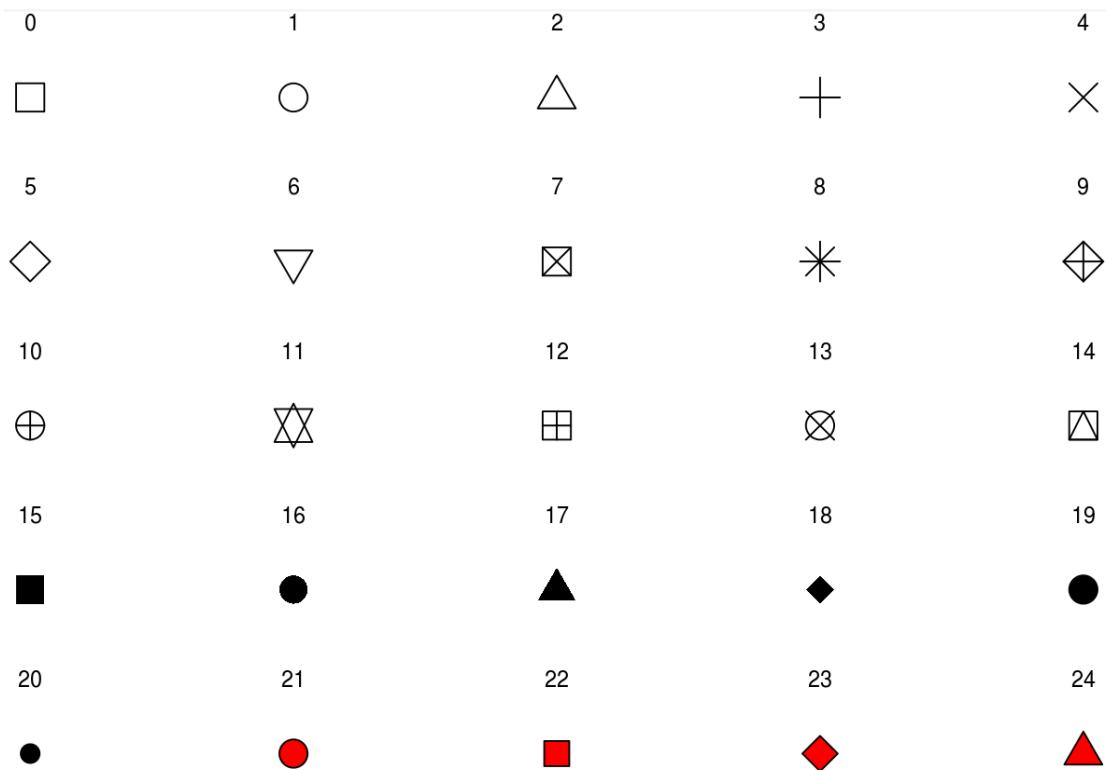
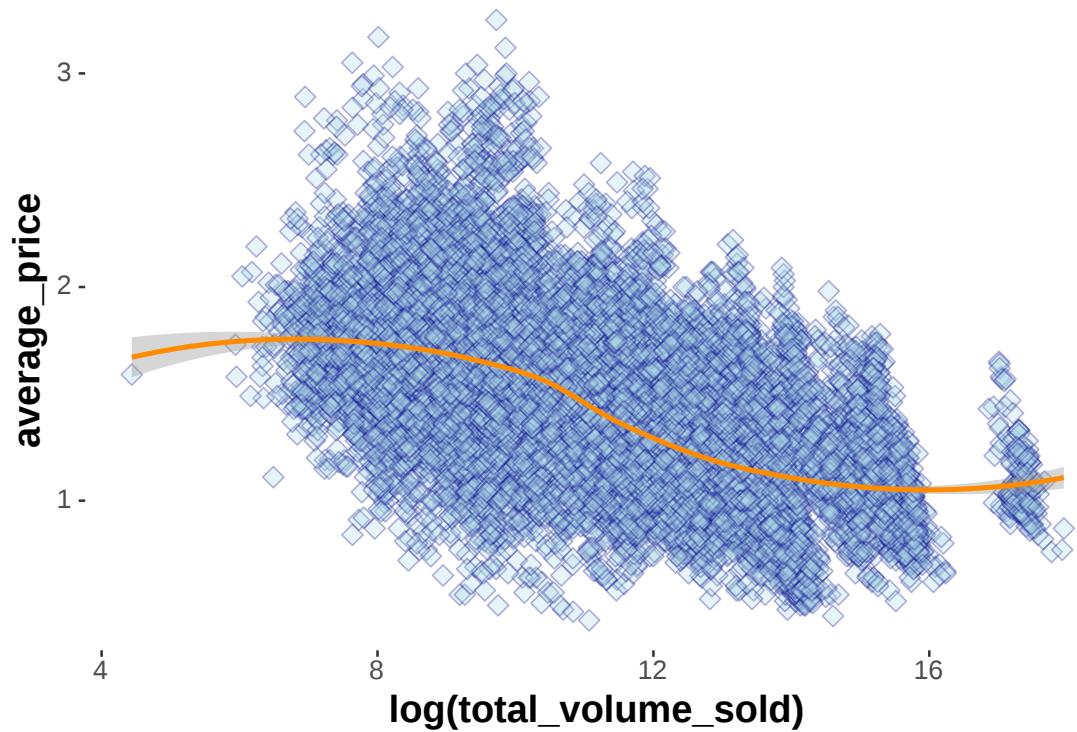


Figure 6.5.: The numerical coding of different shapes in ‘ggplot’ Notice that objects 21-24 are sensitive to both ‘color’ and ‘fill’, but the others are only sensitive to ‘color’.

6. Data Visualization

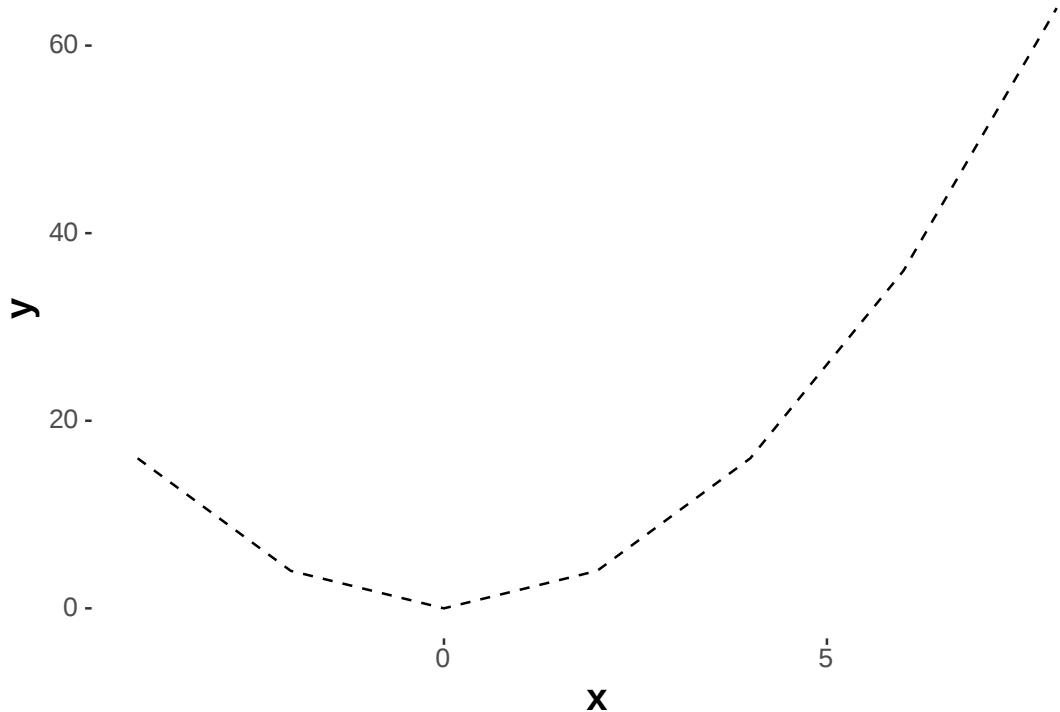
```
avocado_data %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(
    shape = 23,
    color = "darkblue",
    fill = "lightblue",
    size = 3,
    alpha = 0.3
  ) +
  geom_smooth(
    # fitting a smoothed curve to the data
    method = "loess",
    # display standard error around smoothing curve
    se = T,
    color = "darkorange"
  )
```



6.4.3. Line

Use `geom_line` to display a line for your data if that data has associated (ordered) metric values. You can use argument `linetype` to specify the kind of line to draw.

```
tibble(
  x = seq(-4,8, by = 2),
  y = x^2
) %>%
  ggplot(aes(x,y)) +
  geom_line(
    linetype = "dashed"
)
```

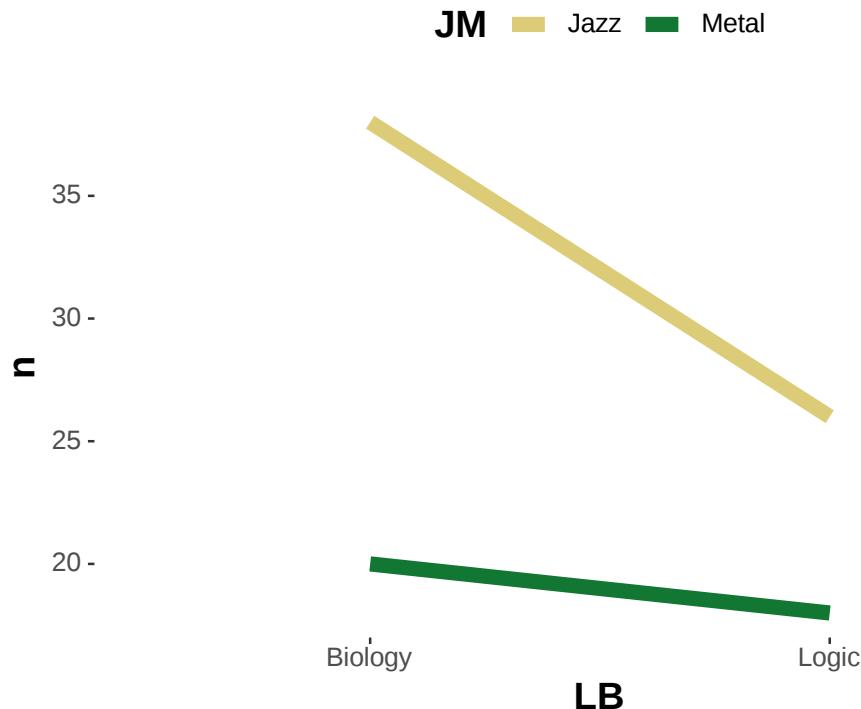


Sometimes you may want to draw lines between items that are grouped:

```
BLJM_associated_counts %>%
  ggplot(
    aes(
      x = LB,
      y = n,
```

6. Data Visualization

```
    color = JM,  
    group = JM  
)  
) +  
geom_line(size = 3)
```



6.4.4. Barplot

A barplot, plotted with `geom_bar` or `geom_col`, displays a single number for each of several groups for visual comparison by length. The difference between these two functions is that `geom_bar` relies on an implicit counting, while `geom_col` expects the numbers that translate into the length of the bars to be supplied for it. This book favors the use of `geom_col` by first wrangling the data to show the numbers to be visualized, since often this is the cleaner approach and the numbers are useful to have access to independently (e.g., for referring to in the text).

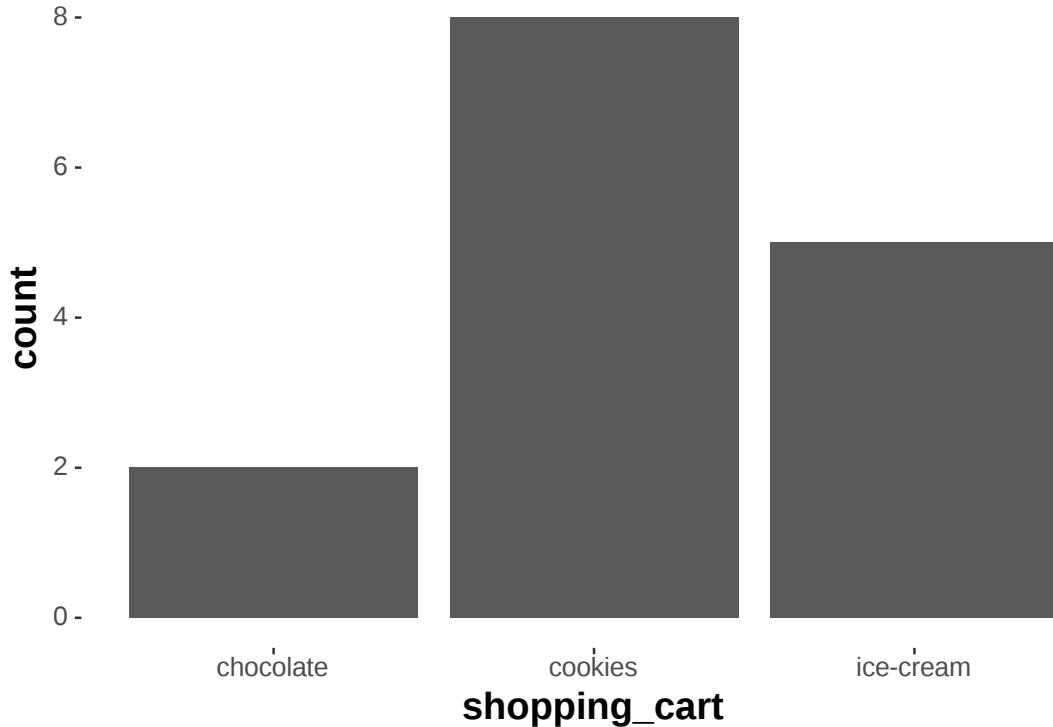
Here's an example of how `bar_plot` works (implicitly counting numbers of occurrences):

```
tibble(  
  shopping_cart = c(  
    rep("chocolate", 2),
```

```

    rep("ice-cream", 5),
    rep("cookies", 8)
)
) %>%
ggplot(aes(x = shopping_cart)) +
geom_bar()

```



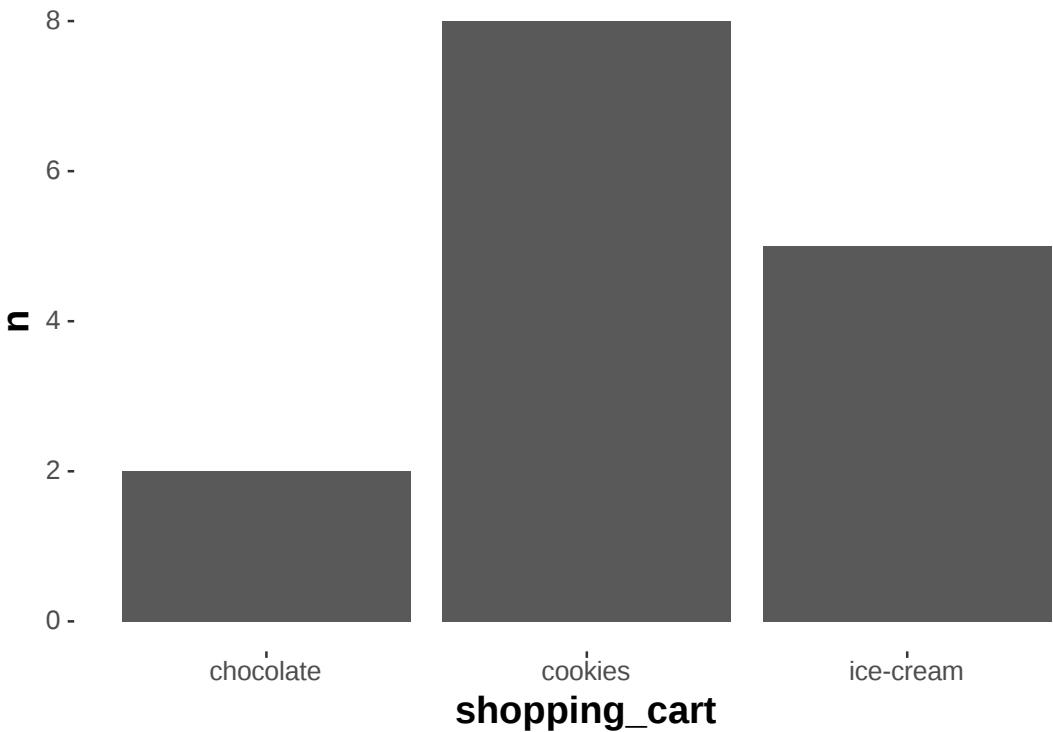
To display this data with `geom_col` we need to count occurrences first ourselves:

```

tibble(
  shopping_cart = c(
    rep("chocolate", 2),
    rep("ice-cream", 5),
    rep("cookies", 8)
  )
) %>%
dplyr::count(shopping_cart) %>%
  ggplot(aes (x = shopping_cart, y = n) ) +
  geom_col()

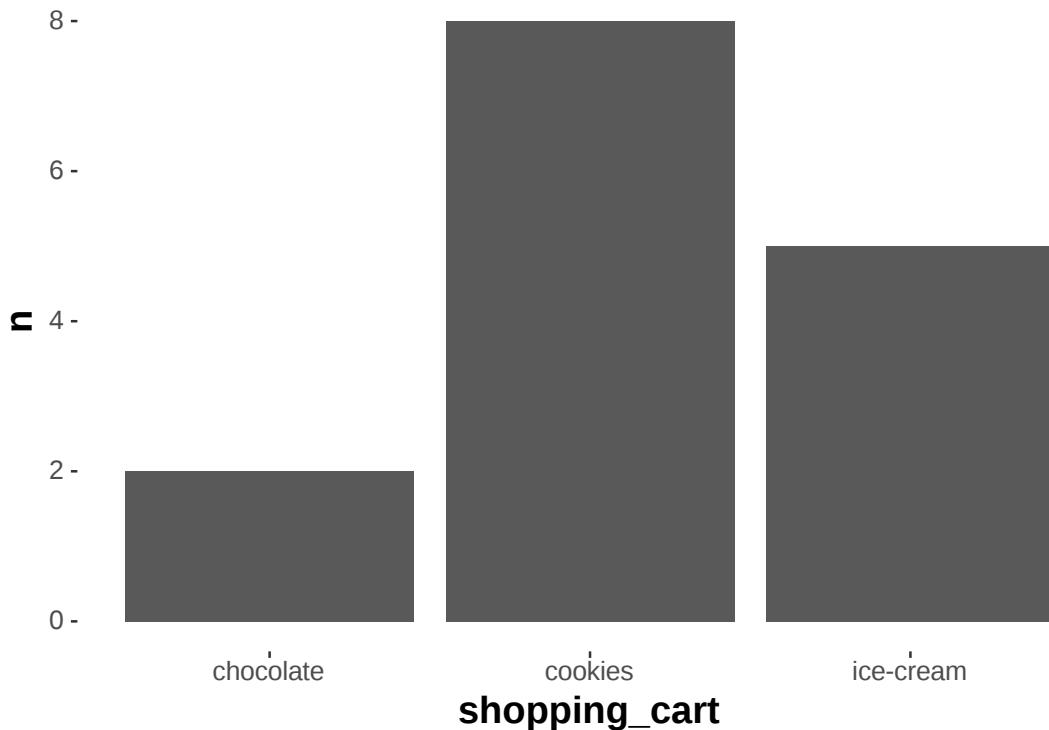
```

6. Data Visualization



To be clear, `geom_col` is essentially `geom_bar` when we overwrite the default statistical transformation of counting to “identity”:

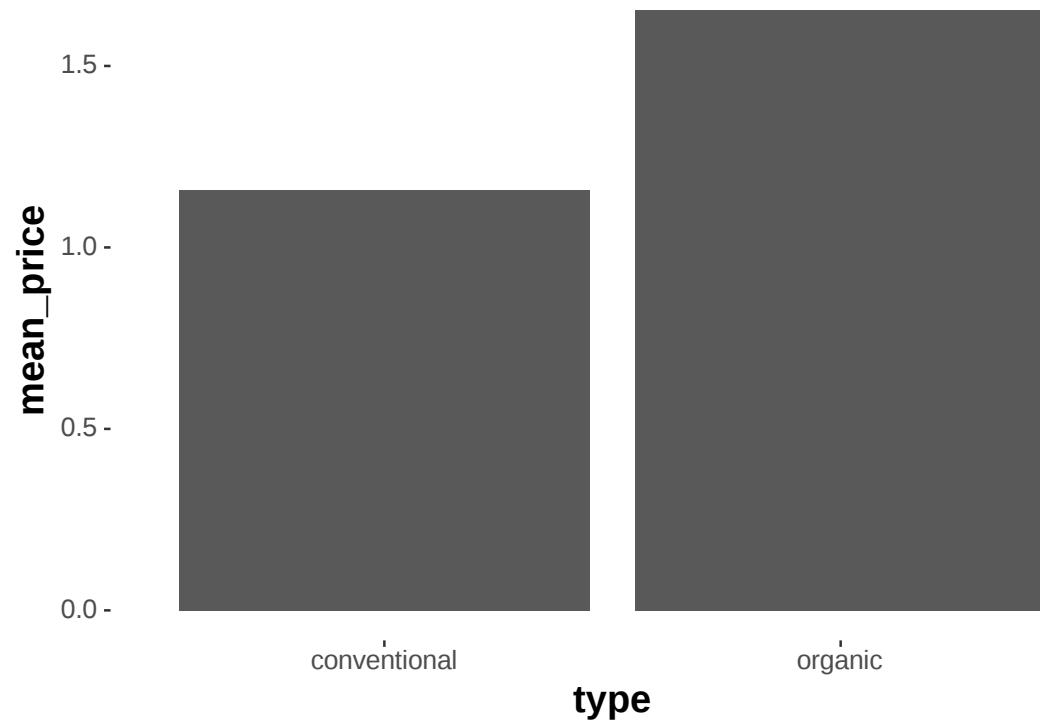
```
tibble(
  shopping_cart = c(
    rep("chocolate", 2),
    rep("ice-cream", 5),
    rep("cookies", 8)
  )
) %>%
  dplyr::count(shopping_cart) %>%
  ggplot(aes (x = shopping_cart, y = n) ) +
  geom_bar(stat = "identity")
```



Barplots are a frequent sight in psychology papers. They are also controversial. They often fare badly with respect to the data-ink ratio. Especially, when what is plotted are means of grouped variables. For example, the following plot is rather uninformative (even if the research question is a comparison of means):

```
avocado_data %>%
  group_by(type) %>%
  summarise(
    mean_price = mean(average_price)
  ) %>%
  ggplot(aes(x = type, y = mean_price)) +
  geom_col()
```

6. Data Visualization



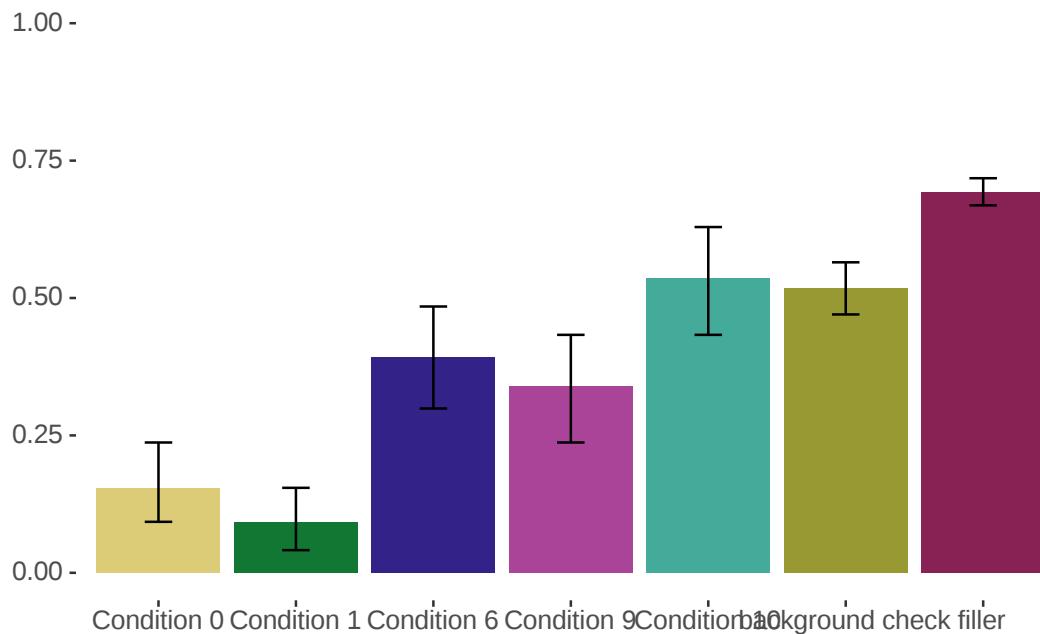
It makes sense to use the available space for a more informative report about the distribution of data points around the means, e.g., by using `geom_violin` or `geom_histogram` etc.

But barplots may also be good enough if there is not more of immediate relevance, such as when we look at counts or proportions. Still, it might help to include a measure of certainty. For instance, using the King of France data set, we can display proportions of 'true' answers with 95% bootstrapped confidence intervals like in the plot below. Notice the use of the `geom_errorbar` function to display the intervals in the following example.

```
data_KoF_processed %>%
  # drop unused factor levels
  droplevels() %>%
  # get means and 95% bootstrapped CIs for each condition
  group_by(condition) %>%
  nest() %>%
  summarise(
    CIs = map(data, function(d) bootstrapped_CI(d$response == "TRUE")))
  ) %>%
  unnest(CIs) %>%
  # plot means and CIs
  ggplot(aes(x = condition, y = mean, fill = condition)) +
  geom_col() +
```

```
geom_errorbar(aes(ymin = lower, ymax = upper, width = 0.2)) +
  ylim(0,1) +
  ylab("") + xlab("") + ggtitle("Proportion of 'TRUE' responses per condition") +
  theme(legend.position = "none") +
  scale_fill_manual(values = project_colors)
```

Proportion of 'TRUE' responses per condition



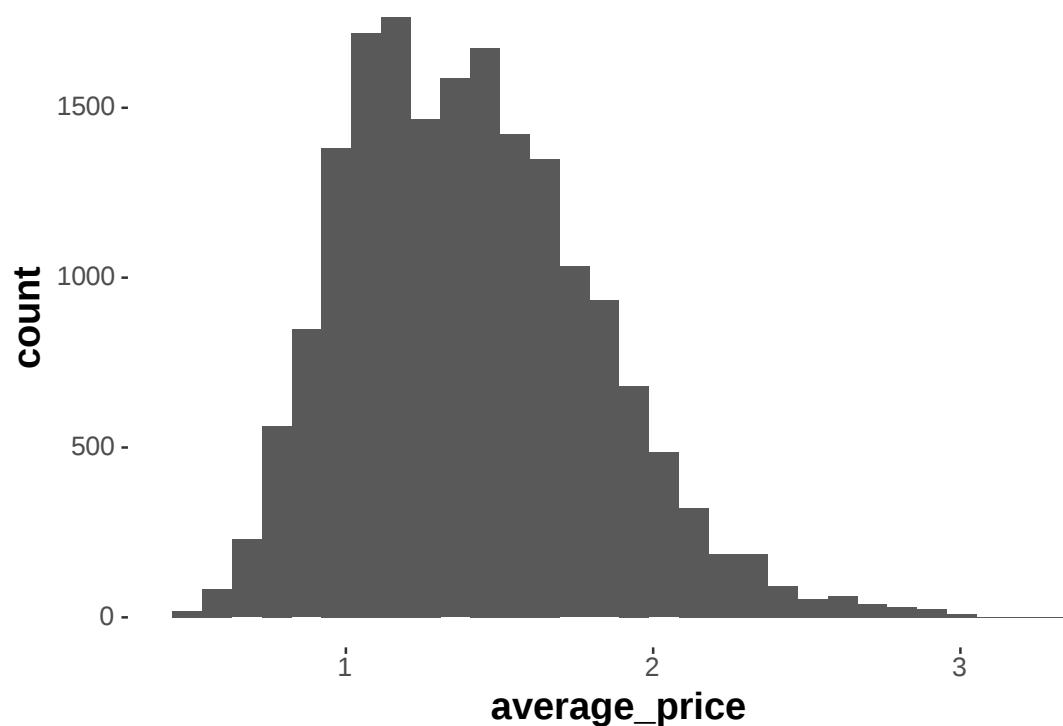
6.4.5. Plotting distributions: histograms, boxplots, densities and violins

There are different ways for plotting the distribution of observations in a one-dimensional vector, each with its own advantages and disadvantages: the histogram, a box plot, a density plot, and a violin plot. Let's have a look at each, based on the `average_price` of different types of avocados.

The histogram displays the number of occurrences of observations inside of prespecified bins. By default the function `geom_histogram` uses 30 equally spaced bins to display counts of your observations.

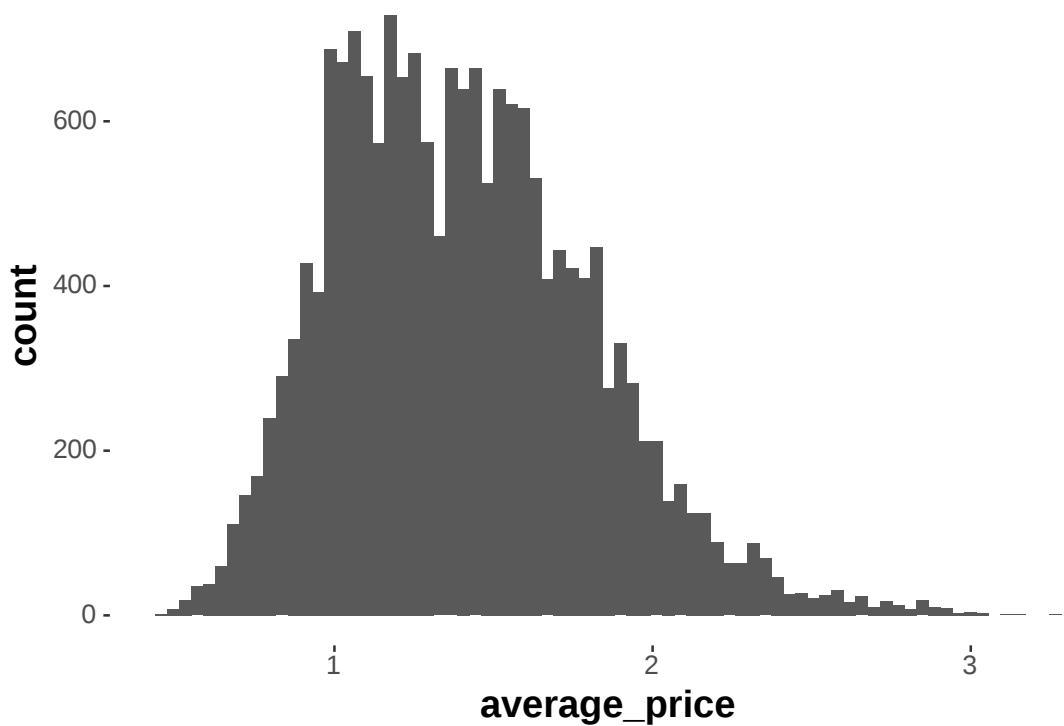
```
avocado_data %>%
  ggplot(aes(x = average_price)) +
  geom_histogram()
```

6. Data Visualization



If we specify more bins, we get a more fine-grained picture. (But notice that such a high number of bins works for the present data set, which has many observations, but it would not necessarily work for a small data set.)

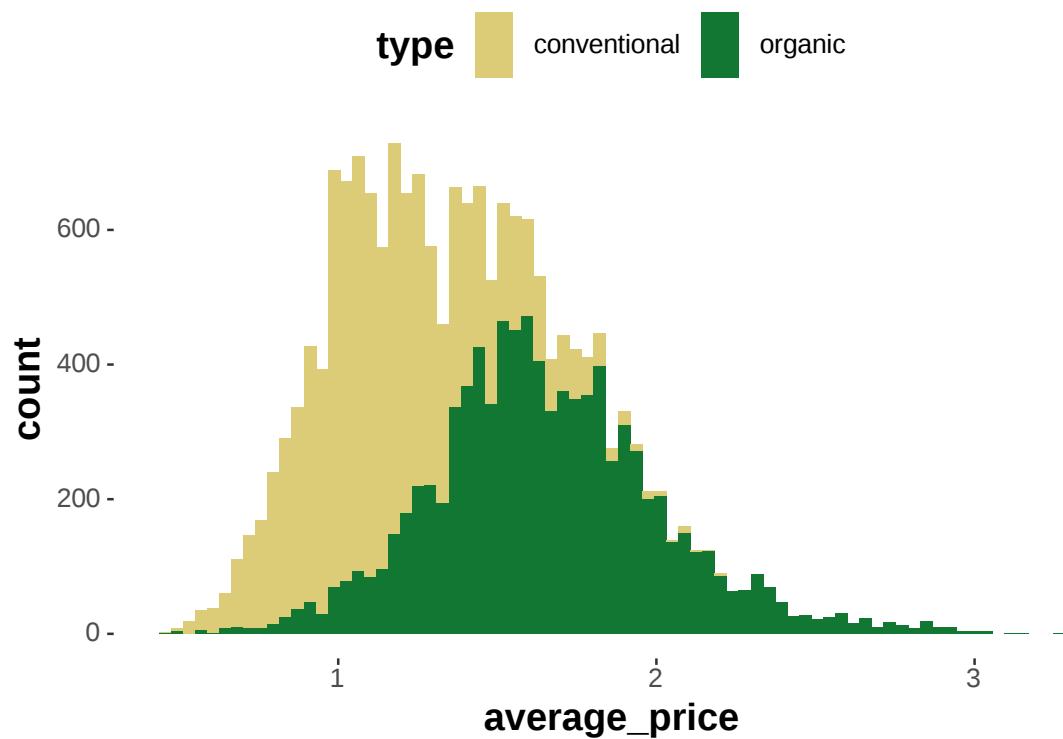
```
avocado_data %>%
  ggplot(aes(x = average_price)) +
  geom_histogram(bins = 75)
```



We can also layer histograms but this is usually a bad idea (even if we tinker with opacity) because a higher layer might block important information from a lower layer:

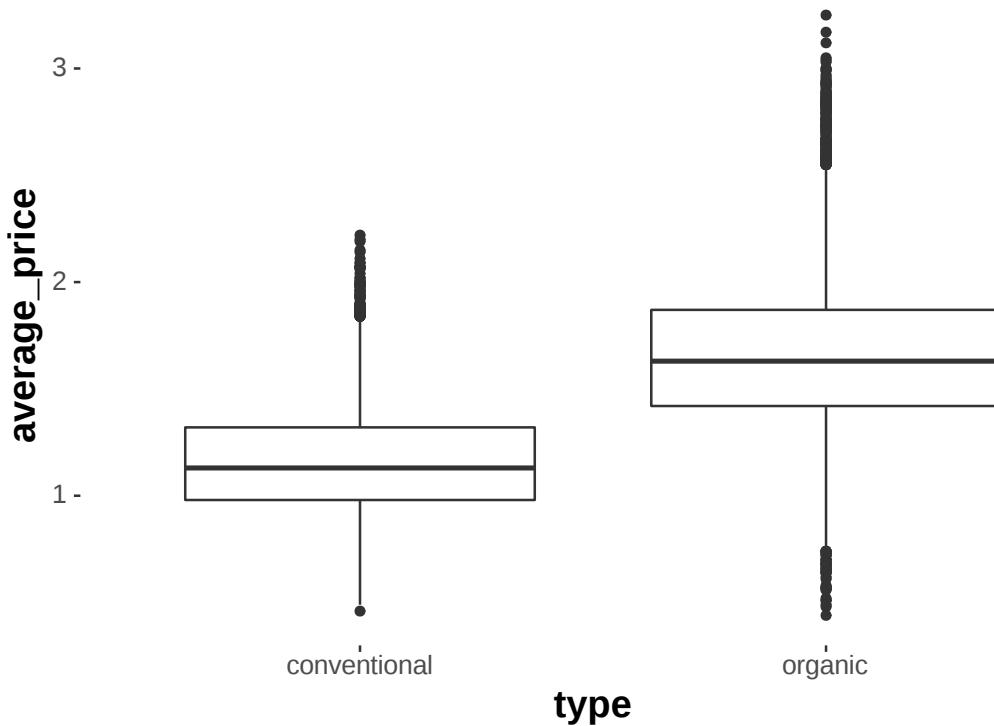
```
avocado_data %>%
  ggplot(aes(x = average_price, fill = type)) +
  geom_histogram(bins = 75)
```

6. Data Visualization



An alternative display of distributional metric information is a **box plot**. Box plots are classics, also called *box-and-whiskers plots*, and they basically visually report key summary statistics of your metric data. These do work much better than histograms for direct comparison:

```
avocado_data %>%
  ggplot(aes(x = type , y = average_price)) +
  geom_boxplot()
```



What we see here is the median for each group (thick black line) and the 25% and 75% quantiles (boxes). The straight lines show the range from the 25% or 75% quantiles to the values given by $\text{median} + 1.58 * \text{IQR} / \sqrt{n}$, where the IQR is the “interquartile range”, i.e., the range between the 25% and 75% quantiles (boxes).

To get a better picture of the shape of the distribution, `geom_density` uses a kernel estimate to show a smoothed line, roughly indicating ranges of higher density of observations with higher numbers. Using opacity, `geom_density` is useful also for the close comparison of distributions across different groups:

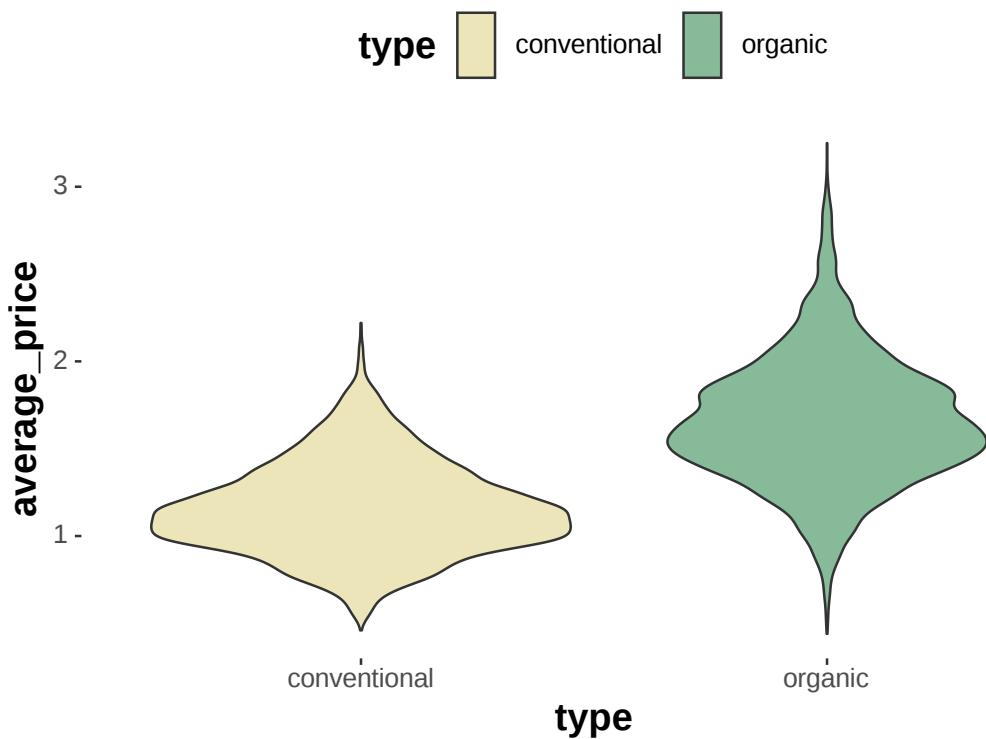
```
avocado_data %>%
  ggplot(aes(x = average_price, color = type, fill = type)) +
  geom_density(alpha = 0.5)
```

6. Data Visualization



For many groups to compare, density plots can become cluttered. **Violin plots** are like mirrored density plots and are better for comparison of multiple groups:

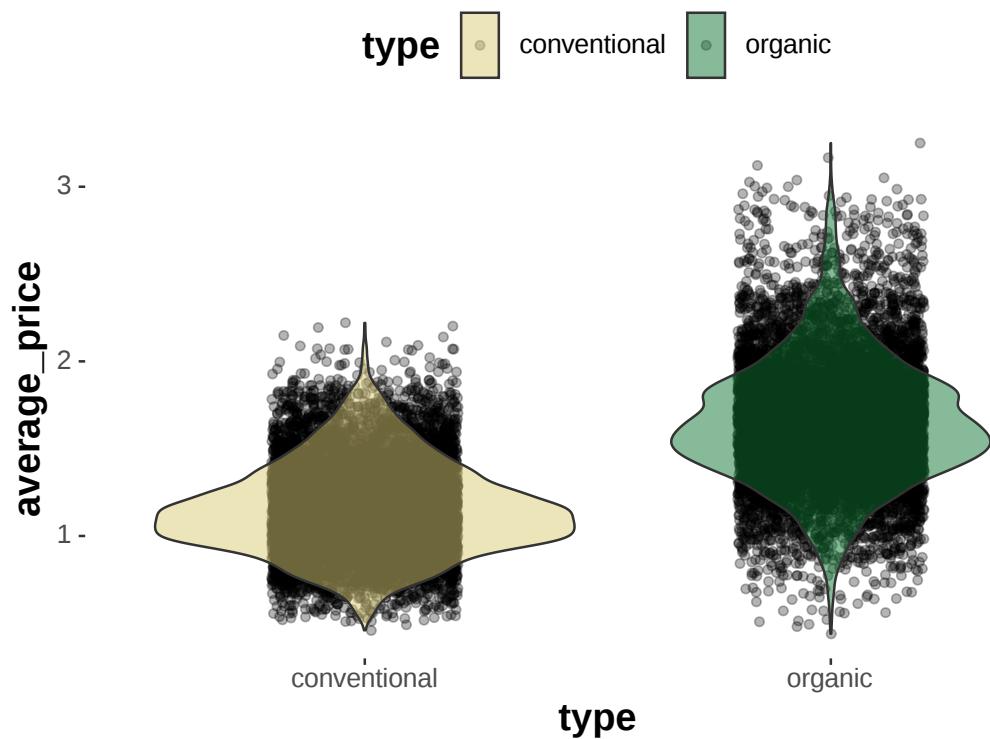
```
avocado_data %>%
  ggplot(aes(x = type, y= average_price, fill = type)) +
  geom_violin(alpha = 0.5)
```



A frequently seen method of visualization is to layer a jittered distribution of points under a violin plot, like so:

```
avocado_data %>%
  ggplot(aes(x = type, y= average_price, fill = type)) +
  geom_jitter(alpha = 0.3, width = 0.2) +
  geom_violin(alpha = 0.5)
```

6. Data Visualization

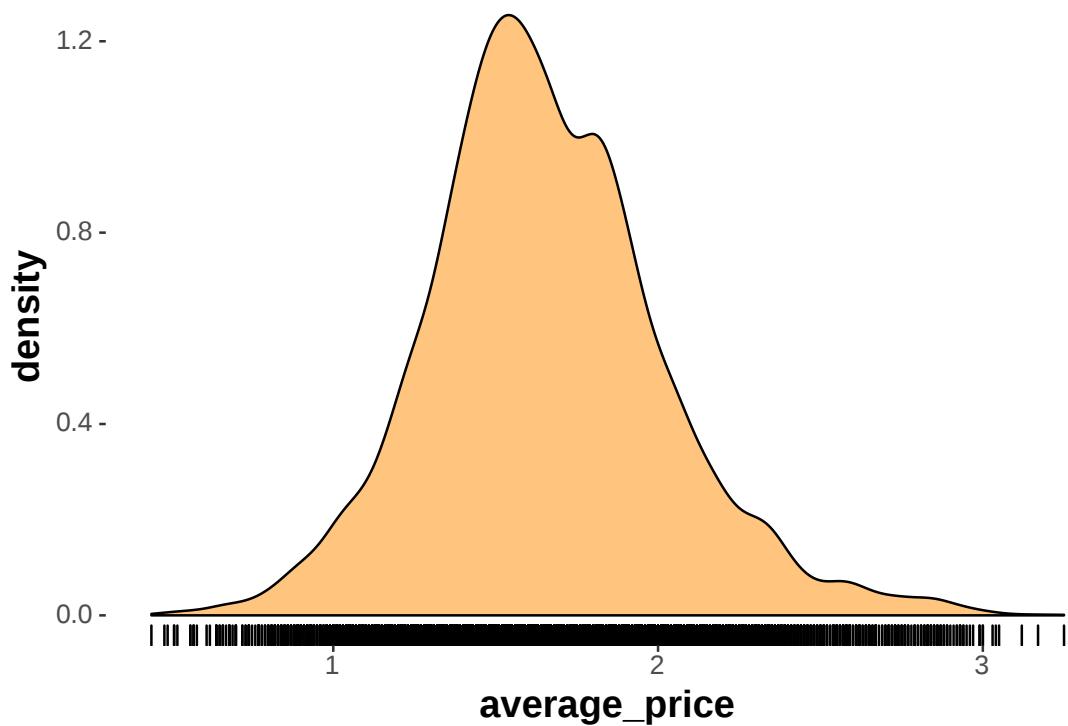


6.4.6. Rugs

Since plotting distributions, especially with high-level abstract smoothing as in `geom_density` and `geom_violin` fails to give information about the actual quantity of the data points, rug plots are useful additions to such plots. `geom_rug` add marks along the axes where different points lie.

Here is an example of `geom_rug` combined with `geom_density`:

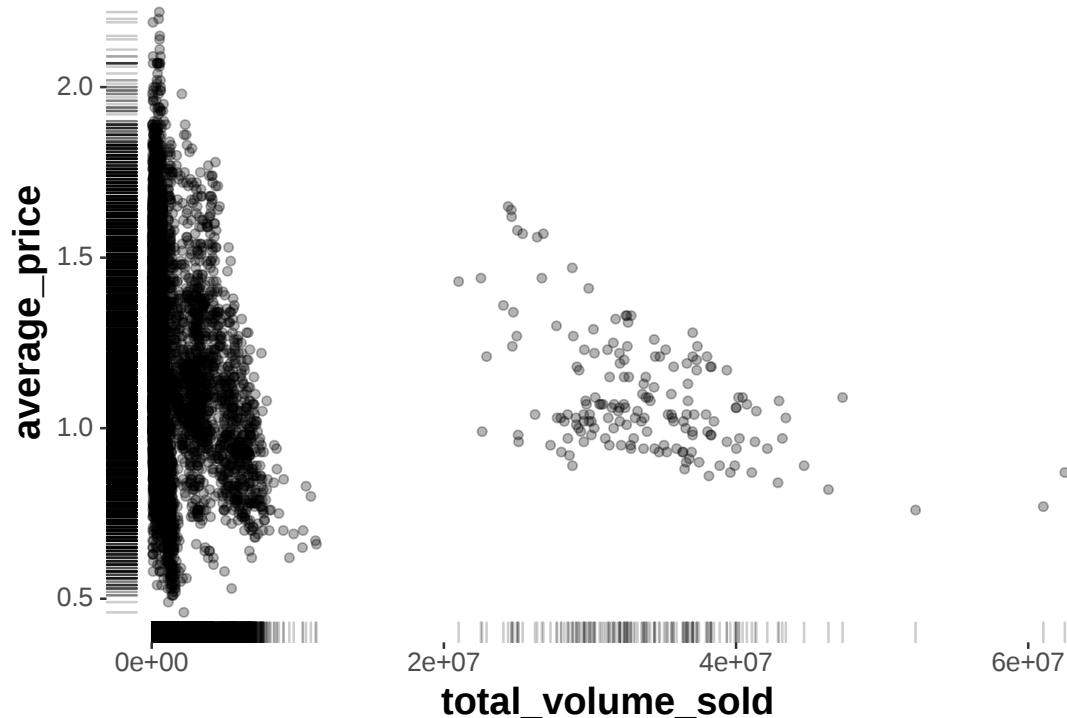
```
avocado_data %>%
  filter(type == "organic") %>%
  ggplot(aes(x = average_price)) +
  geom_density(fill = "darkorange", alpha = 0.5) +
  geom_rug()
```



Here are rugs on a two-dimensional scatter plot:

```
avocado_data %>%
  filter(type == "conventional") %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point(alpha = 0.3) +
  geom_rug(alpha = 0.2)
```

6. Data Visualization



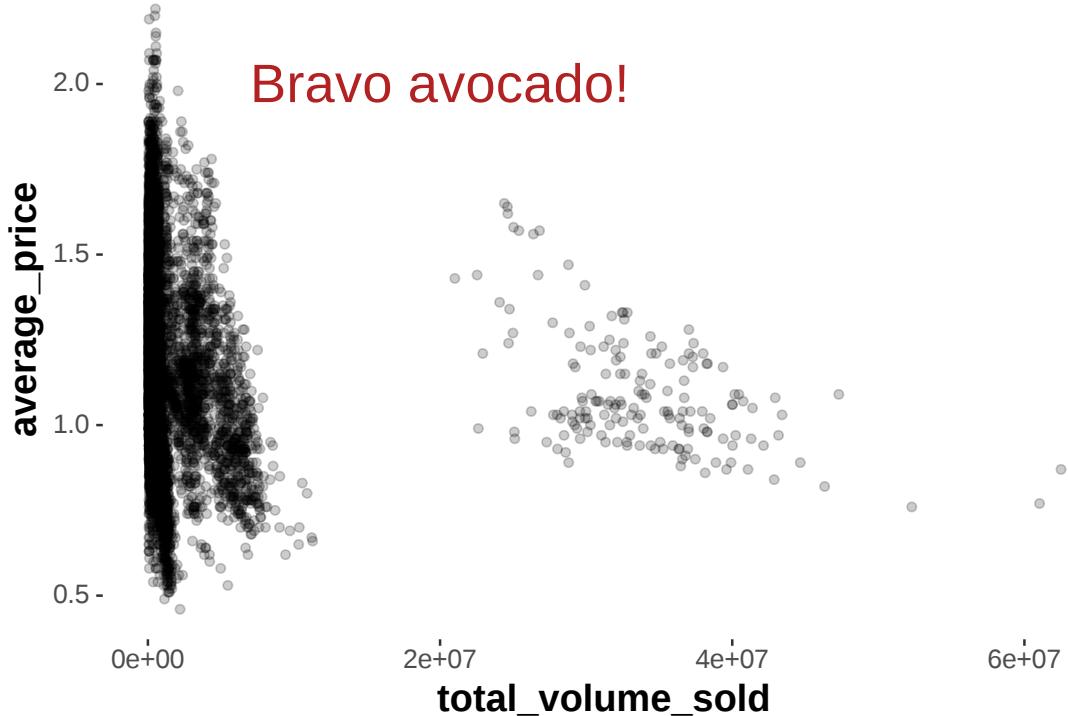
6.4.7. Annotation

It can be useful add further elements to a plot. We might want to add text, or specific geometrical shapes to highlight aspects of data. The most general function for doing this is `annotate`. The function `annotate` takes as a first argument a `geom` argument, e.g., `text` or `rectangle`. It is therefore not a wrapper function in the `geom_` family of functions, but the underlying function around which convenience functions like `geom_text` or `geom_rectangle` are wrapped. The further arguments that `annotate` expects depend on the `geom` it is supposed to realize.

Suppose we want to add textual information at a particular coordinate. We can do this with `annotate` as follows:

```
avocado_data %>%
  filter(type == "conventional") %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point(alpha = 0.2) +
  annotate(
    geom = "text",
    # x and y coordinates for the text
    x = 2e7,
    y = 2,
```

```
# text to be displayed
label = "Bravo avocado!",
color = "firebrick",
size = 8
)
```

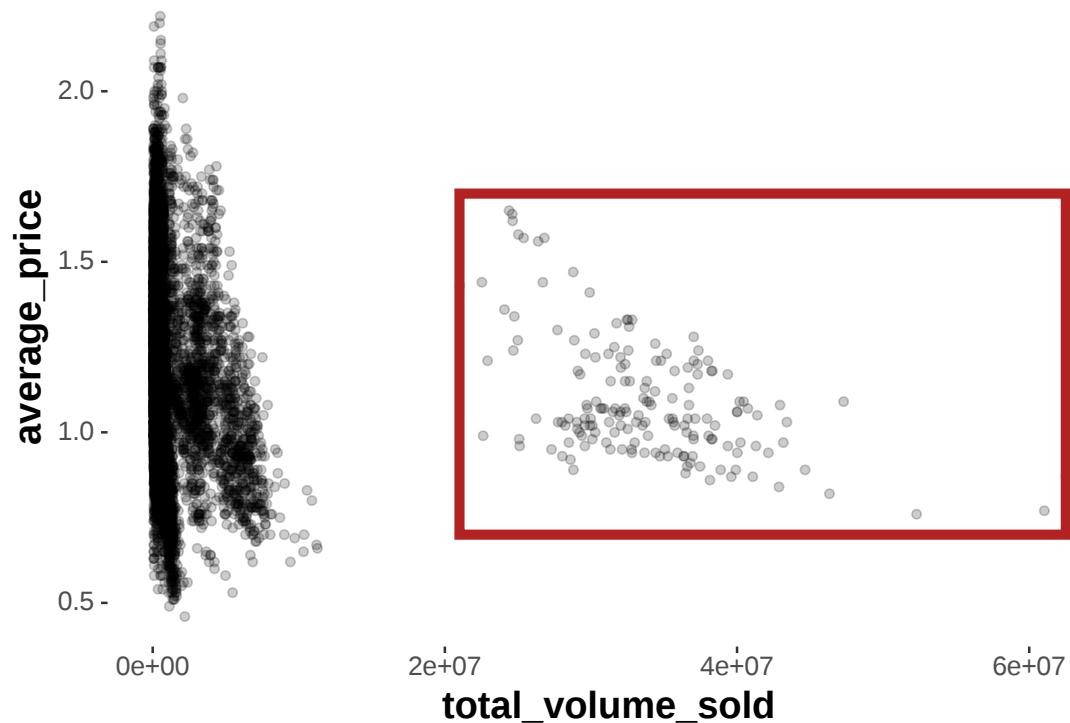


We can also single out some data points, like so:

```
avocado_data %>%
  filter(type == "conventional") %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point(alpha = 0.2) +
  annotate(
    geom = "rect",
    # coordinates for the rectangle
    xmin = 2.1e7,
    xmax = max(avocado_data$total_volume_sold) + 100,
    ymin = 0.7,
    ymax = 1.7,
    color = "firebrick",
```

6. Data Visualization

```
alpha = 0,  
size = 2  
)
```



6.5. Faceting

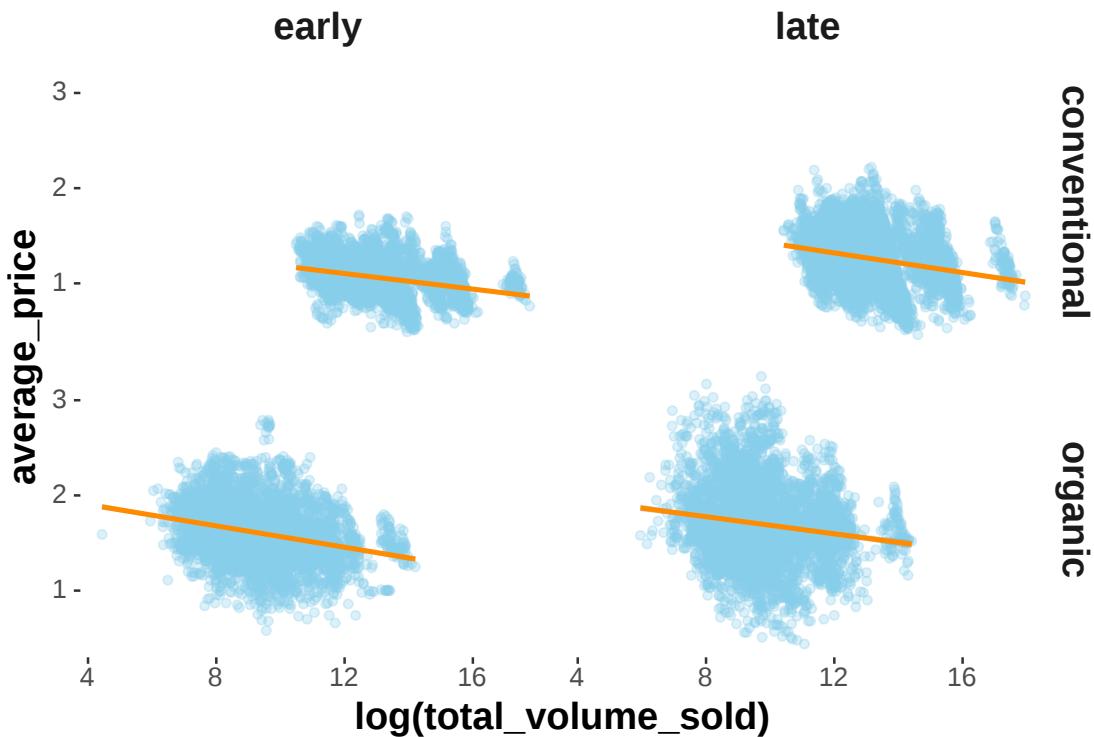
If we have grouping information, sometimes it can just get too much to put all of the information in a single plot, even if we use colors, shapes or linetypes for disambiguation. Facets are a great way to separately repeat the same kind of plot for different levels of relevant factors.

The functions `facet_grid` and `facet_wrap` are used for faceting. They both expect a formula-like syntax (we have not yet introduced formulas) using the notation `~` to separate factors. The difference between these functions shows most clearly when we have more than two factors. So let's introduce a new factor `early` to the avocado price data, representing whether a recorded measurement was no later than the median date or not.

```
avocado_data_early_late <- avocado_data %>%  
  mutate(early = ifelse(Date <= median(Date), "early", "late"))
```

Using `facet_grid` we get a two-dimensional grid, and we can specify along which axis of this grid the different factor levels are to range by putting the factors in the formula notation like this: `row_factor ~ col_factor`.

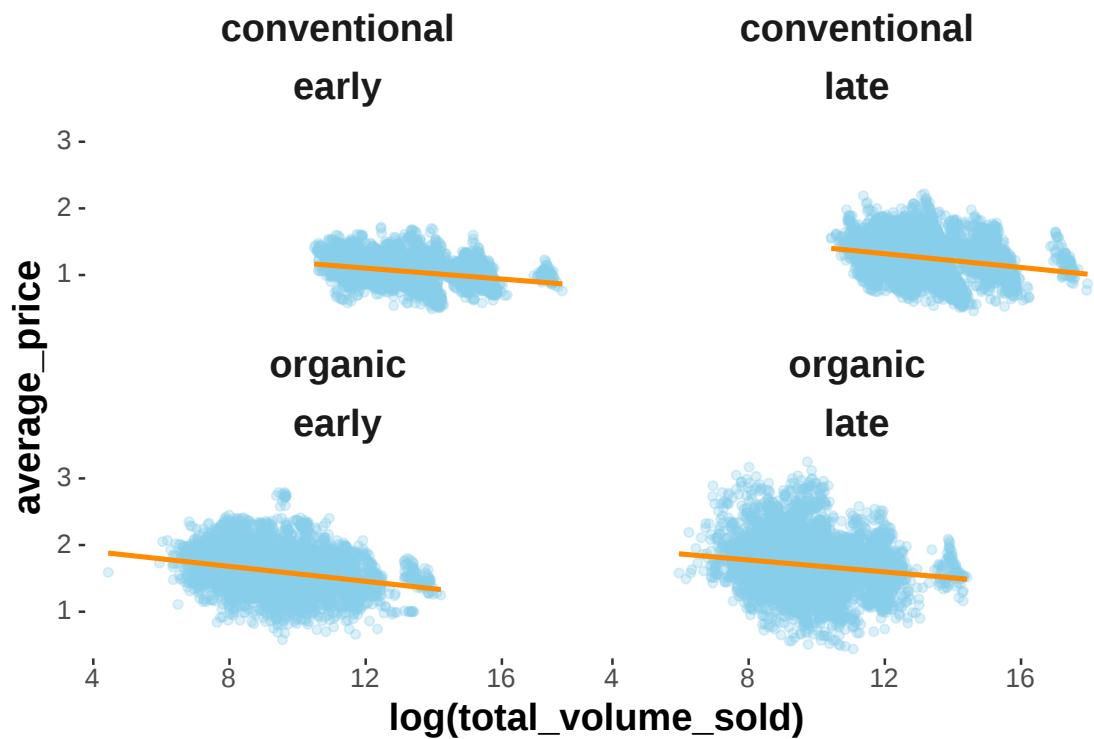
```
avocado_data_early_late %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(alpha = 0.3, color = "skyblue") +
  geom_smooth(method = "lm", color = "darkorange") +
  facet_grid(type ~ early)
```



The same kind of plot realized with `facet_wrap` looks slightly different. The different factor level combinations are mashed together into a pair.

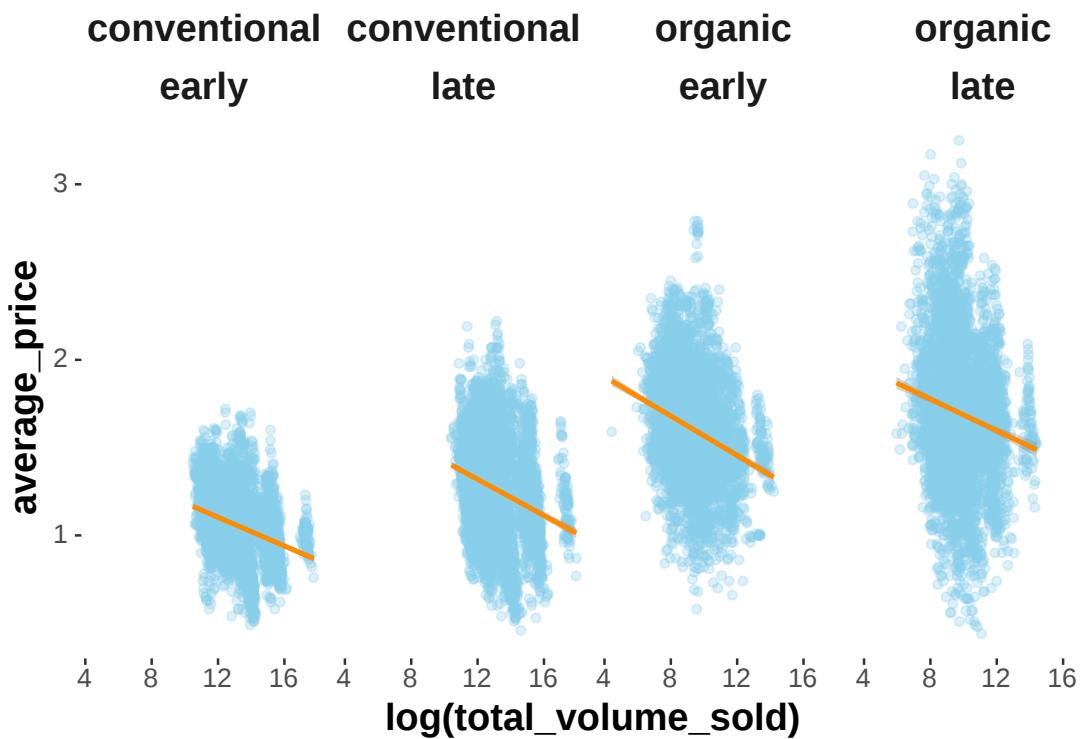
```
avocado_data_early_late %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(alpha = 0.3, color = "skyblue") +
  geom_smooth(method = "lm", color = "darkorange") +
  facet_wrap(type ~ early)
```

6. Data Visualization



With `facet_wrap` it is possible to specify the desired number of columns or rows:

```
avocado_data_early_late %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(alpha = 0.3, color = "skyblue") +
  geom_smooth(method = "lm", color = "darkorange") +
  facet_wrap(type ~ early, nrow = 1)
```



6.6. Customization etc.

There are many ways in which graphs can (and often: ought to) be tweaked further. The following can only cover a small, but hopefully useful selection.

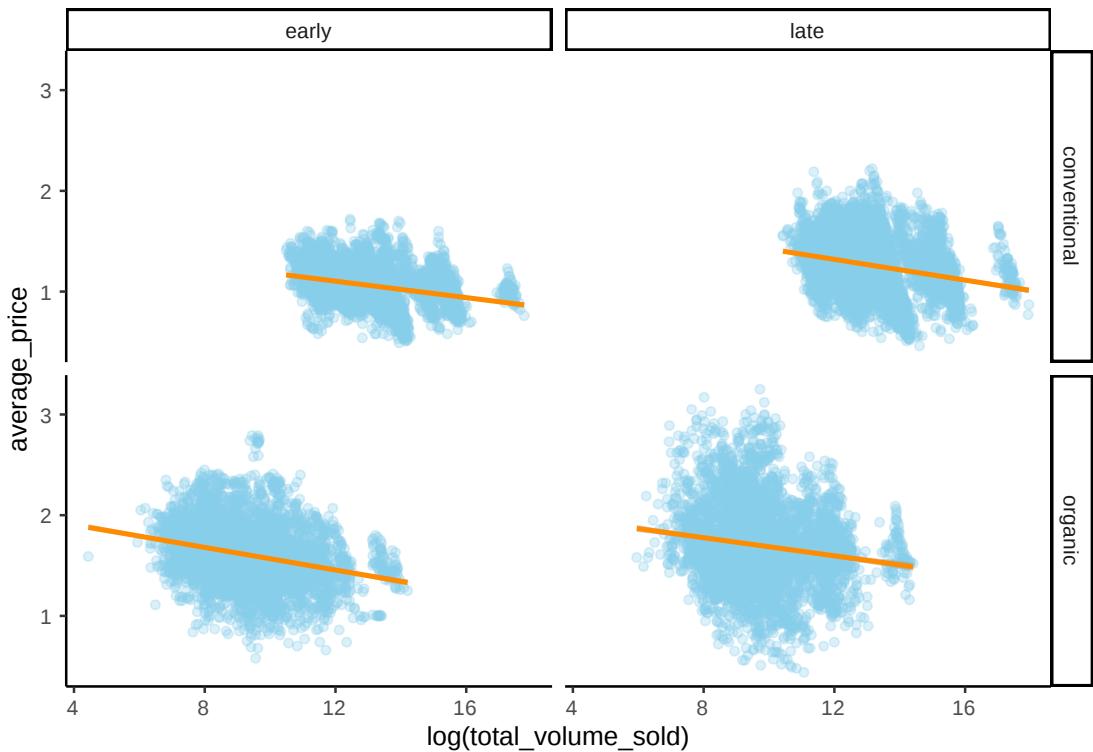
6.6.1. Themes

The general appearance of a plot is governed by its **theme**. There are many ready-made themes already in the `ggplot` package, as listed here, and there are more in several other packages. If we store a plot in a variable we can look at how different themes affect it.

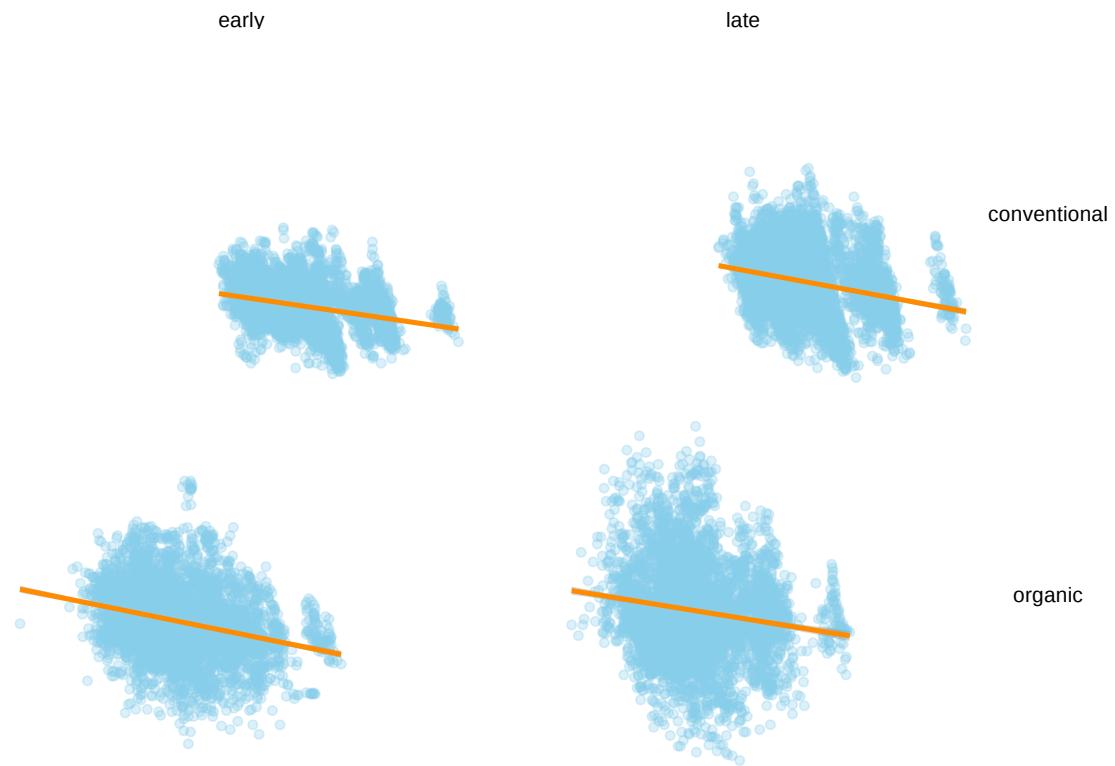
```
avocado_grid_plot <- avocado_data_early_late %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(alpha = 0.3, color = "skyblue") +
  geom_smooth(method = "lm", color = "darkorange") +
  facet_grid(type ~ early)

avocado_grid_plot + theme_classic()
```

6. Data Visualization

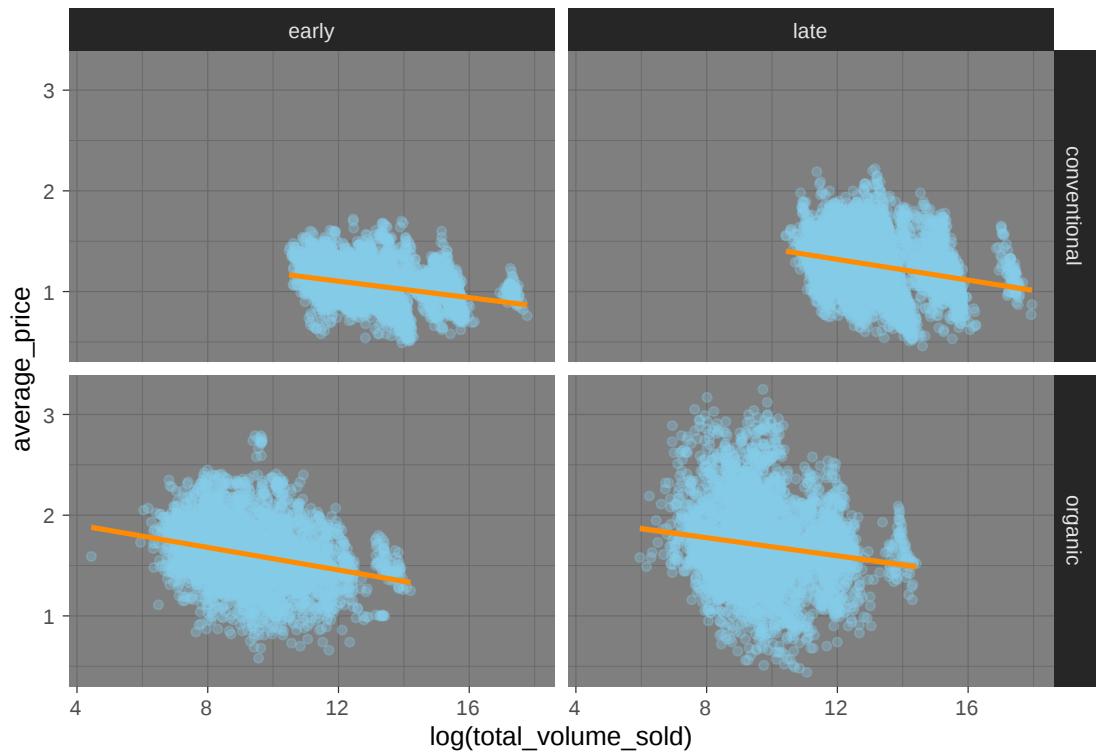


```
avocado_grid_plot + theme_void()
```



```
avocado_grid_plot + theme_dark()
```

6. Data Visualization

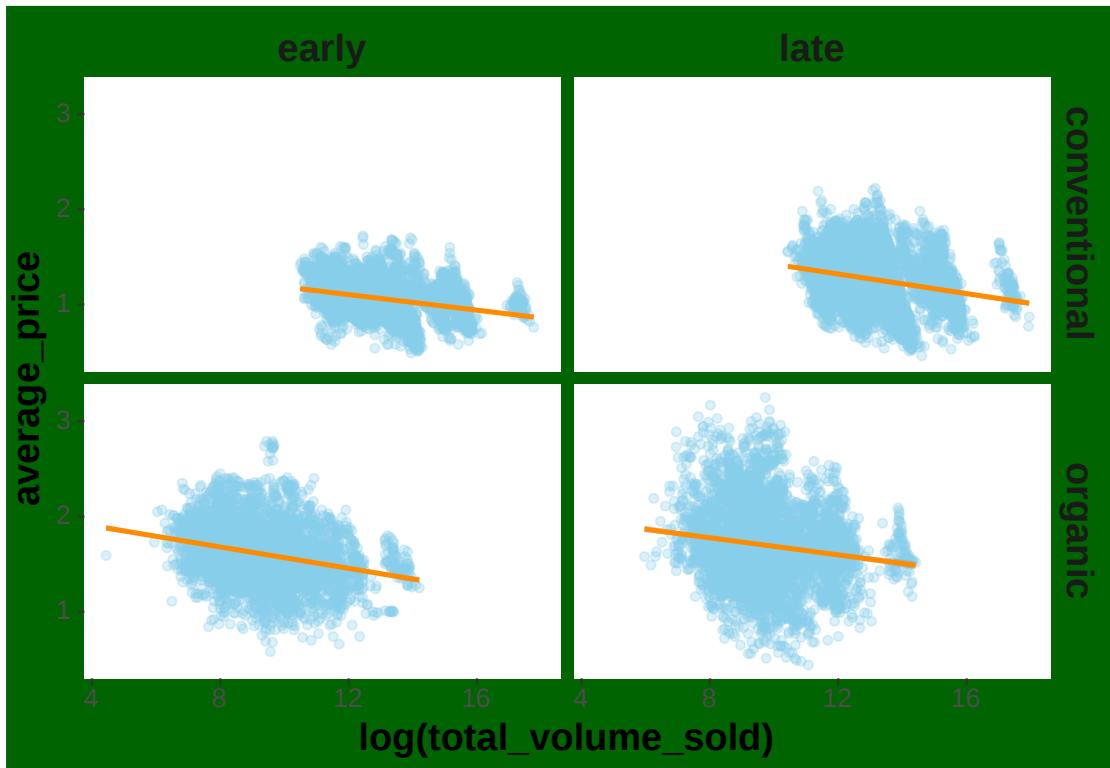


The plots in this book use the theme `hrbrthemes::theme_ipsum` from the `hrbrthemes` package as a default. You can set the default theme for all subsequent plots using a command like this:

```
# set the 'void' theme as global default
theme_set(
  theme_void()
)
```

More elaborate tweaking of a plot's layout can be achieved by the `theme` function. There are many options. Some let you do crazy things:

```
avocado_grid_plot + theme(plot.background = element_rect(fill = "darkgreen"))
```

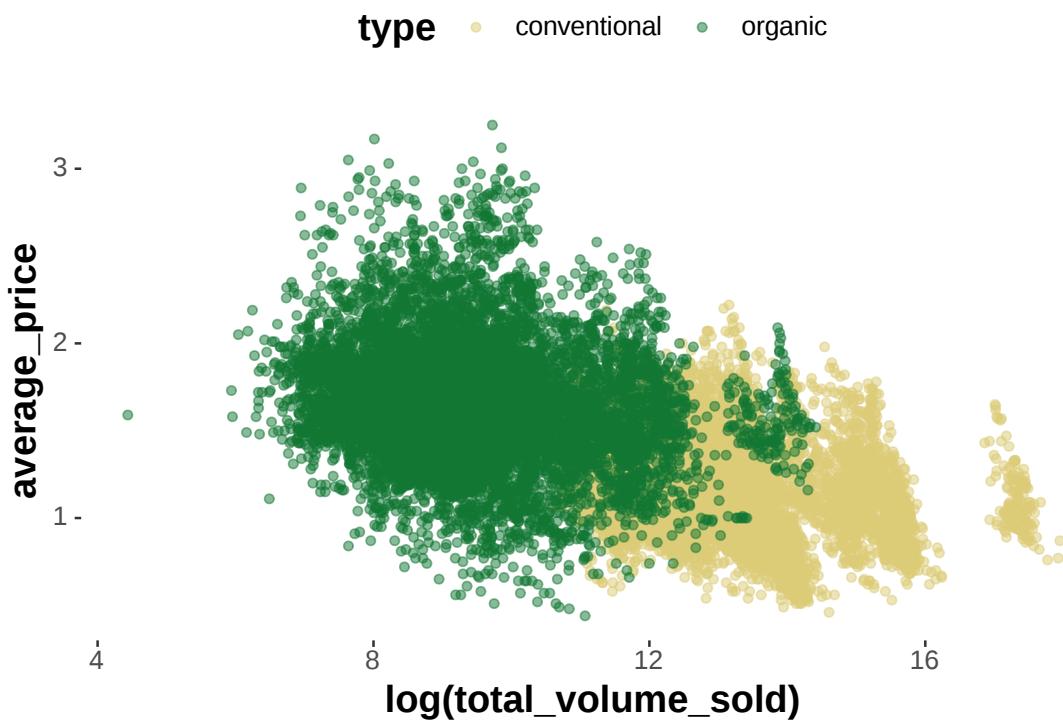


6.6.2. Guides

When using grouped variables (by color, shape, linetype, group, ...) ggplot creates a legend automatically.

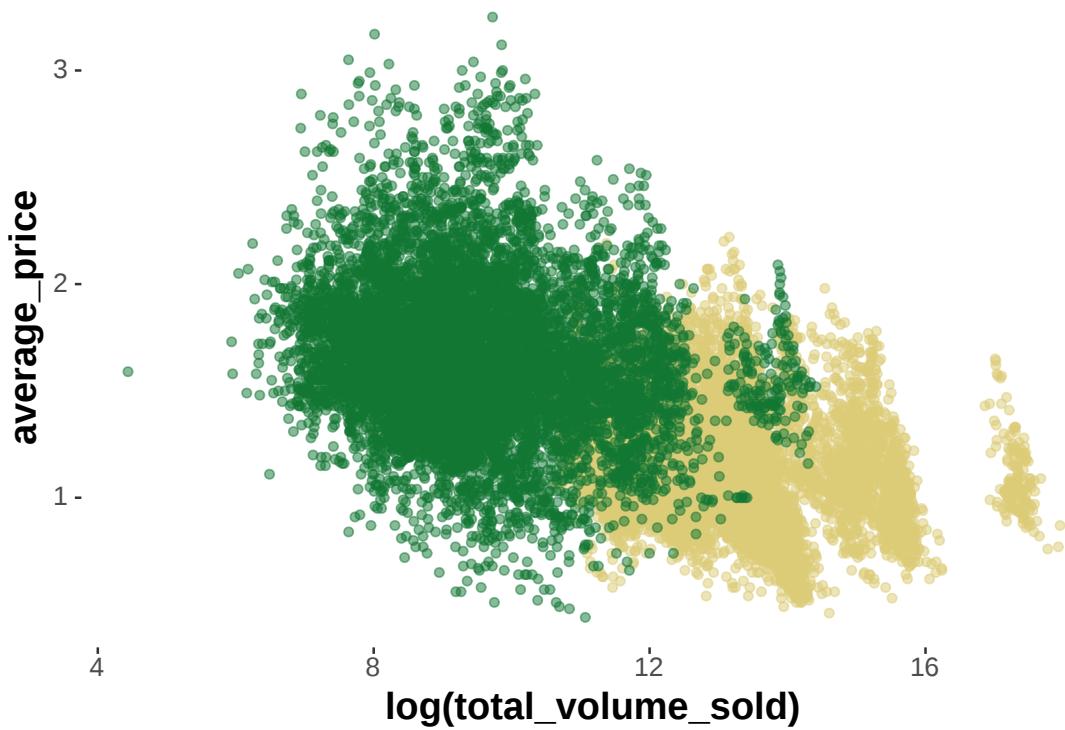
```
avocado_data %>%
  ggplot(
    mapping = aes(
      x = log(total_volume_sold),
      y = average_price,
      color = type
    )
  ) +
  geom_point(alpha = 0.5)
```

6. Data Visualization



The legend can be suppressed with the `guides` command. It takes as arguments the different types of grouping variables (like `color`, `group`, etc.)

```
avocado_data %>%
  ggplot(
    mapping = aes(
      x = log(total_volume_sold),
      y = average_price,
      color = type
    )
  ) +
  geom_point(alpha = 0.5) +
  # no legend for grouping by color
  guides(color = "none")
```

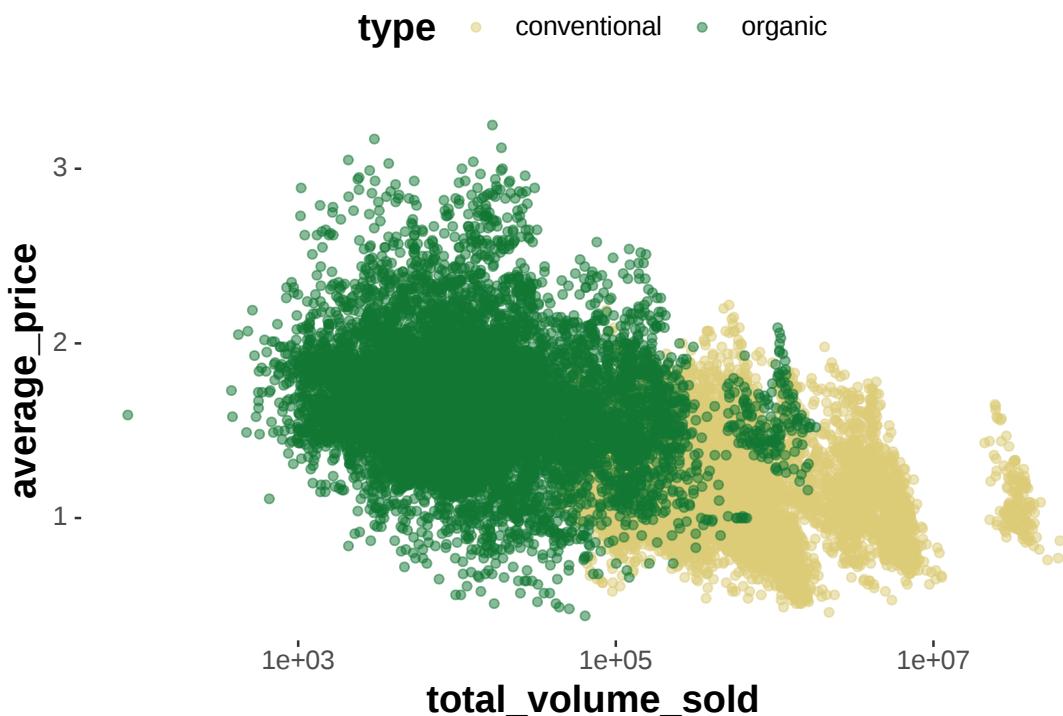


6.6.3. Axes, ticks and tick labels

If you need to use a non-standard (Cartesian) axis, you can do so, e.g., to change the x -axis to a log scale (with base 10):

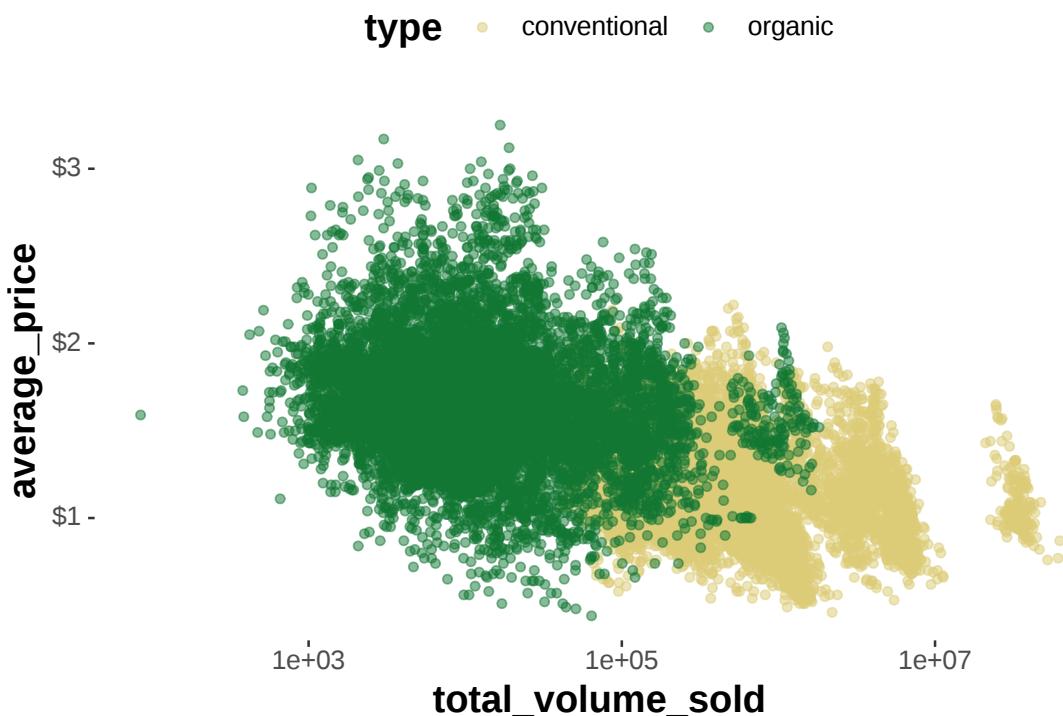
```
avocado_data %>%
  ggplot(
    mapping = aes(
      x = total_volume_sold,
      y = average_price,
      color = type
    )
  ) +
  geom_point(alpha = 0.5) +
  scale_x_log10()
```

6. Data Visualization



The `scales` package has a number of nice convenience functions for tweaking axis ticks (the places where axes are marked and possibly labelled) and tick labels (the labels applied to the tick marks). For example, we can add dollar signs to the price information, like so:

```
avocado_data %>%
  ggplot(
    mapping = aes(
      x = total_volume_sold,
      y = average_price,
      color = type
    )
  ) +
  geom_point(alpha = 0.5) +
  scale_x_log10() +
  scale_y_continuous(labels = scales::dollar)
```



6.6.4. Labels

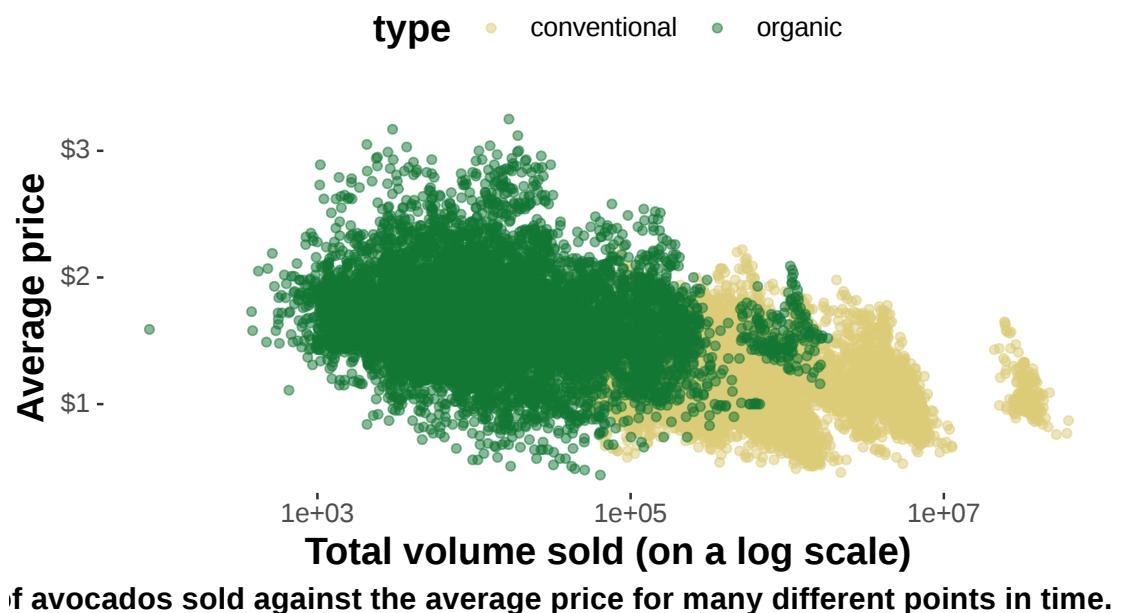
To change any other kind of labeling information (aside from tick mark labels on axes), the `labs` function can be used. It is rather self-explanatory:

```
avocado_data %>%
  ggplot(
    mapping = aes(
      x = total_volume_sold,
      y = average_price,
      color = type
    )
  ) +
  geom_point(alpha = 0.5) +
  scale_x_log10() +
  scale_y_continuous(labels = scales::dollar) +
  # change axis labels and plot title & subtitle
  labs(
    x = 'Total volume sold (on a log scale)',
    y = 'Average price',
    title = "Avocado prices plotted against the amount sold per type",
  )
```

6. Data Visualization

```
    subtitle = "With linear regression lines",
    caption = "This plot shows the total volume of avocados sold against the average price
)
```

Avocado prices plotted against the amount sold With linear regression lines



If avocados sold against the average price for many different points in time.

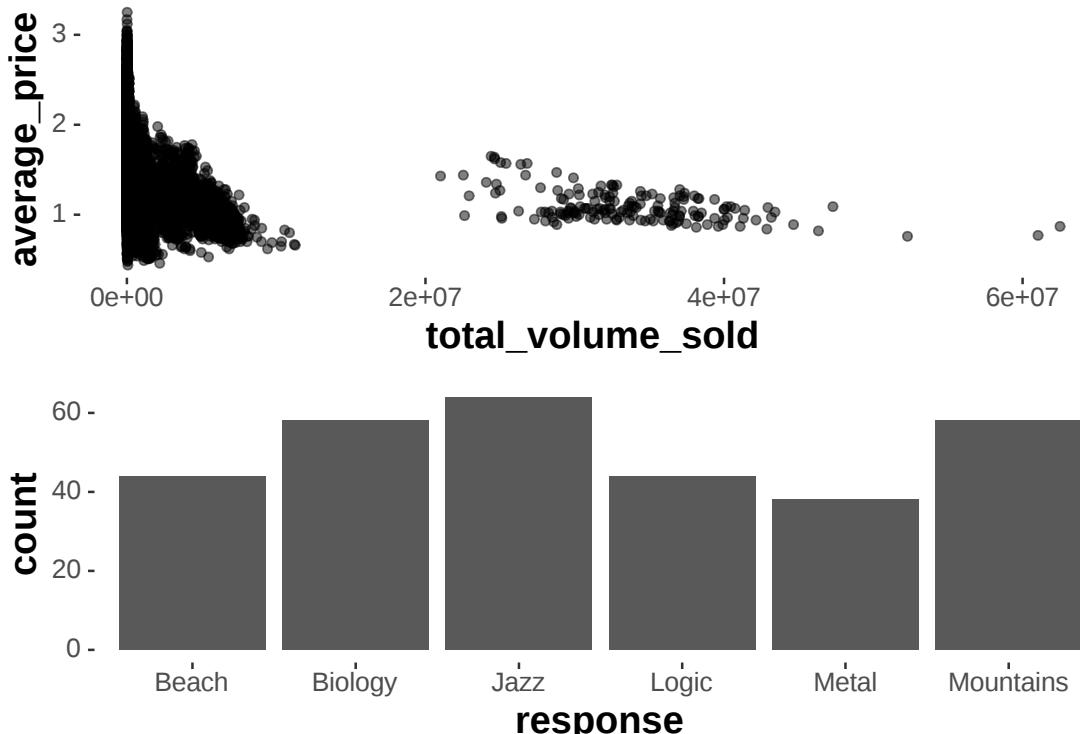
6.6.5. Combining & arranging plots

Presenting visual information in a tightly packed spatial arrangement can be helpful for the spectator. Everything is within a single easy saccade, so to speak. Therefore it can be useful to combine different plots into a single combined plot. The cowplot package helps with this, in particular the function `cowplot::plot_grid` as shown here:

```
# create an avocado plot
avocado_plot <- avocado_data %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point(alpha = 0.5)

# create a BLJM bar plot
BLJM_plot <- data_BLJM_processed %>%
  ggplot(aes(x = response)) +
  geom_bar()
```

```
# combine both into one
cowplot::plot_grid(
  # plots to combine
  avocado_plot,
  BLJM_plot,
  # number columns
  ncol = 1
)
```



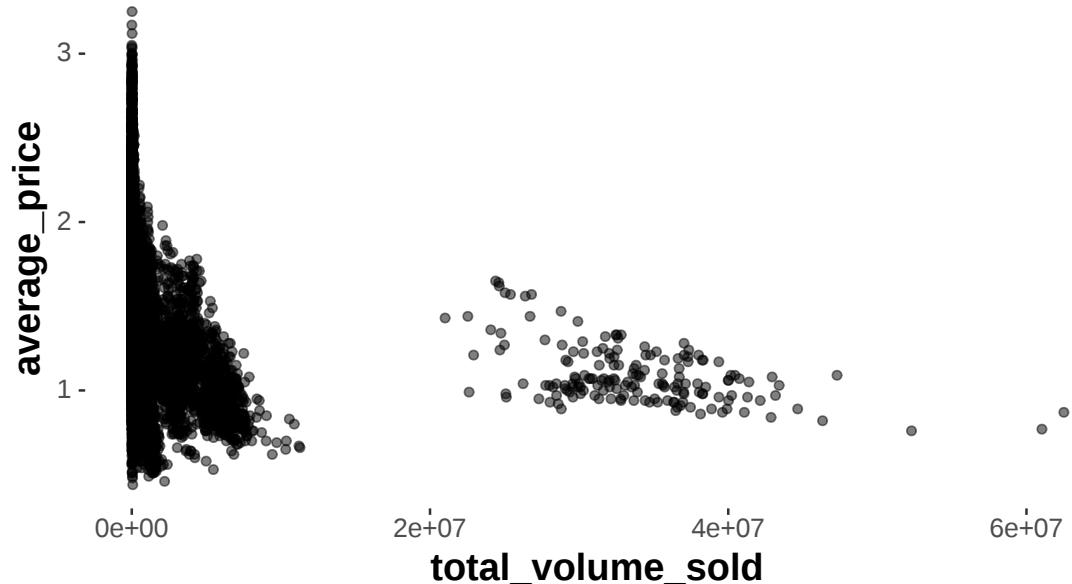
6.6.6. LaTeX expressions in plot labels

If you are enthusiastic about LaTeX, you can also use it inside of plot labels. The `latex2exp` package is useful here, which provides the function `latex2exp::TeX` to allow you to include LaTeX formulas. Just make sure that you double all backslashes, as in the following example:

```
avocado_data %>%
  ggplot(aes(x = total_volume_sold, y = average_price)) +
  geom_point(alpha = 0.5) +
  labs(title = latex2exp::TeX("We can use $\\LaTeX$ here: $\\sum_{i = 0}^n \\alpha^i$"))
```

6. Data Visualization

We can use L^AT_EX here: $\sum_{i=0}^n \alpha^i$



Part III.

Models and inferences

7. Basics of Probability Theory

Probability is the basic ingredient of statistical inference.

In this chapter we will cover the very basics of probability theory. We will visit its axiomatic definition and some common interpretations in Section 7.1, where we also start with the main mental exercise of this section: seeing how **probability distributions can be approximately represented by samples**. We will cover important concepts such as joint and marginal probability in Section 7.2. This paves the way for learning about conditional probability and Bayes rule in Section 7.3. Section 7.4 introduces the notion of a random variable. Finally Section 7.5 briefly covers how information about common probability distributions can be accessed in R.

Learning goals:

- become familiar with the notion of probability and also:
 - its axiomatic definition
 - the notion of joint, marginal and conditional probability
- understand and apply Bayes rule
- get comfortable with the notion of and notation for random variables
- become able to handle probability distributions in R
- understand how probability distributions are approximately represented by samples

7.1. Probability

7.1.1. Outcomes, events, observations

We are interested in the space Ω of all **elementary outcome** $\omega_1, \omega_2, \dots$ of a process or event whose execution is (partially) random or unknown. Elementary outcomes are mutually exclusive. The set Ω exhausts all possibilities.¹

¹For simplicity of exposure, we gloss over subtleties arising when dealing with infinite sets Ω . We make up for this when we define probability density functions for continuous random variables, which is all the uncountable infinity that we will usually be concerned with in applied statistics.

7. Basics of Probability Theory

Example. The set of elementary outcomes of a single coin flip is $\Omega_{\text{coin flip}} = \{\text{heads, tails}\}$. The elementary outcomes of tossing a six-sided die is $\Omega_{\text{standard die}} = \{\square, \square, \square, \square, \square, \square\}$.²

An **event** A is a subset of Ω . Think of an event as a (possibly partial) observation. We might observe, for instance, not the full outcome of tossing a die, but only that there is a dot in the middle. This would correspond to the event $A = \{\square, \square, \square\}$, i.e., observing an odd numbered outcome. The *trivial observation* $A = \Omega$ and the *impossible observation* $A = \emptyset$ are counted as events, too. The latter is included for technical reasons.

For any two events $A, B \subseteq \Omega$, standard set operations correspond to logical connectives in the usual way. For example, the conjunction $A \cap B$ is the observation of both A and B ; the disjunction $A \cup B$ is the observation that it is either A or B ; the negation of A , $\overline{A} = \{\omega \in \Omega \mid \omega \notin A\}$, is the observation that it is not A .

7.1.2. Probability distributions

A **probability distribution** P over Ω is a function $P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ that assigns to all events $A \subseteq \Omega$ a real number (from the unit interval, see A1 below), such that the following (so-called Kolmogorov axioms) are satisfied:

A1. $0 \leq P(A) \leq 1$

A2. $P(\Omega) = 1$

A3. $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$ whenever A_1, A_2, A_3, \dots are mutually exclusive³

Occasionally we encounter notation $P \in \Delta(\Omega)$ to express that P is a probability distribution over Ω . (E.g., in physics, theoretical economics or game theory. Less so in psychology or statistics.) If $\omega \in \Omega$ is an elementary event, we often write $P(\omega)$ as a shorthand for $P(\{\omega\})$. In fact, if Ω is finite, it suffices to assign probabilities to elementary outcomes.

A number of rules follow immediately from this definition (prove this!):

C1. $P(\emptyset) = 0$

C2. $P(\overline{A}) = 1 - P(A)$

C3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any $A, B \subseteq \Omega$

7.1.3. Interpretations of probability

It is reasonably safe, at least preliminarily, to think of probability, as defined above, as a handy mathematical primitive which is useful for certain applications. There are at least three ways of thinking about where this

²Think of Ω as a partition of the space of all possible ways in which the world could be, where we lump together into one partition cell all ways in which the world could be that are equivalent regarding those aspects of reality that we are interested in. We do not care whether the coin lands in the mud or in the sand. It only matters whether it came up heads or tails. Each elementary event can be realized in myriad ways. Ω is our, the modellers', first crude simplification of nature, abstracting away aspects we currently do not care about.

³A3 is the axiom of *countable additivity*. Finite additivity may be enough for finite or countable sets Ω , but infinite additivity is necessary for full generality in the uncountable case.

primitive probability might come from, roughly paraphrasable like so:

1. **Frequentist:** Probabilities are generalizations of intuitions/facts about frequencies of events in repeated executions of a random event.
2. **Subjectivist:** Probabilities are subjective beliefs by a rational agent who is uncertain about the outcome of a random event.
3. **Realist:** Probabilities are a property of an intrinsically random world.

7.1.4. Urns, frequencies & distributions as samples

No matter what your metaphysics of probability are, it is useful to realize that probability distributions can be approximately represented by sampling.

Think of an **urn** as a container with differently colors balls of different proportions (see Figure 7.1). In the simplest case, there is a number of $N > 1$ balls of which $k > 0$ are black and $N - k > 0$ are white. (There is at least one black and one white ball.) For a single random draw from our urn we have: $\Omega_{\text{our urn}} = \{\text{white, black}\}$. We now draw from this urn with replacement. That is, we shake the urn, draw one ball, observe its color, take note of the color and put it back into the urn. Each ball has the same chance of being sampled. If we imagine an infinite sequence of single draws from our urn, putting whichever ball we drew back in after every draw, the limiting proportion with which we draw a black ball is $\frac{k}{N}$. This statement about frequency is what motivates saying that the probability of drawing a black ball on a single trial is (or should be⁴) $P(\text{black}) = \frac{k}{N}$.

The following code demonstrates how the proportion of black balls drawn from an urn like in Figure 7.1 with $k = 7$ black balls and $N = 10$ balls in total, gravitates to the probability 0.7 when we keep drawing and drawing.

```
# urn with 7 black and 3 white balls
urn <- c(
  rep("black", 7),
  rep("white", 3)
)

# number of samples to take
n_samples <- 10000

# take `n_samples` samples from the urn (with replacement)
draws <- sample(
  # vector to sample from (default probability is uniform)
  x = urn,
  # take a million samples
  size = n_samples,
  # put each ball back after drawing
```

⁴If probabilities are subjective beliefs, a rational agent is, in a sense, normatively required to assign exactly this probability.

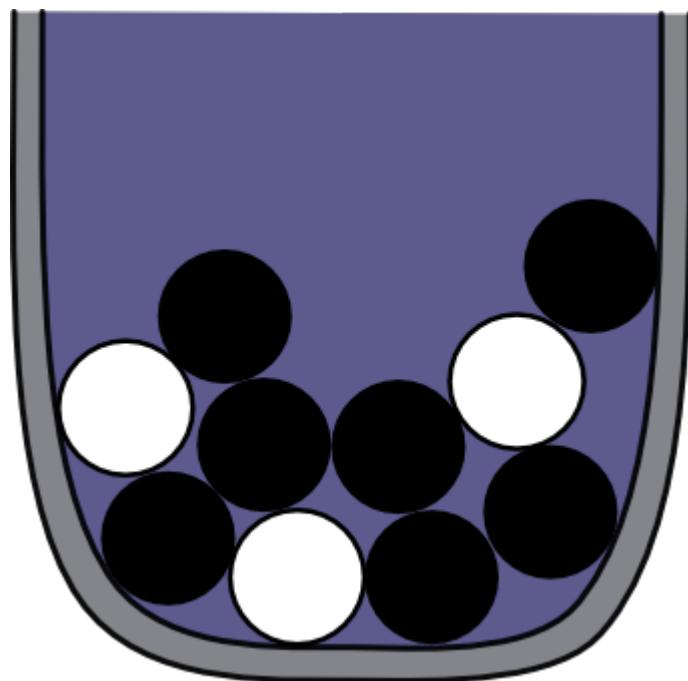


Figure 7.1.: An urn with seven black balls and three white balls. Imagine shaking this container, and then drawing blindly a single ball from it. If every ball has equal probability of being drawn, what is the probability of drawing a black ball? That would be 0.7.

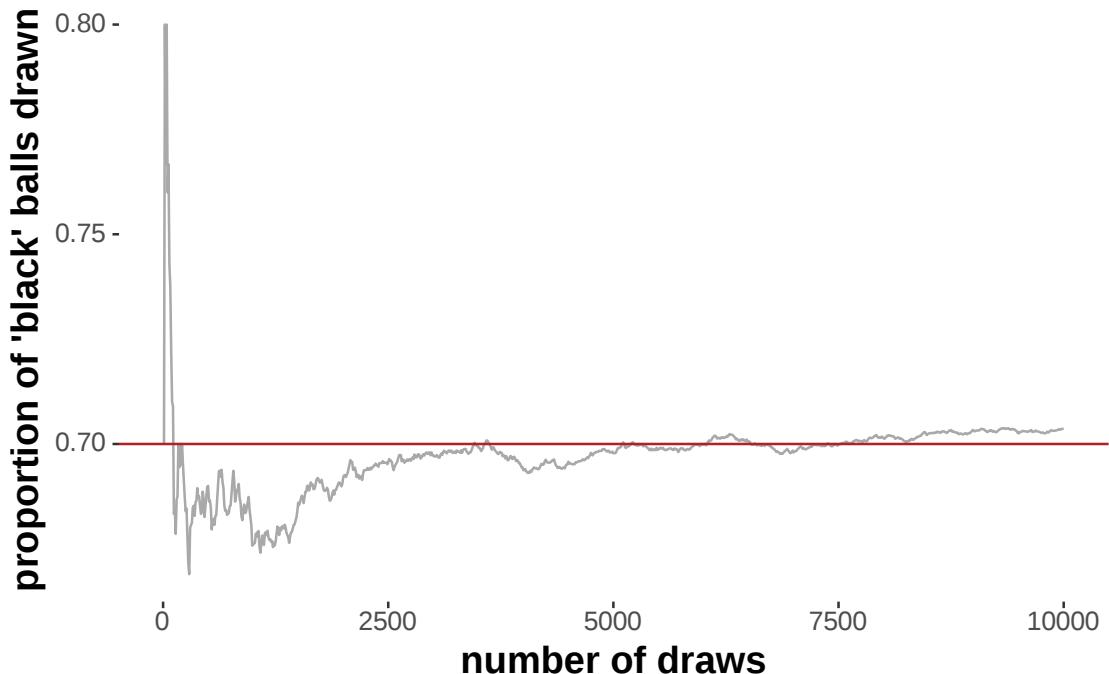
```

replace = TRUE
)

# plotting the development of proportion 'black'
tibble(
  draw_nr = 1:n_samples,
  draw = draws,
  prop_black = cumsum(draw == "black") / draw_nr
) %>%
  filter(draw_nr %% 10 == 0) %>%
  ggplot(aes(x = draw_nr, y = prop_black)) +
  geom_line(color = "darkgray") +
  # add a red line for the true limiting probability
  geom_hline(aes(yintercept = 0.7), color = "firebrick") +
  labs(
    x = "number of draws",
    y = "proportion of 'black' balls drawn",
    title = "Temporal development of the proportion of draws from an urn"
)

```

Temporal development of the proportion of drawn black balls



To sum this up concisely, we have a random process (drawing once from the urn) whose outcome is uncer-

Table 7.1.: Joint probability table for the flip-and-draw scenario

	heads	tails
black	$0.5 \times 0.2 = 0.1$	$0.5 \times 0.4 = 0.2$
white	$0.5 \times 0.8 = 0.4$	$0.5 \times 0.6 = 0.3$

tain, and we convinced ourselves that the probability of an outcome corresponds to the relative frequency it occurs, in the limit of repeatedly executing the random process (i.e., sampling from the urn). From here it is only a small step to a crucial but ultimately very liberating realization. If probability of an event occur can be approximated by its frequency in a large sample, then we can represent (say: internally in a computer) a probability distribution as one of two things:

1. a large set of (what is called: representative) samples; or even better as
2. an oracle (e.g., in the form of a clever algorithm) which quickly returns a representative sample.

This means that, for approximately computing with probability, we can represent distributions through samples or a sample-generating function. We do not need to know precise probability, or be able to express them in a mathematical formula. Samples or sampling is enough to approximate probability distributions.

7.2. Structured events & marginal distributions

7.2.1. Probability table for a flip-and-draw scenario

Suppose we have two urns. Both have $N = 10$ balls. Urn 1 has $k_1 = 2$ black and $N - k_1 = 8$ white balls. Urn 2 has $k_2 = 4$ black and $N - k_2 = 6$ white balls. We sometimes draw from urn 1, sometimes from urn 2. To decide, we flip a fair coin. If it comes up heads, we draw from urn 1; if it comes up tails, we draw from urn 2. The process is visualized in Figure 7.2 below.

An elementary outcome of this two-step process of flip-and-draw is a pair \langle outcome-flip, outcome-draw \rangle . The set of all possible such outcomes is:

$$\Omega_{\text{flip-and-draw}} = \{\langle \text{heads}, \text{black} \rangle, \langle \text{heads}, \text{white} \rangle, \langle \text{tails}, \text{black} \rangle, \langle \text{tails}, \text{white} \rangle\} .$$

The probability of event $\langle \text{heads}, \text{black} \rangle$ is given by multiplying the probability of seeing “heads” on the first flip, which happens with probability 0.5, and then drawing a black ball, which happens with probability 0.2, so that $P(\langle \text{heads}, \text{black} \rangle) = 0.5 \times 0.2 = 0.1$. The probability distribution over $\Omega_{\text{flip-draw}}$ is consequently as in Table 7.1. (If in doubt, start flipping & drawing and count your outcomes.)

7.2.2. Structured events and joint-probability distributions

Table 7.1 is an example of a **joint probability distribution** over a structured event space, which here has two dimensions. Since our space of outcomes is the Cartesian product of two simpler outcome spaces, namely

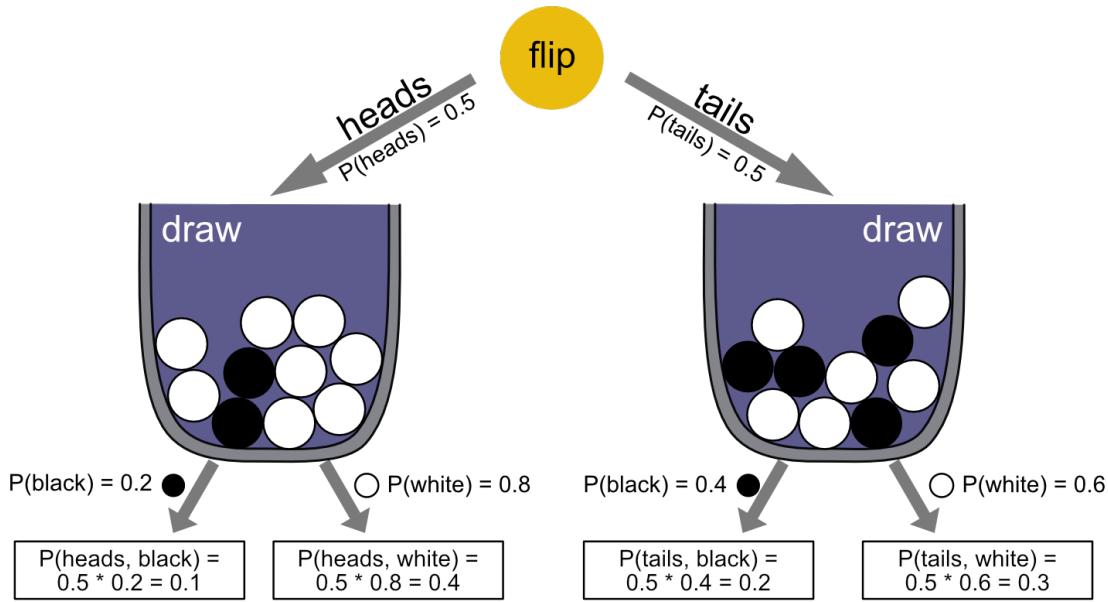


Figure 7.2.: The flip-and-draw scenario, with transition and full path probabilities.

$\Omega_{\text{flip-and-draw}} = \Omega_{\text{flip}} \times \Omega_{\text{draw}}$,⁵ we can use notation $P(\text{heads, black})$ as shorthand for $P(\langle \text{heads, black} \rangle)$. More generally, if $\Omega = \Omega_1 \times \dots \Omega_n$, we can think of $P \in \Delta(\Omega)$ as a joint probability distribution over n subspaces.

7.2.3. Marginalization

If P is a joint-probability distribution over event space $\Omega = \Omega_1 \times \dots \Omega_n$, the **marginal distribution** over subspace Ω_i , $1 \leq i \leq n$ is the probability distribution that assigns to all $A_i \subseteq \Omega_i$ the probability:⁶

$$P(A_i) = \sum_{A_1 \subseteq \Omega_1, \dots, A_{i-1} \subseteq \Omega_{i-1}, A_{i+1} \subseteq \Omega_{i+1}, \dots, A_n \subseteq \Omega_n} P(A_1, \dots, A_{i-1}, A_i, A_{i+1}, \dots, A_n)$$

For example, the marginal distribution over coin flips derivable from the joint probability distribution in Table 7.1 gives $P(\text{heads}) = P(\text{tails}) = 0.5$, since the sum of each column is exactly 0.5. The marginal distribution over flips derivable from Table 7.1 has $P(\text{black}) = 0.3$ and $P(\text{white}) = 0.7$.⁷

⁵With $\Omega_{\text{flip}} = \{\text{heads, tails}\}$ and $\Omega_{\text{draw}} = \{\text{black, white}\}$.

⁶This notation, using \sum , assumes that subspaces are countable. In other cases, a parallel definition with integrals can be used.

⁷The term “marginal distribution” derives from such probability tables, where traditionally the sum of each row/column was written in the margins.

7.3. Conditional probability

Fix probability distribution $P \in \Delta(\Omega)$ and events $A, B \subseteq \Omega$. The conditional probability of A given B , written as $P(A | B)$, gives the probability of A on the assumption that B is true.⁸ It is defined like so:

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

Conditional probabilities are only defined when $P(B) > 0$.⁹

Example. If a die is unbiased, each of its six faces has equal probability to come up after a toss. The probability of event $B = \{\text{odd}\}$ that the tossed number is odd has probability $P(B) = \frac{1}{2}$. The probability of event $A = \{\text{even}\}$ that the tossed number is even has probability $P(A) = \frac{2}{3}$. The probability that the tossed number is even and odd is $P(A \cap B) = P(\{\text{odd}\}) = \frac{1}{3}$. The conditional probability of tossing a number that is even, when we know that the toss is odd, is $P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}$.

Algorithmically, conditional probability first rules out all events in which B is not true and then simply renormalizes the probabilities assigned to the remaining events in such a way that the relative probabilities of surviving events remains unchanged. Given this, another way of interpreting conditional probability is that $P(A | B)$ is what a rational agent should *should* believe about A after observing (nothing more than) that B is true. The agent rules out, possibly hypothetically, that B is false, but otherwise does not change opinion about the relative probabilities of anything that is compatible with B .

7.3.1. Bayes rule

Looking back at the joint-probability distribution in Table 7.1, the conditional probability $P(\text{black} | \text{heads})$ of drawing a black ball, given that the initial coin flip showed heads, can be calculated as follows:

$$P(\text{black} | \text{heads}) = \frac{P(\text{black, heads})}{P(\text{heads})} = \frac{0.1}{0.5} = 0.2$$

This calculation, however, is quite spurious. We knew that already from the way the flip-and-draw scenario was set up. After flipping heads, we draw from urn 1, which has $k = 2$ out of $N = 10$ black balls, so clearly: if the flip is heads, then the probability of a black ball is 0.2. Indeed, in a step-wise random generation process like the flip-and-draw scenario, some conditional probabilities are very clear, and sometimes given by definition. These are, usually, the conditional probabilities that define how the process unfolds forward in time, so to speak.

⁸We also verbalize this as "the conditional probability of A conditioned on B ."

⁹Updating with events which have probability zero entails far more severe adjustments of the underlying belief system than just ruling out information hitherto considered possible. Formal systems that capture such *belief revision* are studied in formal epistemology Halpern (2003).

Bayes rule is a way of expressing, in a manner of speaking, conditional probabilities in terms of the “reversed” conditional probabilities:

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

Bayes rule is straightforward corollary of the definition of conditional probabilities, according to which $P(A \cap B) = P(A | B) \times P(B)$, so that:

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

Bayes rule allows for reasoning backwards from observed causes to likely underlying effects. When we have a feed-forward model of how unobservable effects probabilistically constrain observable outcomes, Bayes rule allows us to draw inferences about *latent/unobservable variables* based on the observation of their downstream effects.

Consider yet again the flip-and-draw scenario. But now assume that Jones flipped the coin and drew a ball. We see that it is black. What is the probability that it was drawn from urn 1, equivalently, that the coin landed heads? It is not $P(\text{heads}) = 0.5$, the so-called *prior probability* of the coin landing heads. It is a conditional probability, also called the *posterior probability*,¹⁰ namely $P(\text{heads} | \text{black})$, but one that is not as easy and straightforward to write down as the reverse $P(\text{black} | \text{heads})$ of which we said above that it is an almost trivial part of the set up of the flip-and-draw scenario. It is here that Bayes rule has its purpose:

$$P(\text{heads} | \text{black}) = \frac{P(\text{black} | \text{heads}) \times P(\text{heads})}{P(\text{black})} = \frac{0.2 \times 0.5}{0.3} = \frac{1}{3}$$

This result is quite intuitive. Drawing a black ball from urn 2 (i.e., after seeing tails) is twice as likely as drawing a black ball from urn 1 (i.e., after seeing heads). Consequently, after seeing a black ball drawn, with equal probabilities of heads and tails, the probability that the coin landed tails is also twice as large as that it landed heads.

Excursion: Bayes rule for data analysis In later chapters we will use Bayes rule for data analysis. The flip-and-draw scenario structurally “preflects” what will happen later. Think of the color of the ball drawn as the *data D* which we observe. Think of the coin as a *latent parameter θ* of a statistical model. Bayes rule for data analysis then looks like this:

$$P(\theta | D) = \frac{P(D | \theta) P(\theta)}{P(D)}$$

We will discuss this at length in Chapter 8 and thereafter.

¹⁰The terms *prior* and *posterior* make sense when we think about an agent’s belief state before (prior to) and after (posterior to) an observation.

7. Basics of Probability Theory

Table 7.2.: Joint probability table for a flip-and-draw scenario where the coin has a bias of 0.8 towards heads and where each of the two urns hold 3 black and 7 white balls.

	heads	tails	Σ rows
black	$0.8 \times 0.3 = 0.24$	$0.2 \times 0.3 = 0.06$	0.3
white	$0.8 \times 0.7 = 0.56$	$0.2 \times 0.7 = 0.14$	0.7
Σ columns	0.8	0.2	1.0

7.3.2. Stochastic (in-)dependence

Event A is **stochastically independent** of B if, intuitively speaking, learning B does not change one's beliefs about A : $P(A | B) = P(A)$. If A is stochastically independent of B , then B is stochastically independent of A because:

$$\begin{aligned} P(B | A) &= \frac{P(A | B) P(B)}{P(A)} && [\text{Bayes rule}] \\ &= \frac{P(A) P(B)}{P(A)} && [\text{by ass. of independence}] \\ &= P(B) && [\text{cancellation}] \end{aligned}$$

For example, imagine a flip-and-draw scenario where the initial coin flip has a bias of 0.8 towards heads, but each of the two urns has the same number of black balls, namely 3 black and 7 white balls. Intuitively and formally, the probability of drawing a black ball is then *independent* of the outcome of the coin flip; learning that the coin landed heads, does not change our beliefs about how likely the subsequent draw will result in a black ball. The probability table for this example is in Table 7.2.

Independence shows in Table 7.2 in the fact that the probability in each cell is the product of the two marginal probabilities. This is a direct consequence of stochastic independence:

Proposition 7.1 (Probability of conjunction of stochastically independent events). *For any pair of events A and B with non-zero probability:*

$$P(A \cap B) = P(A) P(B) \quad [\text{if } A \text{ and } B \text{ are stoch. independent}]$$

Proof. By assumption of independence, it holds that $P(A | B) = P(A)$. But then:

$$\begin{aligned} P(A \cap B) &= P(A | B) P(B) && [\text{def. of conditional probability}] \\ &= P(A) P(B) && [\text{by ass. of independence}] \end{aligned}$$

□

7.4. Random variables

We have so far defined a probability distribution as a function that assigns a probability to each subset of the space Ω of elementary outcomes. A special case occurs when we are interested in a space of numeric outcomes.

A **random variable** is a function $X : \Omega \rightarrow \mathbb{R}$ that assigns to each elementary outcome a numerical value. It is reasonable to think of this number as a **summary statistic**: a number that captures one aspect of relevance of what is actually a much more complex chunk of reality.

Example. For a single flip of a coin we have $\Omega_{\text{coin flip}} = \{\text{heads, tails}\}$. A usual way of mapping this onto numerical outcomes is to define $X_{\text{coin flip}} : \text{heads} \mapsto 1; \text{tails} \mapsto 0$. Less trivially, consider flipping a coin two times. Elementary outcomes should be individuated by the outcome of the first flip and the outcome of the second flip, so that we get:

$$\Omega_{\text{two flips}} = \{\langle \text{heads, heads} \rangle, \langle \text{heads, tails} \rangle, \langle \text{tails, heads} \rangle, \langle \text{tails, tails} \rangle\}$$

Consider the random variable $X_{\text{two flips}}$ that counts the total number of heads. Crucially, $X_{\text{two flips}}(\langle \text{heads, tails} \rangle) = 1 = X_{\text{two flips}}(\langle \text{tails, heads} \rangle)$. We assign the same numerical value to different elementary outcomes.

7.4.1. Notation & terminology

Traditionally random variables are represented by capital letters, like X . Variables for the numeric values they take on are written as small letters, like x .

We write $P(X = x)$ as a shorthand for the probability $P(\{\omega \in \Omega \mid X(\omega) = x\})$ that an event occurs that is mapped onto x by random variable X . For example, if our coin is fair, then $P(X_{\text{two flips}} = x) = 0.5$ for $x = 1$ and 0.25 otherwise. Similarly, we can also write $P(X \leq x)$ for the probability of observing an event that X maps to a number not bigger than x .

If the range of X is countable, we say that X is **discrete**. For ease of exposition, we may say that if the range of X is an interval of real numbers, X is called **continuous**.

7.4.2. Cumulative distribution functions, mass & density

For a discrete random variable X , the **cumulative distribution function** F_X associated with X is defined as:

$$F_X(x) = P(X \leq x) = \sum_{x' \in \{\text{Rng}(X) \mid x' \leq x\}} P(X = x')$$

The **probability mass function** f_X associated with X is defined as:

$$f_X(x) = P(X = x)$$

7. Basics of Probability Theory

Example. Suppose we flip a coin with a bias of θ n times. What is the probability that we will see heads k times? If we map the outcome of heads to 1 and tails to 0, this probability is given by the Binomial distribution, as follows:

$$\text{Binom}(K = k; n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. It gives the number of possibilities of drawing an unordered set with k elements from a set with a total of n elements. Figure 7.3 gives an example of the Binomial distribution, concretely its probability mass function, for two values of the coin's bias, $\theta = 0.25$ or $\theta = 0.5$, when flipping the coin $n = 24$ times. Figure 7.4 gives the corresponding cumulative distributions.

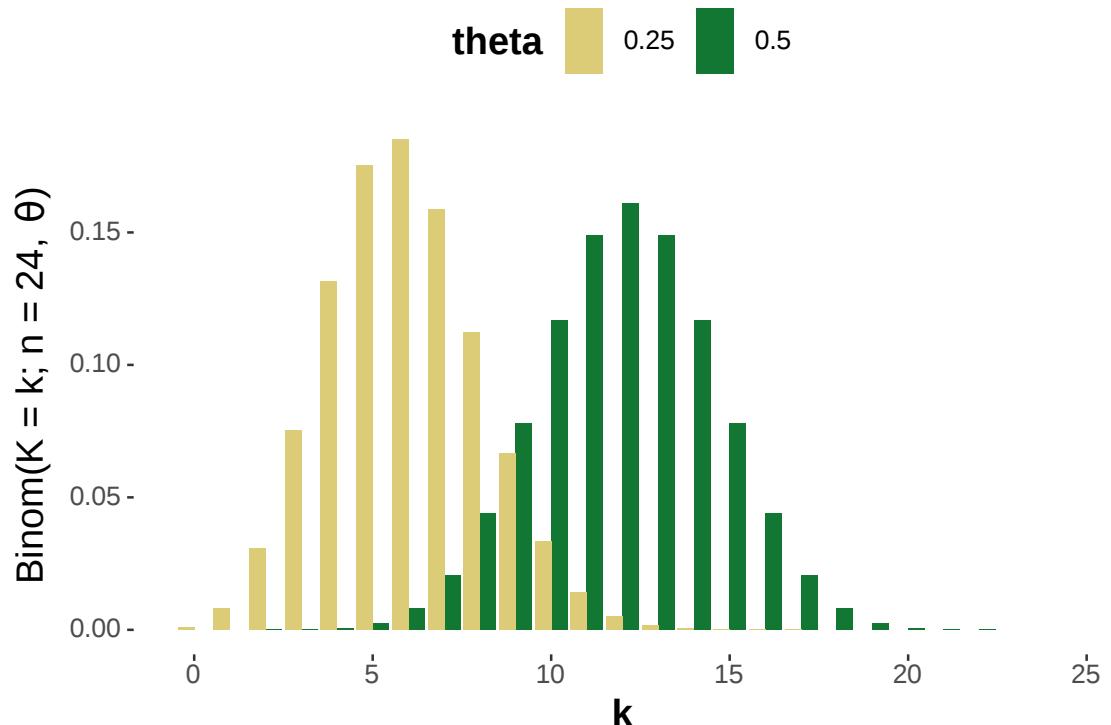


Figure 7.3.: Examples of the Binomial distribution. The y -axis give the probability of seeing k heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

For a continuous random variable X , the probability $P(X = x)$ will usually be zero: it is virtually impossible that we will see precisely the value x realized in a random event that can realize uncountably many numerical values of X . However, $P(X \leq x)$ does take workable values and so we define the cumulative distribution function F_X associated with X as:

$$F_X(x) = P(X \leq x)$$

Instead of a probability **mass** function, we derive a **probability density function** from the cumulative function as:

$$f_X(x) = F'(x)$$

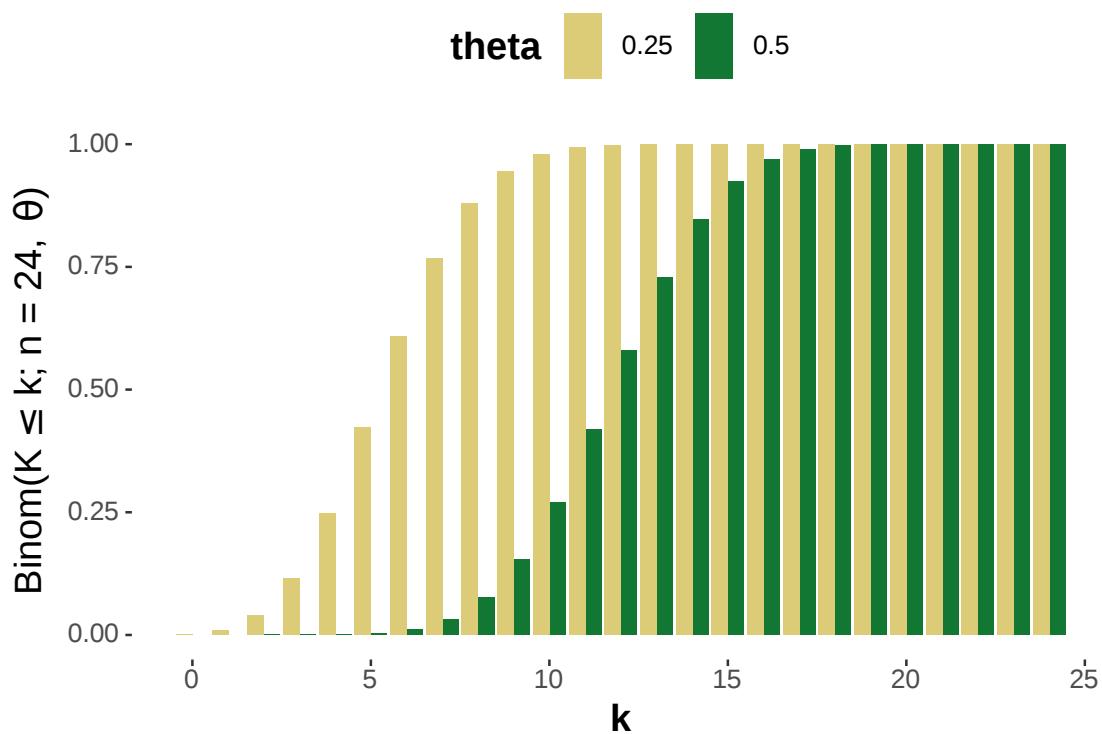


Figure 7.4.: Examples of the cumulative distribution of the Binomial. The y -axis gives the probability of seeing k or less outcomes of heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

7. Basics of Probability Theory

A probability density function can take values greater than one, unlike a probability mass function.

Example. The **Gaussian or Normal distribution** characterizes many natural distributions of measurements which are symmetrically spread around a central tendency. It is defined as:

$$\mathcal{N}(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

where parameter μ is the *mean*, the central tendency, and parameter σ is the *standard deviation*. Figure 7.5 gives examples of the probability density function of two normal distributions. Figure 7.6 gives the corresponding cumulative distribution functions.

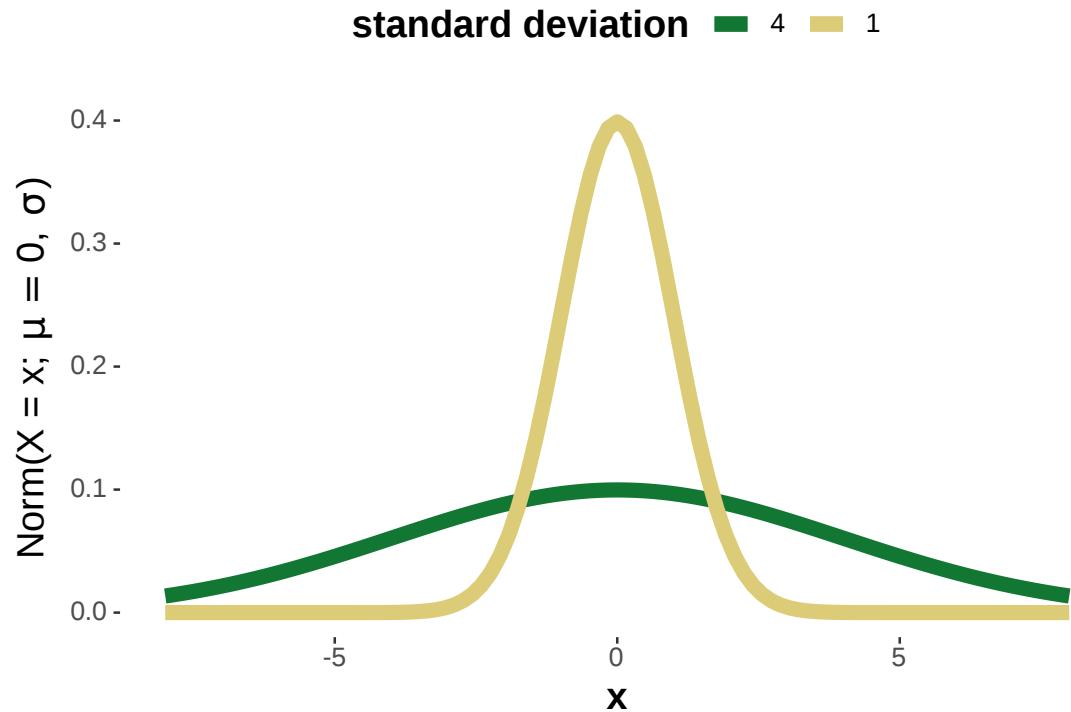


Figure 7.5.: Examples of the Normal distribution. In both cases $\mu = 0$, once with $\sigma = 1$ and once with $\sigma = 4$

7.4.3. Expected value & variance

The **expected value** of a random variable X is a measure of central tendency. It tells us, like the name suggests, which average value of X we can expect when repeatedly sampling from X . If X is continuous, the expected value is:

$$\mathbb{E}_X = \sum_x x \times f_X(x)$$

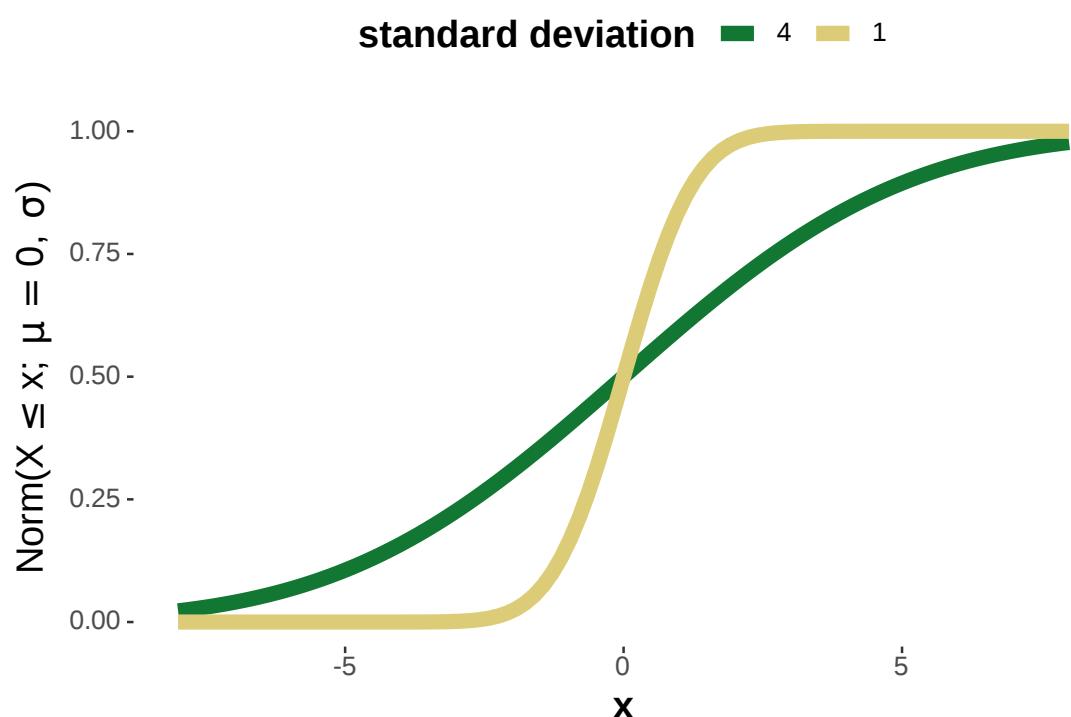


Figure 7.6.: Examples of the cumulative normal distribution corresponding to the previous probability density functions.

7. Basics of Probability Theory

If X is continuous, it is:

$$\mathbb{E}_X = \int x \times f_X(x) dx$$

The expected value is also frequently called the **mean**.

The **variance** of a random variable X is a measure of how much likely values of X are spread or clustered around the expected value. If X is discrete, the variance is:

$$\text{Var}(X) = \sum_x (\mathbb{E}_X - x)^2 \times f_X(x)$$

If X is continuous, it is:

$$\text{Var}(X) = \int (\mathbb{E}_X - x)^2 \times f_X(x) dx$$

Example. If we flip a coin with bias $\theta = 0.25$ a total of $n = 24$, we expect on average to see $n \times \theta = 24 \times 0.25 = 6$ outcomes showing heads.¹¹ The variance is $n \times \theta \times (1 - \theta) = 24 \times 0.25 \times 0.75 = \frac{24 \times 3}{16} = \frac{18}{4} = 4.5$.

The expected value of a normal distribution is just its mean μ and its variance is σ^2 .

7.4.4. Composite random variables

Composite random variables are random variables generated by mathematical operations conjoining other random variables. For example, if X and Y are random variables, then we can define a new derived random variable Z using notation like:

$$Z = X + Y$$

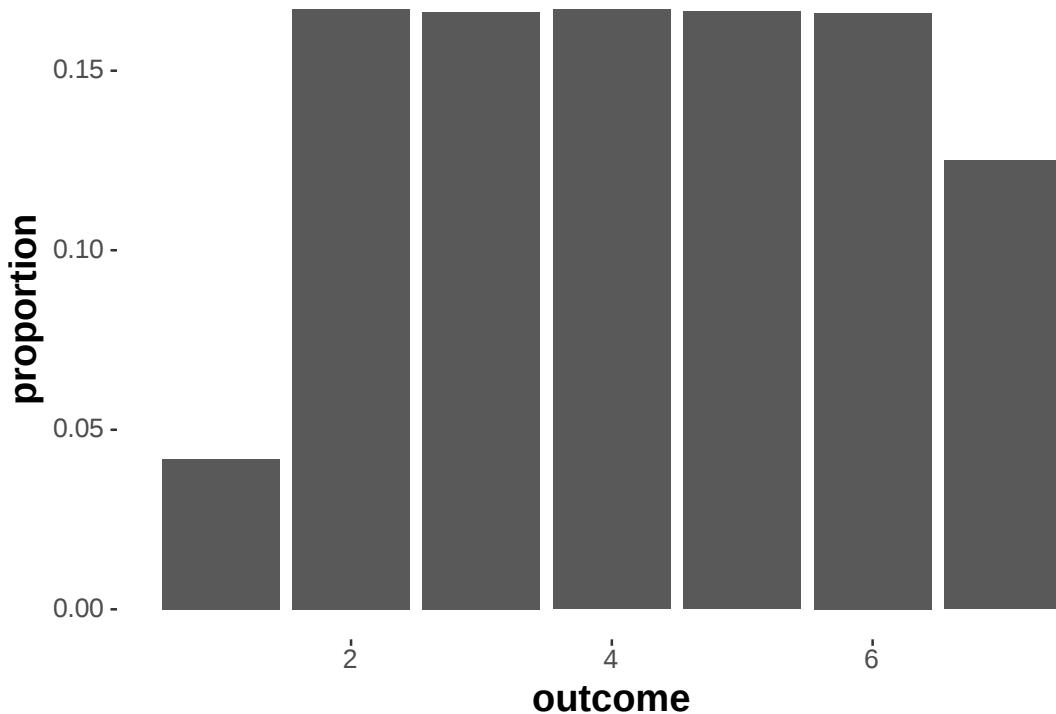
This notation looks innocuous, but is conceptually tricky yet ultimately very powerful. On the face of it, we are doing as if we are adding with $+$ two functions. But a sampling-based perspective makes this quite intuitive. We can think of X and Y are large samples, representing the probability distributions in question. Then we built a sample by just adding elements in X and Y . (If samples are of different size, just add a random element of Y to each X .)

Consider the following concrete example. X is the probability distribution of rolling a fair dice with six sides. Y is the probability distribution of flipping a biased coin which lands heads (represented as number 1) with probability 0.75. The derived probability distribution $Z = X + Y$ can be approximately represented by samples derived as follows:

```
n_samples <- 1e6
# `n_samples` rolls of a fair dice
samples_x <- sample(
  1:6,
  size = n_samples,
  replace = T
```

¹¹This is not immediately obvious from our definition, but it is intuitive and you can derive it.

```
)  
  
# `n_samples` flips of a biased coin  
samples_y <- sample(  
  c(0,1),  
  prob = c(0.25, 0.75),  
  size = n_samples,  
  replace = T  
)  
  
samples_z <- samples_x + samples_y  
  
tibble(outcome = samples_z) %>%  
  dplyr::count(outcome) %>%  
  mutate(n = n/sum(n)) %>%  
  ggplot(aes(x = outcome, y = n)) +  
  geom_col() +  
  labs(y = "proportion")
```



7.5. Probability distributions in R

Appendix B covers a number of common probability distributions that are relevant for the purposes of this course. Appendix C furthermore provides additional theoretical background on the *exponential family*, an important class of probability distributions widely used in statistics.

R has built-in functions for most common probability distributions. Further distributions are covered in additional packages. If `mydist` is the name of a probability distribution, then R routinely offers four functions for `mydist`, distinguished by the first letter:

1. `dmydist(x, ...)` the *density function* gives the probability (mass/density) $f(x)$ for x
2. `pmydist(x, ...)` the *cumulative probability function* gives the cumulative distribution function $F(x)$ for x
3. `qmydist(p, ...)` the *quantile function* gives the value x for which $p = \text{pmydist}(x, ...)$
4. `rmydist(n, ...)` the *random sample function* returns n samples from the distribution

For example, the family of functions for the normal distribution has the following functions:

```
# density of standard normal at x = 1
dnorm(x = 1, mean = 0, sd = 1)

## [1] 0.2419707

# cumulative density of standard normal at q = 0
pnorm(q = 0, mean = 0, sd = 1)

## [1] 0.5

# point where the cumulative density of standard normal is p = 0
qnorm(p = 0.5, mean = 0, sd = 1)

## [1] 0

# n = 3 random samples from a standard normal
rnorm(n = 3, mean = 0, sd = 1)

## [1] 1.1333607 -0.8555162 0.4880749
```

8. Models

Uninterpreted data is uninformative. We cannot generalize, draw estimations or attempt to make predictions unless we make (however minimal) assumptions about the data at hand: what it represents, how it came into existing, which parts relate to which other parts etc. One way of explicitly acknowledging these assumptions is to engage in model-based data analysis. **A statistical model is a conventionally condensed formal representation of the assumptions we make about what the data is and how it might have been generated.** In this way, model-based data analysis is more explicit about the analyst's assumptions than other approaches, such as test-based approaches, which we will encounter in Chapter 10.

There is room for divergence in how to think about a statistical model, the assumptions it encodes and truth. Some will want to reason with models using language like "if we assume that model M is true, then ..." or "this shows convincingly that M is likely to be the true model". Others feel very uncomfortable with such language. In times of heavy discomfort they might repeat their soothing mantra:

All models are wrong, but some are useful. – Box (1979)

To become familiar with model-based data analysis, Section 8.1 introduces the concept of a **probabilistic statistical model**. Section 8.2 enlarges on the crucial aspects of parameters and priors. Section 8.3 introduces the three types of statistical goals that we can use models to help us with. Section 8.4 expands on the notation, both formulaic and graphical, which we will use in this book to communicate about models. Finally, Section 8.5 discusses examples of interesting models, some of which we will use frequently in the following chapters.

The learning goals for this chapter are:

- become acquainted with statistical models
- understand what parameters are and what priors can do
- meet pivotal exemplars:
 - Binomial Model, T-Test Model, Simple Linear Regression
- understand notation to communicate models
 - formulas & graphs

8.1. Probabilistic models in statistics

In its most common natural sense, a "model" is a model of something. It intends to represent something else in a condensed, abstract and more practical form; where what is practical is conditioned by a given pur-

8. Models

pose. Often the purpose of a model is epistemic, i.e., related to knowledge gain or a deeper understanding of the world: even a model plane arguably serves the epistemic purpose of getting a better idea of what real planes look like; a model of bridge to be constructed helps us imagine how the real thing would turn out to be; this is how models differ from a more ordinary picture or sculpture without such an epistemic purpose. For any given purpose, a good model will try to represent some aspects of reality and abstract away from irrelevant features which might otherwise blur our vision.

A statistical model is a model of (what we imagine to be) a random process R . In most common parlance, however, we often speak of “a model of the data” or of “modeling the data”, but this is only sloppy slang for “a model of (what we assume is) a random process that could generate data of this kind”.

Let D be (rectangular, tidy) data that represents the kind of data that random process R is assumed to generate. A model M for random process R fixes which variables (columns) of D are to be modelled as dependent and which are independent variables (and which do not matter at all). Let D_{DV} by the subset of D containing the dependent variables and D_{IV} the subset of D containing the independent variables.¹ We want D_{DV} to contain at least one variable. A model with empty D_{IV} is fine.

A model M for data D also fixes a **likelihood function** for D_{DV} . The likelihood function determines how likely any potential data observation D_{DV} is, given the corresponding observations in D_{IV} . Most often, the likelihood function also has **free parameters**, represented by a parameter vector θ . The basic (and yet rather uninformative) notation for a likelihood function of model M for data D with parameter vector θ is therefore:

$$P_M(D_{DV} \mid D_{IV}, \theta)$$

Bayesian models have an additional component, namely a **prior distribution** over parameter values, commonly written as:

$$P_M(\theta)$$

In sum, let's adopt the following definition. A **statistical model** M for data D consists of:

1. a selection of (disjoint) sets of dependent and independent variables D_{DV} and D_{IV} , where the latter is possibly empty, but the former is not; and
2. a parameterized likelihood function: $P_M(D_{DV} \mid D_{IV}, \theta)$.

A statistical model is a **Bayesian model** if it also contains:

3. a prior distribution: $P_M(\theta)$.

The Bayesian prior over parameter values can be used to regularize inference and/or to represent any motivated and justifiable *a priori* assumptions about parameter values that are plausible given our knowledge so far. The next section elaborates on parameters, priors and key differences between frequentist and Bayesian models. But first, we should take a look at two simple examples of models.

¹Naturally, D_{DV} and D_{IV} are disjoint: it makes no sense to predict or explain x based on an observation of x .

8.1.1. Example 1: a single draw from an urn

In front of us is an urn. We cannot see what is inside. We assume (!) that there are $N = 10$ balls in the urn and that any number $0 \leq k \leq 10$ is black, the rest white. Our data is minimal. There is only one variable, which therefore is also our dependent variable. We have drawn a ball from the urn once, and we observed that it is black.

```
minimal_data_from_an_urn <-
  tribble(~ draw, c("black"))
```

So far, so boring. But what is a reasonable parameterized likelihood function for this case? – Well, we do not know what the content of the urn is but, given our (modelling) assumptions, there are only eleven possible states of the world $k \in \{0, 1, \dots, 10\}$. If we assume (as part of the model structure) that the total number of balls N in the urn is known $N = 10$, then the number k of black balls in the urn straightforwardly entails the likelihood of the data:

$$P_M(D = \text{"black"} \mid k) = \frac{k}{N}$$

It is important to realize that this (and any other likelihood function) defines the probability, not only of the observed data, but for the whole class of *observable data*, including observations that are only logically conceivable, but possibly ruled out by the model.

The model of the single-draw random process has a single free parameter k , which feeds into the likelihood function. We naturally think of the likelihood of the data as probabilistically dependent on the parameter value k .

A Bayesian model of this situation would additionally also include a *prior* over parameter values. There are only eleven possible values for k , so this is a discrete probability distribution. If we do not have any relevant *a priori* knowledge of the process, we might want to assign the same probability to each value of k :

$$P_M(k = i) = \frac{1}{11}; \text{ for all } i \in \{0, 1, \dots, 10\}$$

8.1.2. Example 2: avocado prices by type

We must also consider a slightly less minimalist example. The avocado data set is useful for that. As before, we load the data into a variable named `avocado_data` and do some minor data wrangling (see also Appendix Chapter D.6):

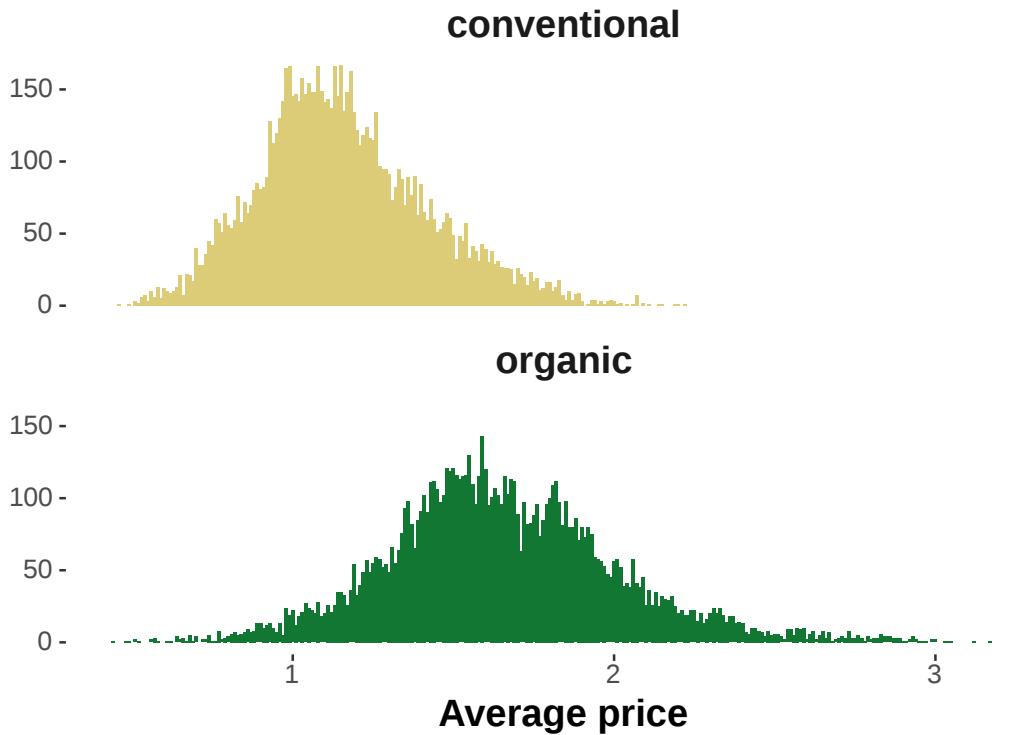
```
avocado_data <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-data-anal...'))
# remove currently irrelevant columns
select( -X1 , - contains("Bags") , - year , - region) %>%
```

8. Models

```
# rename variables of interest for convenience
rename(
  total_volume_sold = `Total Volume`,
  average_price = `AveragePrice`,
  small = '4046',
  medium = '4225',
  large = '4770',
)
```

We are interested in the random process that generates avocado prices. The data relevant for modeling this random process contains `average_price` as the dependent variable and `type` as the independent variable. We could also say that we are interested in predicting / explaining the average prices in terms of the avocado type. To get a feeling for how the data to be modeled looks like, here are histograms for the price data for conventionally and organically grown avocados:

```
avocado_data %>%
  ggplot(aes(x = average_price, fill = type)) +
  geom_histogram(binwidth = 0.01) +
  facet_wrap(type ~ ., ncol = 1) +
  ylab('') +
  xlab('Average price') +
  theme(legend.position = "none")
```



Our model assumes that the data observations in `average_price` are samples from a normal distribution, whose mean μ and standard deviation σ are free parameters, one pair of μ and σ for each type of avocado. So, this model has four free parameters, which constitute the parameter vector $\theta = \langle \mu_c, \sigma_c, \mu_o, \sigma_o \rangle$. As for the likelihood function, if \vec{y} is the vector `average_price`, so that $y_i \in \mathbb{R}^+$ is the average price observed in row i , and if \vec{x} is an indicator variable such that $x_i \in \{1, 0\}$ is the entry for the type of avocado in line i where 1 represents conventionally grown and 0 represents organically grown, and if there are k rows in the data set, the likelihood function can be written as:

$$P_M(\vec{y} | \vec{x}, \theta) = \prod_{i=1}^k x_i \text{Normal}(y_i, \mu_c, \sigma_c) + (1 - x_i) \text{Normal}(y_i, \mu_o, \sigma_o)$$

If we aspire to handle a Bayesian model, we need to supply a prior for parameters θ as well. General strategies of fixing priors for Bayesian data analysis are discussed in the next subsection. To give a concrete example for this case, we could assume that all parameter vectors are independent of each other and assume that the means μ_c and μ_o are themselves normally distributed. We could use a *truncated normal distribution*² as the priors for the standard deviations σ_c and σ_o :

²A truncated normal distribution is like a normal distribution, but restricted to a certain range of possible values. In general, if P is a continuous probability distribution on some interval which properly contains $[a; b]$, a truncated version of P to the interval

8. Models

$$\begin{aligned}
P(\mu_c, \sigma_c, \mu_o, \sigma_o) &= P(\mu_c) P(\sigma_c) P(\mu_o) P(\sigma_o), \text{ where} \\
P(\mu_c) &= \text{Normal}(\mu_c, \mu = 1.5, \sigma = 0.25) \\
P(\mu_o) &= \text{Normal}(\mu_o, \mu = 1.5, \sigma = 0.25) \\
P(\sigma_c) &= \text{Trunc-Normal}(\sigma_c, \mu = 0.2, \sigma = 0.05, \text{lower} = 0) \\
P(\sigma_o) &= \text{Trunc-Normal}(\sigma_o, \mu = 0.25, \sigma = 0.1, \text{lower} = 0)
\end{aligned}$$

8.2. Parameters, priors, probability and predictions

We defined a model as containing a parameterized likelihood function, and, if it is a Bayesian model, also a prior distribution over parameter values:

$$\begin{aligned}
\text{Likelihood: } &P_M(D \mid \theta) \\
\text{Prior: } &P_M(\theta) \quad [\text{if Bayesian}]
\end{aligned}$$

More needs to be said about what a parameter is, what a prior distribution $P_M(\theta)$ is, and what the difference is between a non-Bayesian / frequentist model without priors over θ and a Bayesian model with priors over θ .

The running example for this section is the **Binomial Model**, which is also included in the cast of fine characters in Section 8.5.³ The Binomial Model is a generalization of the urn-model covered in the previous section. We imagine a flip of a coin with bias $\theta \in [0; 1]$, which is flipped N times. We are interested in the number k of heads (where each outcome of heads is represented as the number 1, and a tails outcome is represented by number 0). The likelihood function for this model is the Binomial distribution:

$$P_M(k \mid \theta, N) = \text{Binomial}(k, N, \theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}$$

As a concrete example, we consider a case with $N = 24$ flips and $k = 7$ head outcomes.

8.2.1. What's a model parameter?

Parameters are defined by their role: changing a model parameter changes the likelihood of the data. Figure ?? shows the likelihood function of the Binomial Model for $\theta \in [0; 1]$.

Several things are important to note, even though their significance might only reveal itself in full much later. Firstly, since there is a direct influence of a model parameter on the likelihood of data observations,

$[a; b]$ such that:

$$\text{Trunc-}P(x, \min = a, \max = b) = \begin{cases} \frac{P(x)}{\int_a^b P(x') dx'} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

³We should actually speak of a class of (infinitely many) models, all sharing the same likelihood function. Sloppily, we still speak of "the" Binomial Model.

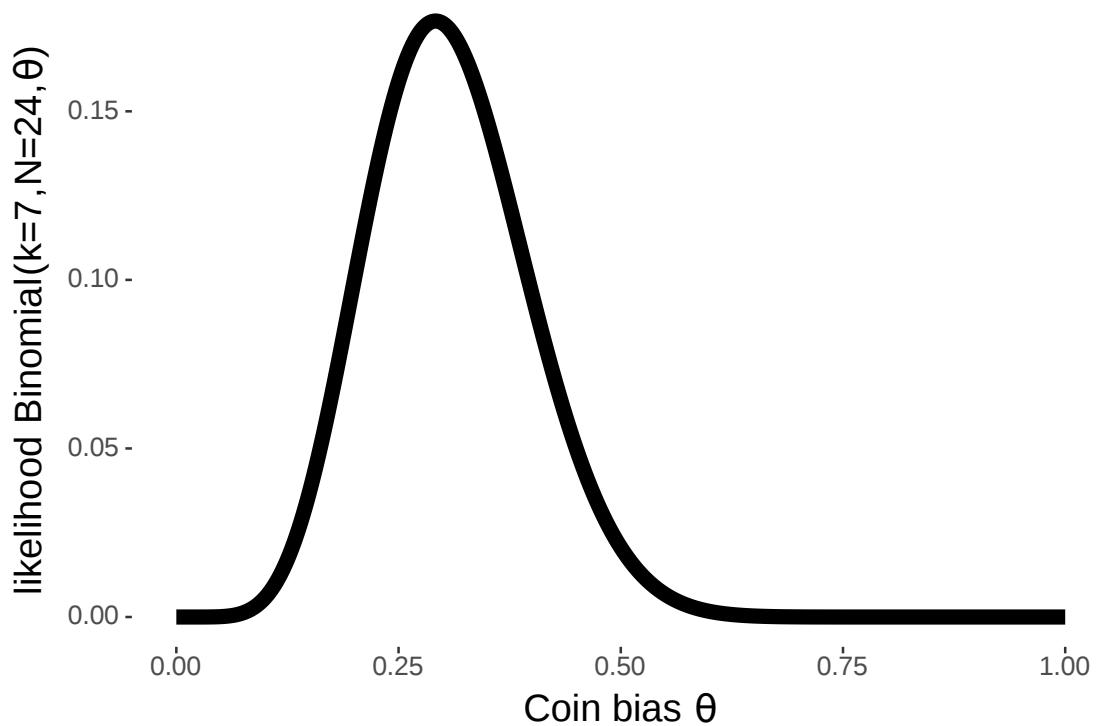


Figure 8.1.: Likelihood function for the Binomial Model, for $k = 7$ and $N = 24$.

8. Models

we might be able to express subjective beliefs about that parameter value, even if the parameter itself is not clearly interpretable in intuitive ways. In other words, we might not have intuitions about numerical values of abstract parameters, but we might nonetheless hold beliefs about the likelihood of the data that results from setting abstract parameter values one way or another. Therefore, we might get access to our (implicit) beliefs about parameter values via (more explicable) intuitions about the data. Secondly, not all parameters are equal. Some parameters may have a much stronger effect on the likelihood than others (all else equal). Just counting the number of model parameters as an indicator of model complexity will therefore not be enough. (This will be important when we look at different ways of comparing models in Section 11.)

8.2.2. Priors over parameters

The prior distribution over parameter values $P_M(\theta)$ is an integral part of a model, when we adopt a Bayesian approach to data analysis. This entails that two (Bayesian) models can share the same likelihood function, and yet ought to be considered as different models.

In Bayesian data analysis priors $P_M(\theta)$ are most saliently interpreted as encoding the modeller's prior beliefs about the parameters in question. As mentioned, this need not necessarily be beliefs derived from fundamental convictions about coin biases, means or standard deviations; these could be beliefs held because of the effects certain parameter values have on the likelihood of the data. Ideally, the beliefs that support the specification of a prior should be supported by argument, results of previous research or other clearly justifiable motivations. But these informed subjective priors are really just one way in which priors over parameters can be justified.

There are three main types of motivations for priors $P_M(\theta)$, and it is possible that these motivations are mixed in the choice of a particular prior for a particular application:

1. **Subjective priors** capture the modeller's subjective beliefs in the sense described above.
2. **Objective priors** are priors that, as some argue, *should* be adopted for a given likelihood function to avoid conceptually paradoxical consequences.
3. **Practical priors** are priors that are used pragmatically because of their specific usefulness, e.g., because they simplify a mathematical calculation or a computer simulation, or because they help in statistical reasoning, such as when *skeptical priors* are formulated that work against a particular conclusion.

Orthogonal to the kind of motivation given for a prior, we can distinguish different priors based on how strongly they commit the modeller to a particular range of parameter values. The most extreme case are **uninformative priors** which assign the same level of credence to all parameter values. Uninformative priors are also called *flat priors* because they express themselves as flat lines for discrete probability distributions and continuous distributions defined over an interval with finite lower and upper bounds. It is possible to maneuver uninformative priors also for continuous distributions defined over an unbounded interval, in which case we speak of *improper priors* (to remind ourselves that, mathematically, we are doing something tricky). Informative priors, on the other hand, can be *weakly informative* or *strongly informative*, depending on how much commitment they express. The most extreme case of commitment would be expressed in a **point-valued prior**, which puts all probability on a single value of a parameter. Since this is no longer a respectable probability distribution, although it satisfies the definition, we speak of a *degenerate prior* here.

Figure 8.2 shows examples of objective, uninformative, as well as weakly or strongly informative priors for the Binomial Model.⁴ The priors shown here (resulting in four different Bayesian models all falling inside the family of Binomial Models) are as follows:

- *objective* : $\theta \sim \text{Beta}(0.5, 0.5)$
- *uninformative* : $\theta \sim \text{Beta}(1, 1)$
- *weakly informative* : $\theta \sim \text{Beta}(2, 3)$
- *strongly informative* : $\theta \sim \text{Beta}(20, 30)$

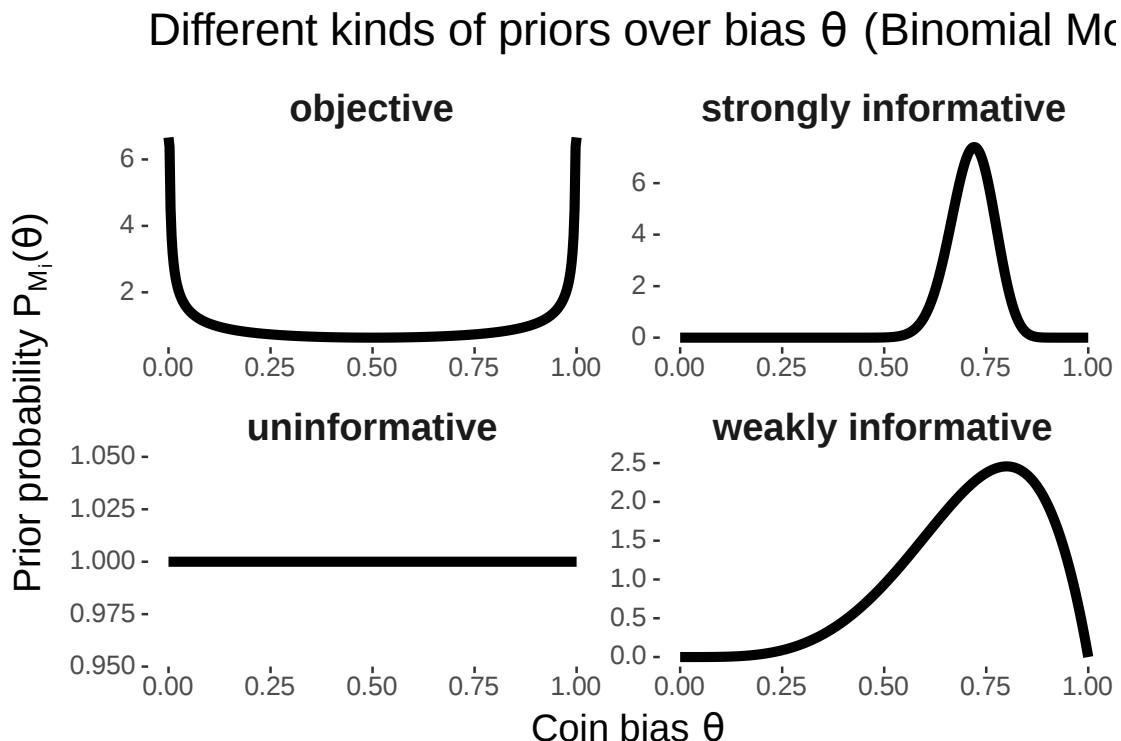


Figure 8.2.: Examples of different kinds of Bayesian priors for coin bias *theta* in the Binomial Model.

8.2.3. Two notions of probability (revisited)

To understand why frequentist models do not have priors, but Bayesian models do, we need to revisit the interpretation of the notion of probability, in particular frequentism and subjectivism. We should be reminded that although both notions imply different approaches to the problem of how to deal with probabilities, the mathematical properties are quite similar (Kruschke 2015).

⁴We will not go into the topic of objective priors in this course. This is a topic we must reserve for a follow-up course. It is therefore also not important to “see” from Figure 8.2 how or why the curve shown in an objective prior.

8. Models

A crude and overly simplified version of why frequentist models do not contain priors is this. Extreme frequentism denies that a probability distribution over a latent parameter like θ is meaningful. It cannot be justified or defended in a scientifically rigorous manner. The only statements about probabilities that are conceptually sound, according to a fundamentalist frequentist interpretation, are those that derive from intuitions about limiting frequencies when (hypothetically) performing a random process (like throwing a dice or drawing a ball from an urn). Bluntly put, there is no “(thought) experiment” which can be repeated so that its objective results, on average, align with whatever subjective prior beliefs the Bayesian analysis needs. As a result, the frequentist approach to statistical inference needs alternative methods for parameter estimation – methods that do *not* rely on (subjective) priors $P(\theta)$.

Of course, the objections to the use of priors could be less fundamentalist. Researchers who have no metaphysical troubles with subjective priors in principle might reject the use of priors in data analysis because they feel that the necessity to specify priors is a burden or a spurious degree of freedom in empirical science that is best avoided in order to stay as objective, procedurally systematic, defaultist and streamlined as possible.

In contrast, adopters of Bayesian practices see added value in using priors (expression of prior knowledge, practicality) that exceeds any conceptual or procedural difficulties they see in the process of specification of priors. When uncertain about a particular choice of prior, a good Bayesian practice is a *robustness analysis*: exploring how conclusions or decisions actually depend on the specification of the prior. As we will see, given enough data, the effect of prior specifications can be washed out. This is intuitive: no matter what you believe initially, given enough empirical evidence you will eventually be persuaded to leave your original beliefs behind.

8.2.4. Prior predictions

The inclusion or omission of priors also has straightforward technical consequences. Bayesian models make predictions about how likely a particular data outcome is, even before having seen any data at all. Frequentist models do not (unless they assume uninformative priors implicitly, making them a Bayesian model).

The (Bayesian) **prior predictive distribution** of model M is a probability distribution over future or hypothetical data observations:

$$P_M(D) = \sum_{\theta} P_M(D | \theta) P_M(\theta) \quad [\text{discrete parameter space}]$$
$$P_M(D) = \int P_M(D | \theta) P_M(\theta) d\theta \quad [\text{continuous parameter space}]$$

The formula above is obtained by marginalization over parameter values (represented here as an integral for the continuous case). If we do not want to commit to any relative weighing of parameters, not even an uninformative equal weighing of all parameter values, there is no way in which we can derive predictions of a model.

In the case of the Binomial Model when we use a Beta prior over θ , the prior predictive distribution is the prominent that it has its own name and fame. It's called the Beta-binomial distribution. Figure 8.3 shows

the prior predictives for the four kinds of priors we looked at before when $N = 24$.

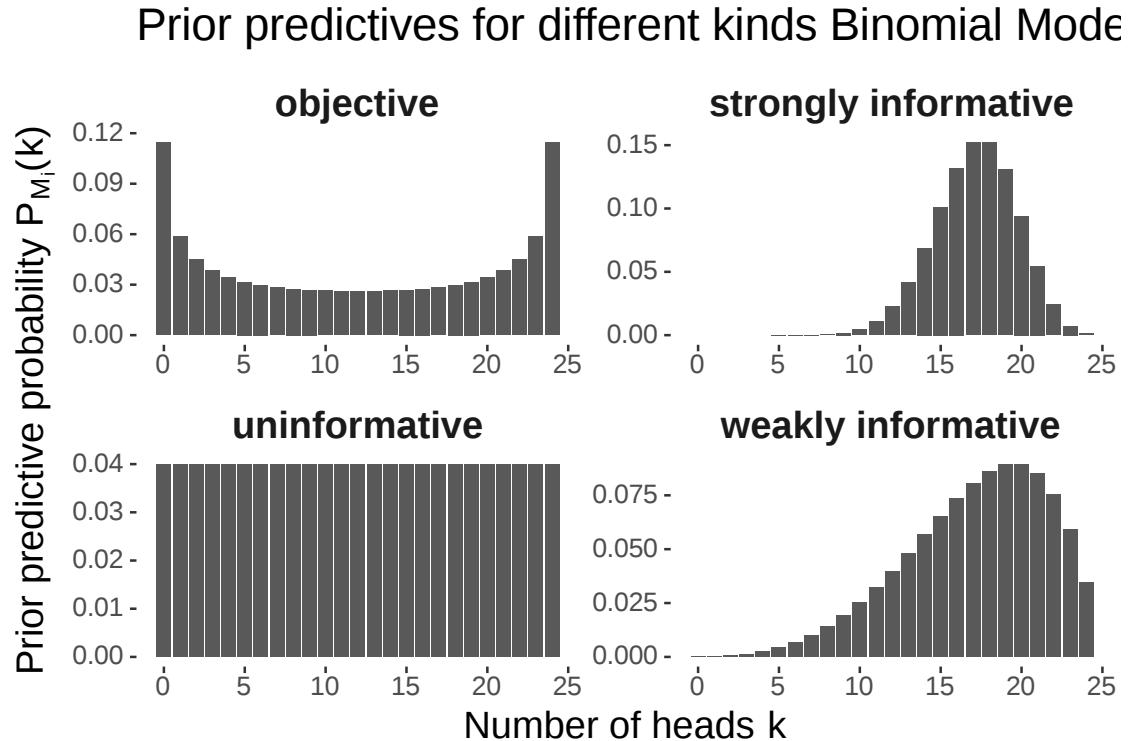


Figure 8.3.: Prior predictive distributions for Binomial Models with different kind of Beta-priors.

8.3. Three pillars of data analysis

There are three main uses for models in statistical data analysis:

1. **Parameter estimation:** Based on model M and data D , we try to infer which value of the parameter vector θ we should believe in, or work with (e.g., base our decision on). Parameter estimation can also serve knowledge gain, especially if (some component of) θ is theoretically interesting.
2. **Model comparison:** If we formulate at least two alternative models, we can ask which model better explains or better predicts some data. In some of its guises, model comparison helps with the question of whether a given data set provides evidence in favor of one model and against another other, and if so, how much.
3. **Prediction:** Models can also be used to make predictions about future or hypothetical data observations.

The frequentist and the Bayesian approach each have their own methods and techniques to do estimation, comparison and prediction. Even within each approach (frequentist or Bayesian) and a particular goal (estimation, comparison or prediction) there is not necessarily unanimity about the best method or technique.

8. Models

Table 8.1.: Most common/salient methods of frequentist and Bayesian approaches for the three major goals of model-based data analysis. The abbreviations used are: MLE for 'maximum likelihood estimate', AIC for 'Akaike information criterion', LR-test for 'likelihood-ratio test' and D_{rep} for 'repeat data'.

inferential goal	target	frequentist	Bayesian
estimation	θ	MLE: $\hat{\theta} = \arg \max_{\theta} P_M(D \theta)$	posterior: $P_M(\theta D)$
comparison	M	AIC, LR-test	Bayes factor
prediction	D	MLE-based: $P_M(D_{rep} \hat{\theta})$	Posterior-based: $P_M(D_{rep} D)$

Table 8.1 lists the most common / salient methods used for each goal in the frequentist and Bayesian approach. Large part of the remainder of this course will be dedicated to understanding the methods name-dropped in this table, and to compare them against each other and further, as of yet unmentioned alternatives. Chapter 11 deals with model comparison, Chapter 9. The second main goal of this course is to understand the relation between model-based data analysis, as summarized in Table 8.1, to test-based approaches as described in Chapter 10 and Chapter 12.

The three pillars of data analysis mentioned above are tightly related, of course. For one, model comparison is often parasitic on prediction: whereas prediction asks which data is to be expected, given the model, model comparison looks at how well a given data set is or would have been predicted by different models. For another, parameter estimation and data predictions are something like each other's reverse operations. Let's elaborate on the latter briefly.

If we flip a coin flip once, the likelihood of each outcome $x \in X = \{0; 1\}$ (representing heads or tails, encoded as 1 and 0, respectively) can be modeled with the **Bernoulli distribution** as follows, where $\theta \in [0; 1]$ is the coins bias towards landing heads:

$$P_M(X = x | \theta) = \theta^{[x]}(1 - \theta)^{1-[x]}$$

This likelihood function relates two variables of interest: the coin flip outcome (= data D) x and the coin's bias (= model parameter θ). Depending on what is given or assumed to be known, we can then use the same likelihood function, to either infer something about the unknown parameter θ or, when θ is given make predictions about the unknown outcome x .

To make this even clearer, we can temporarily use the following bracket notation: the bracket $[\cdot]$ indicates that the bracketed parameter is treated as known, given or assumed. The **predictive distribution** for unknown data x is then:

$$\text{Predictive Distribution: } F(\theta) = P_M(X = x | [\theta]) = [\theta]^x(1 - [\theta])^{1-x}$$

But often the contrary is the case, that is one is interested in the value of θ by a given data set. Then θ is unknown and the data are observed. Treating θ as parameter instead of x leads to the *likelihood function* –

a mathematical formula that specifies the plausibility of the data as a function of θ . It states the probability of any possible observation:

$$\text{Likelihood Function: } F(x) = P_M([X = x] \mid \theta) = \theta^{[x]}(1 - \theta)^{(1-[x])}$$

Figure 8.4 shows the likelihood function associated with the one-coin-flip model when the observed and known outcome is $x = 1$ (heads). Notice that the likelihood function is not a probability distribution and thus does not necessarily integrate or sum to 1, even though it does in the case at hand.

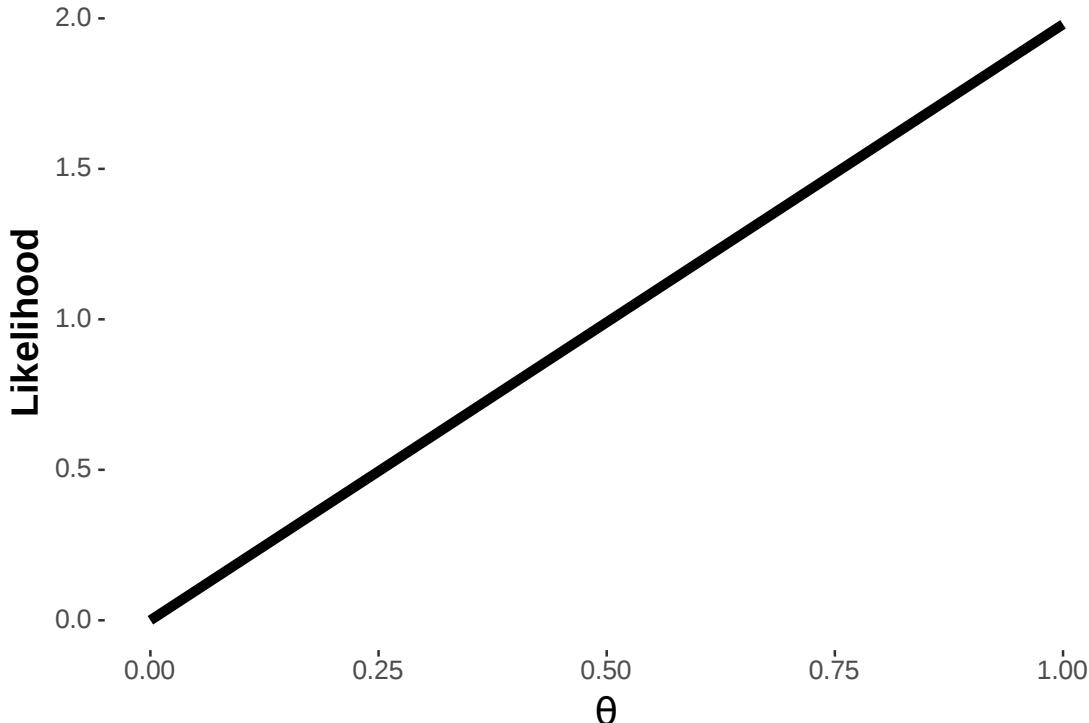


Figure 8.4.: Bernoulli likelihood for the one-coin-flip-model with observed outcome $x = 1$ (heads).

8.4. Notation & graphical representation

If it is important to communicate the assumptions underlying a statistical argument, and if models are means of making formally explicit these assumptions, then it follows that efficient communication of models is important too. We here follow common practice of representing models using both a special purpose formulaic notation and, where useful, a graph-based visual display in which probabilistic dependencies are lucidly represented.

The running example for this section is the **Binomial Model**, which is also included in the cast of main characters in Section 8.5. The Binomial Model is a generalization of the urn-model covered earlier in this

8. Models

chapter. We imagine a flip of a coin with bias $\theta \in [0; 1]$, which is flipped N times. We are interested in the number k of heads (represented as an outcome of 1). The likelihood function for this model is the Binomial distribution:

$$P_M(k \mid \theta, N) = \text{Binomial}(k, N, \theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}$$

For purposes of illustration we use a Beta distribution for the prior of θ , but set its parameters so that the ensuing distribution is flat (uninformative priors):

$$P_M(\theta) = \text{Beta}(\theta, 1, 1)$$

8.4.1. Formula notation

To concisely represent models, we use a special notation, which is very intuitive when we think about sampling. Instead of the above notation for the prior we write:

$$\theta \sim \text{Beta}(1, 1)$$

The symbol “ \sim ” is often read as “is distributed as”. You can also think of it as meaning that θ is sampled from a uniform distribution.

Similarly, for the likelihood function, we would just write:

$$k \sim \text{Binomial}(\theta, N).$$

8.4.2. Graphical notation

When models get very complex and incorporate many parameters it can be difficult to tease out all relations between the model components. In such a situation a graphical notation of a model might be helpful. We here adopt the convention described in Wagenmakers and Lee’s *Bayesian Cognitive Modelling* (2014): the graph structure is used to indicate dependencies between the variables, with children depending on their parents (Lee and Wagenmakers 2014). We represent every relevant variable as a node in a directed acyclic graph structure (a probabilistic network). In visualizing this, we use the following general conventions:

- known or unknown (= latent) variable
 - *shaded nodes*: observed variables
 - *unshaded nodes*: unobserved / latent variables
- kind of variable:
 - *circular nodes*: continuous variables
 - *square nodes*: discrete variables

- kind of dependency:
 - *single line*: stochastic dependency
 - *double line*: deterministic dependency

For the Binomial Model this results in the relevant variables:

- number of trials (N)
- number of success (k)
- probability for a success (θ)

Of these N and k are observed and discrete variables, and θ is a latent continuous variable. Clearly, the number of heads k depends on the coin bias θ as well as on the number of trials N . This yields a graphical and formulaic notation as in Figure 8.5.

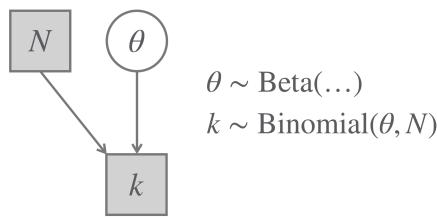


Figure 8.5.: The Binomial Model.

8.4.3. Multiple observations

When we have several observations, like in the avocado price data set, we use indexed variables. For example, the likelihood function handled in Section 8.1 for modeling avocado prices as a function the type of avocado was previously written like so:

$$P_M(\vec{y} \mid \vec{x}, \theta) = \prod_{i=1}^k x_i \text{Normal}(y_i, \mu_c, \sigma_c) + (1 - x_i) \text{Normal}(y_i, \mu_o, \sigma_o)$$

An alternative notation for this likelihood function achieves conciseness with the help of the \sim symbol:

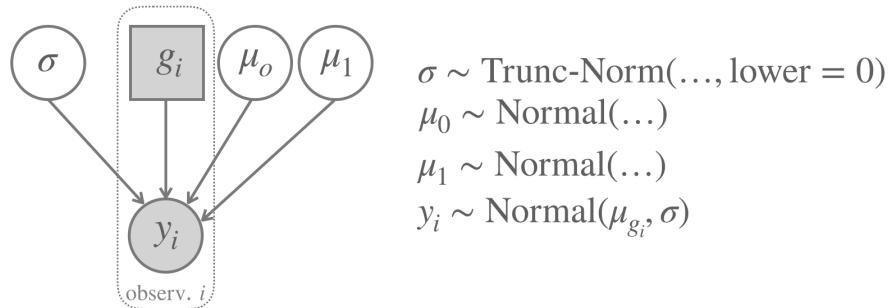
$$y_i \sim \begin{cases} \text{Normal}(\mu_c, \sigma_c) & \text{if } x_i = 1 \\ \text{Normal}(\mu_o, \sigma_o) & \text{otherwise} \end{cases}$$

Notice that this latter notation makes clearer than the previous that each observation in vector \vec{y} is an **independent draw** from a given distribution in the sense that the likelihood of y_i does not depend on the value of any other y_j . We see independence of each y_i from any other observation y_j in the notation $y_i \sim \dots$ when the distribution on the RHS does not contain any reference to y_j , as is the case in this example. If all y_i are independent samples, we also understand implicitly from this notation that the likelihood of the whole

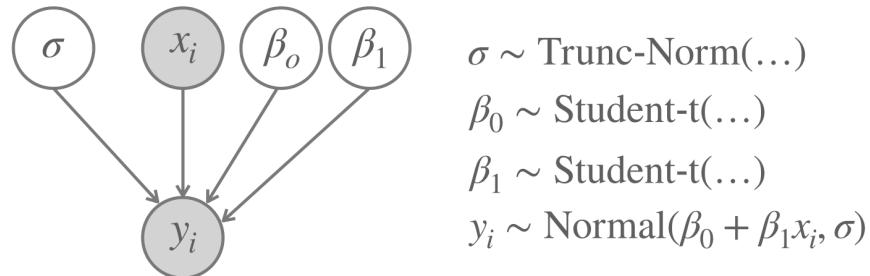
8. Models

data is to be calculated as the product of the likelihood of each individual observation. This is because the probability of the conjunction of two stochastically independent events is the product of their individual probabilities (see Chapter 7.3.2).

In the graphical representation, we may conveniently group variables with the same index together in dotted boxes, like shown below for the T-Test Model, which is introduced in the next section:



The T-Test Model shown above (and elaborated on below) uses different probability distributions for each (independent) draw of y_i . This is, of course, not always the case. The graph below shows the structure of a Linear Regression Model (explained in more detail in the next section), where each y_i is assumed to be an *independent* draw from the *same* distribution. This is often written as "**iid**": independent of all other observations and identically distributed just like all other observations.⁵



8.5. Strolling the zoo of models

Let's have a look at some more models. Some of the characters we will encounter here, will play leading roles in the plot to come. Some are for familiarization and exercise. All are pleasant.

8.5.1. The Binomial Model

We just repeat the Binomial Model discussed above for completeness in Figure 8.6.

⁵With slight but pragmatically justifiable abuse of terminology, we could still speak of the observations y_i in the T-Test Model as iid (independently and identically distributed) if we contextually enrich the intended meaning of "identically distributed" in the obvious way to mean "identically distributed given the group the observation i belongs to".

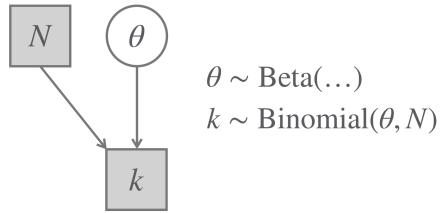


Figure 8.6.: The Binomial Model (repeated from before).

8.5.2. Flip-and-Draw Model

The Flip-and-Draw Model is a model for the flip-and-draw scenario introduced in Section 7.2. Remember that we first flip a coin and then draw from one of two urns, depending on the outcome of the coin flip. Let's generalize this and assume that the coin has a possibly unknown bias θ . The probability of sampling a black or white ball from the first urn is given by a probability vector $\vec{p}_0 = \langle 0.2, 0.8 \rangle$ (where the first entry gives the probability of sampling a black ball). The other urn has a corresponding probability vector $\vec{p}_1 = \langle 0.4, 0.6 \rangle$. The categorical distribution gives the probability of sampling an index from a given probability vector \vec{p} . It is defined as:

$$\text{Categorical}(i) = \vec{p}_i$$

Figure 8.7 gives a concise representation of the model.

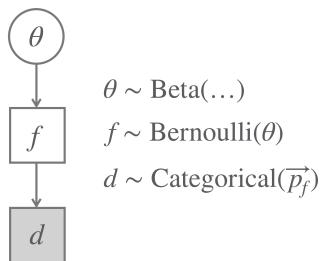


Figure 8.7.: The Flip-and-Draw Model.

8.5.3. Flip-and-Draw-Hypergeometric Model

The Flip-and-Draw-Hypergeometric Model is exactly like the previous Flip-and-Draw Model, except that we allow ourselves to sample repeatedly *without replacement* from urn the coin flip selected for us. The probability of observing k black balls when drawing n balls from an urn which contains N balls in total out of which K balls are black, when we do not put each drawn ball back into the urn is described the the so-called hypergeometric distribution. The hypergeometric distribution assign to each $k \in \{\max(0, n+K-N), \dots, \min(n, K)\}$ the probability mass:

8. Models

$$\text{Hypergeometric}(k \mid n, K, N) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

Suppose we keep the total number of urns fixed to $N = 100$ and always draw $n = 20$ balls. The two urns we have (indexed 0 and 1) hold $K_0 = 20$ and $K_1 = 80$ black balls. For a Bayesian model we could use a Beta prior for θ . Figure 8.8 shows the model.

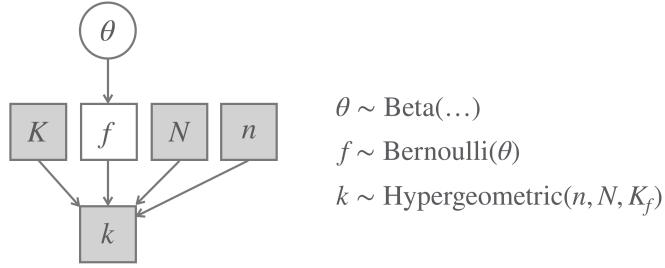


Figure 8.8.: The Flip-and-Draw-Hypergeometric Model.

Below is WebPPL code that allows you to test the models' predictions about data. You can manipulate the parameter values for θ and K_1 and K_2 . The latter are represented in the array / vector $K = [K_1, K_2]$ below.⁶

WebPPL execution from inside the web-book is currently broken, but we are working on this. You can copy-paste the code into the code box at webppl.org.

8.5.4. T-Test Model: comparing two groups

Consider the example introduced in Section 8.1 again, where we set out to compare the means of average prices of different types of avocados. We assume that there is an indicator variable g_i (corresponding to type), such that, for example, $g_i = 0$ is for observations from organic avocados and $g_i = 1$ is for observations from conventionally grown avocados. A first instance from the T-Test Model family for the comparison of two groups assumes that the first group has a mean μ_0 and the second has μ_1 , while both share the same standard deviation σ . Using the notation developed in Section 8.4 we are able to write the likelihood function succinctly like so:

$$y_i \sim \text{Normal}(\mu_{g_i}, \sigma)$$

In a Bayesian approach, the priors over parameter values could use a normal distribution for μ_i and a truncated normal distribution for σ (since σ must be positive). Figure 8.11 shows the resulting model.

A point of general importance can be made in connection to this example. It is possible to build different prior structures around the same likelihood function. Some parameterizations can be more useful for a

⁶To learn about WebPPL the fast way, try this tutorial.

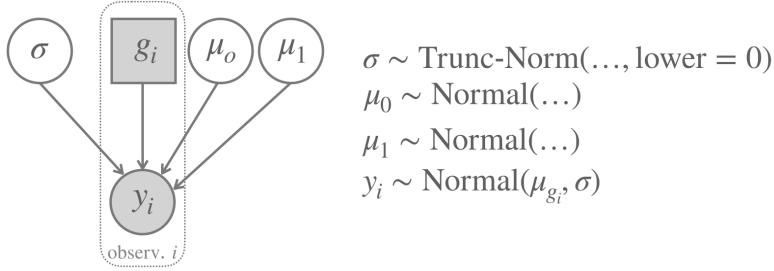


Figure 8.9.: A T-Test Model where each group has its own mean.

given purpose than others. Some parameterizations make it easier to formalize prior domain knowledge than others. To see this, consider the case at hand where we might be specifically interested in the question of how big the difference between the means is. We can therefore also level the model shown in Figure 8.10.

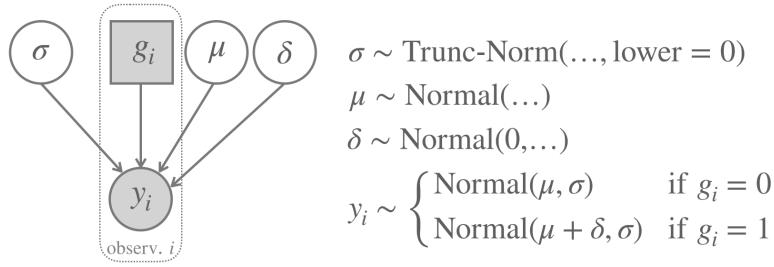


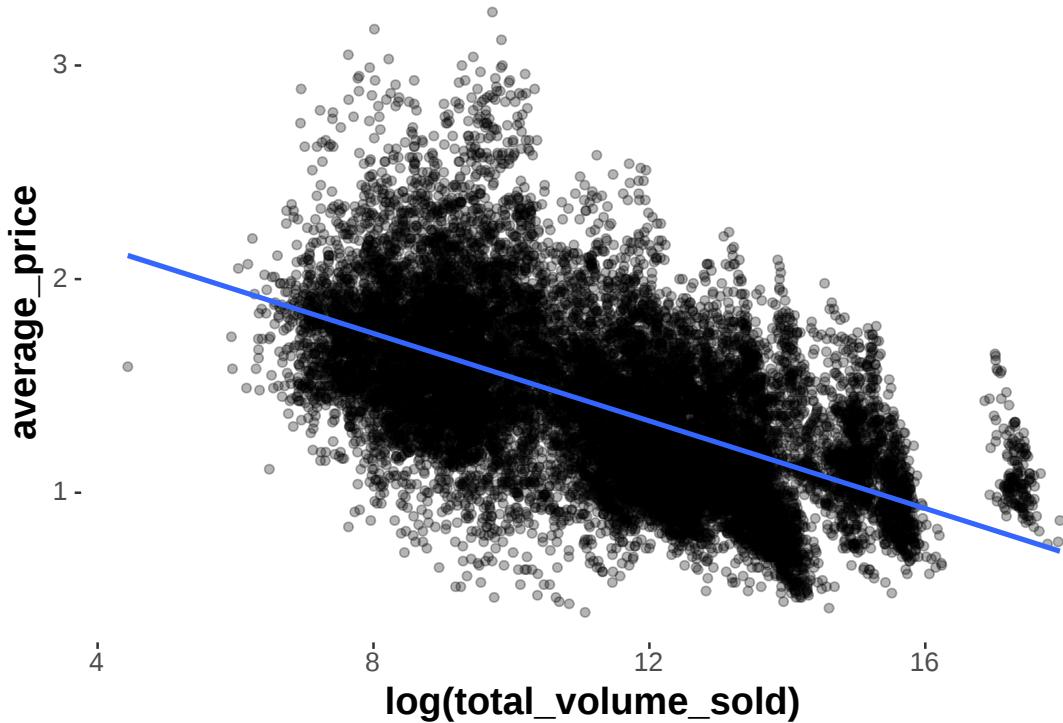
Figure 8.10.: A T-Test Model where one group is the default and the difference between group means is explicitly coded as a parameter.

The variant of the T-Test Model, formulated in terms of differences between groups, in Figure 8.10 is what is usually used in common practice. The advantage is, for example, that it is easy to address the question whether the means are identical, a case which is represented by a single parameter value $\delta = 0$ (rather than an infinite set of pairs $\mu_0 = \mu_1$). It is also easier to specify beliefs about the differences between groups. The prior on δ in Figure 8.10 assumes that we expect *a priori* that $\delta = 0$, but specifying different standard deviations for this prior allows to formalize different degrees of certainty. We could for example use a *skeptical prior*, i.e., an initial model configuration that is skeptical about a group difference, if the standard deviation is set very low.

8.5.5. Simple Linear Regression Model

In Section 6.3 we plotted avocado prices in a scatter plot. In particular we plotted (the log of) average_price as a function of total_amount_sold, and we also added a regression line, like so:

8. Models



A simple linear regression tries to relate pairs of associated observations, frequently denoted as x_i and y_i . Here, i is an index over observations, x_i is the (continuous) independent variable and y_i is the (continuous) dependent variable. In our example, the vector x is the logarithm of `total_amount_sold` and y is the vector `average_price`. The simple linear regression model assumes that there is a simple linear relationship between x and y . A perfect linear relationship would exist if for all i :

$$y_i = \beta_0 + \beta_1 x_i$$

Given measurement error and other lingering uncertainties, we rather handle a **linear relation with stochastic noise**, in which the linear relationship holds, but could be distorted by an ϵ error, which we assume is independently drawn for each i from a normal distribution:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where } \epsilon_i \sim \text{Normal}(\mu = 0, \sigma = \dots)$$

An alternative formulation is to simply write:

$$y_i \sim \text{Normal}(\mu = \beta_0 + \beta_1 x_i, \sigma = \dots)$$

This is the likelihood function that the Simple Linear Regression Model assumes. Figure 8.11 summarizes the full model by also indicating roughly which kinds of prior distributions might be useful in a Bayesian analysis.

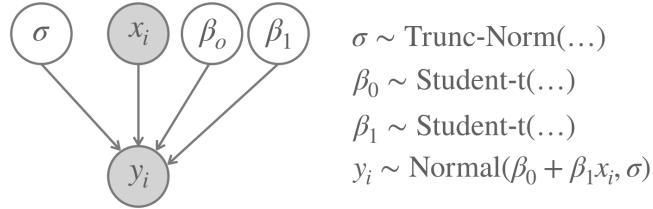
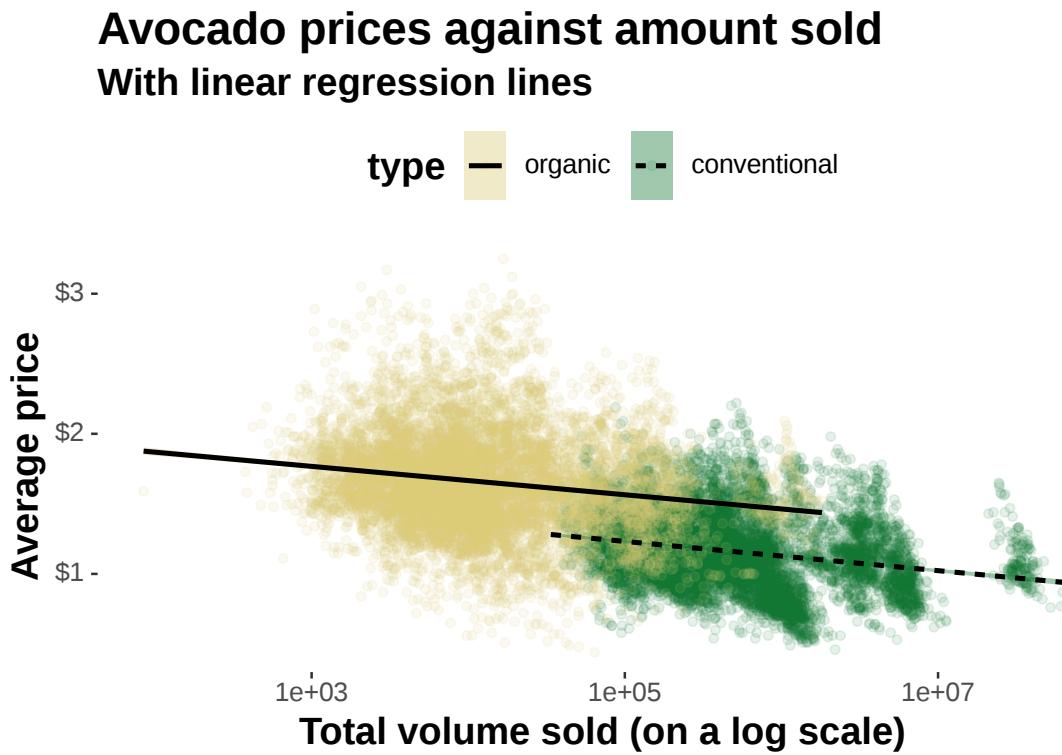


Figure 8.11.: The Simple Linear Regression Model.

8.5.6. Linear Regression with Two Groups

Section 6.3 also plotted avocado prices against the total amount sold and distinguished the type of avocado, also adding a separate regression line for each type. The result looked like this:



Towards a model that computes these linear regression lines, one for each group, that is in line with the idea introduced above that, for the purposes of estimation and testing, we might like to represent differences between groups as δ parameters in a model, let's assume that the type of avocado is our grouping variable g_i . The default group (say: conventional avocados) has $g_i = 0$, the other group has $g_i = 1$. The first group gets "its own" intercept β_0 and slope β_1 . The second group gets additive offsets δ_0 and δ_1 for its slope and intercept, respectively. The full model is shown in Figure 8.12.

8. Models

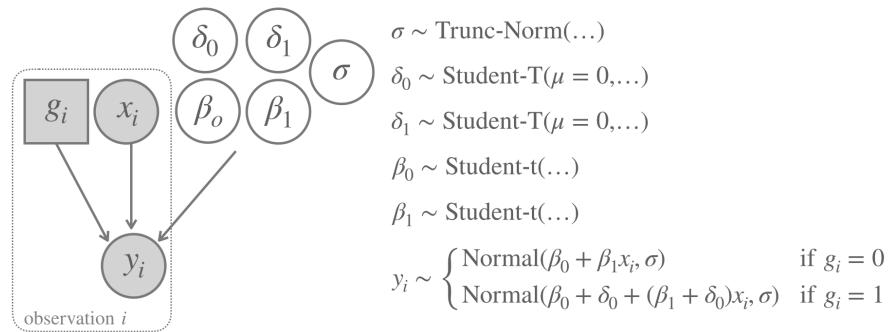


Figure 8.12.: The Linear Regression with Two Groups Model.

8.6. Expressing hypotheses with models

Statistical estimation is used to address research questions of interest. In common practice, these research questions are most often spelled out as specific hypotheses about the values of parameters in a statistical model. Take the case of the mental chronometry data. Figure 8.13 shows the distribution of reaction times for different conditions (= blocks) in the experiment. An obvious research question of interest is whether discrimination takes longer than ‘go/no-go’ responses. We can map this onto a T-Test Model, for example, by focusing on the data from these two conditions only, and asking whether the value of the δ parameter is equal to zero.

The **research hypothesis** in this case would be that there is a difference between the two conditions. But what we consider is its logical negation: the **null hypothesis** is that there is no difference between the two conditions. Using a model-based approach we can cast the null hypothesis as a statement about a single value of a single parameter.

It is possible to formulate hypotheses also about values of parameter tuples. It is also possible to formulate hypothesis as expressions about regions, such as intervals, e.g., asking whether $\delta > 0$. But the majority of applications focuses on point-valued null hypothesis about a single parameter value.

Suppose θ is a single parameter of model M . Our single-parameter, point-valued null hypothesis is that $\theta = x$. There are several approaches to addressing this null hypothesis when we take a model-based approach. These correspond with the three goals of data analysis formulated previously.

Firstly, we can use parameter estimation. We check whether $\theta = x$ squares with the parameter values we infer, given the model M and the data D . This is what we will do in the following Chapter ??.

Secondly, we might want to use a prediction-based approach. In that case, we could ask, for instance, whether a model which assumes that $\theta = x$ would predict the data observed, reversely, whether D would appear rather unlikely given D and the assumption that $\theta = x$. This is the logic of classical hypothesis testing, except that the classical procedures do not routinely make the statistical models explicit. Chapter 10 will deal with this in more detail.

Finally, we can also use model comparison to test hypotheses. In that case, we might compare a model which fixes $\theta = x$ to a model which allows θ to take on other values as well. Chapter 11 covers model

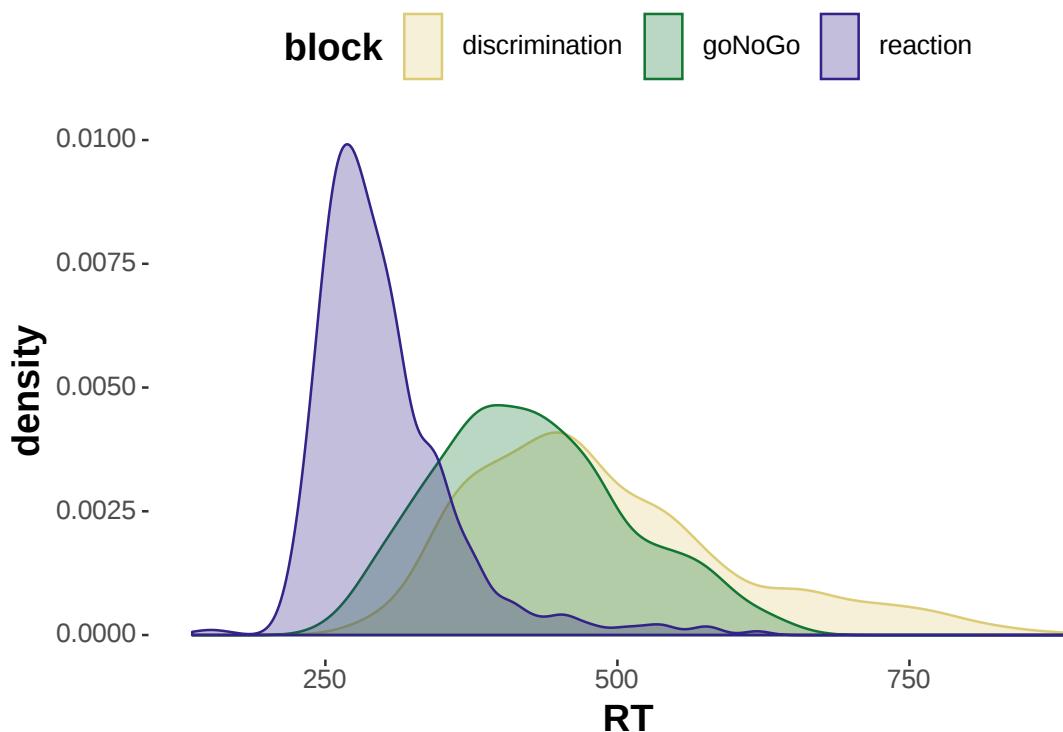


Figure 8.13.: Density plots of the reaction times in the cleaned [mental chronometry data](app-93-data-sets-mental-chronometry) for each type of response measure taken.

8. Models

comparison.

9. Parameter estimation

Based on a model M with parameters θ , parameter estimation addresses the question of which values of θ are good estimates, given some data D . Bayesian and frequentist approaches differ in what they consider a good, or the best estimate. Bayesian approaches take the prior into account, frequentist approaches do not (unsurprisingly).

Parameter estimation is traditionally governed by two measures: (i) a point-estimate for the best parameter value, and (ii) an interval-estimate for a range of values that are considered “good enough”. Table 9.1 gives the most salient answers that each approach gives.

The learning goals for this chapter are:

- understand how Bayes rule applies to parameter estimation
 - role of prior and likelihood
 - understand notion of conjugate prior
- become familiar with and able to compute point-valued estimators
 - frequentist: MLE
 - Bayes: mean of posterior
- become familiar with interval-range estimators
 - frequentist: confidence intervals
 - Bayesian: credible intervals
- be able to implement probabilistic models in `greta` and compute with posterior samples

Table 9.1.: Common methods of obtaining point-valued and interval-range estimates for parameters, given some data, in frequentist and Bayesian approaches.

estimate	Bayesian	frequentist
best value	mean of posterior	maximum likelihood estimate
interval range	credible interval (HDI)	confidence interval

9.1. Bayes rule of parameter estimation

Fix a Bayesian model M with likelihood $P(D \mid \theta)$ for observed data D and prior over parameters $P_M(\theta)$. We then update our prior beliefs $P(\theta)$ to obtained posterior beliefs by Bayes rule:¹

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

The ingredients of this equation are:

- the **posterior distribution** $P(\theta \mid D)$ specifying our beliefs about how likely each value of θ is given fixed k ;
- the **likelihood function** $P(D \mid \theta)$ specifying how likely each observation of k is for a fixed θ (here given by the binomial distribution);
- the **prior distribution** $P(\theta)$ specifying our initial (aka. *a priori*) beliefs about how likely each value of θ might be;
- the **marginal likelihood** $P(D) = \int P(D \mid \theta) P(\theta) d\theta$ specifying how likely an observation of k is under our prior beliefs about θ .

A frequently used shorthand notation for probabilities is this:

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \underbrace{P(D \mid \theta)}_{\text{likelihood}}$$

where the “proportional to” sign \propto indicates that the probabilities on the LHS are defined in terms of the quantity on the RHS after normalization. So, if $F: x \mapsto \mathbb{R}^+$ is a positive function of non-normalized probabilities, $P(x) \propto F(x)$ is equivalent to $P(x) = \frac{F(x)}{\sum_{x'} F(x')}$.

This notation makes particularly clear that the posterior distribution is a “mix” of prior and likelihood. Let’s explore this first, before worrying how to compute posteriors concretely.

9.1.1. The effects of prior and likelihood on the posterior

We consider the case of flipping a coin with unknown bias θ a total of N times and observing k heads (= successes). This is modeled with the **Binomial Model** (see Section 8.5), using priors expressed with a Beta distribution, giving us a model specification as:

¹Since parameter estimation is only about one model, it should do not harm to omit the index M in the probability notation. Moreover, since in many contexts the meaning will be clear enough, we follow common practice and write $P(D \mid \theta)$ as a shortcut for $P(\mathcal{D} = D \mid \Theta = \theta)$. Here \mathcal{D} is the class of all relevant observable data and Θ is the range of a possibly high-dimensional vector of parameter values. We diverge from common practice of using capital roman letters for random variables and small roman letters for values from these random variables, because parameter vectors are traditionally written as θ and the small letter d (albeit non-italic) is reserved for differentials.

$$\begin{aligned}k &\sim \text{Binomial}(N, \theta) \\ \theta &\sim \text{Beta}(\alpha, \beta)\end{aligned}$$

To study the impact of the likelihood function, we compare two data sets. The first one is the contrived “24/7” example where $N = 24$ and $k = 7$. The second example uses a much larger naturalistic data set stemming from the King of France example, namely $k = 109$ for $N = 311$. These numbers are the count of “true” responses for all conditions except for Condition 1, which did not involve a presupposition.

```
data_KoF_cleaned <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-data-science/master/data/KoF_cleaned.csv')) %>%  
  filter(condition != "Condition 1") %>%  
  group_by(response) %>%  
  dplyr::count()  
  
## # A tibble: 2 x 2  
## # Groups:   response [2]  
##   response     n  
##   <lgcl>     <int>  
## 1 FALSE       202  
## 2 TRUE        109
```

We can plot the likelihood function for both data sets like so:

Picking up the example from Section 8.2.2, we will consider the four types of priors shown below:

Combining the four different priors and the two different data sets, we see that the posterior is indeed a mix of prior and likelihood. In particular we see that the strong informative prior has only little effect if there are many data points (the KoF data).

9.1.2. Posterior means and credible intervals

Let's consider the “24/7” example with a flat prior again, concisely repeated in Figure 9.4.

The posterior probability distribution in Figure 9.4 contains rich information. It specifies how likely each value of θ is, obtained by updating the original prior beliefs with the observed data. Such rich information is difficult to process and communicate in language. It is therefore convenient to have conventional means of summarizing the rich information carried in a probability distribution like in Figure 9.4. Customarily, we summarize in terms of a point-estimate and/or an interval estimate. The *point estimate* gives information about a “best value”, i.e., a salient point, such as the expectation (in Bayesian approaches) or the most likely value (in frequentist approaches (see below)). The *interval estimate* gives, usually, an indication of how closely other *good values* are scattered around the *best value*.

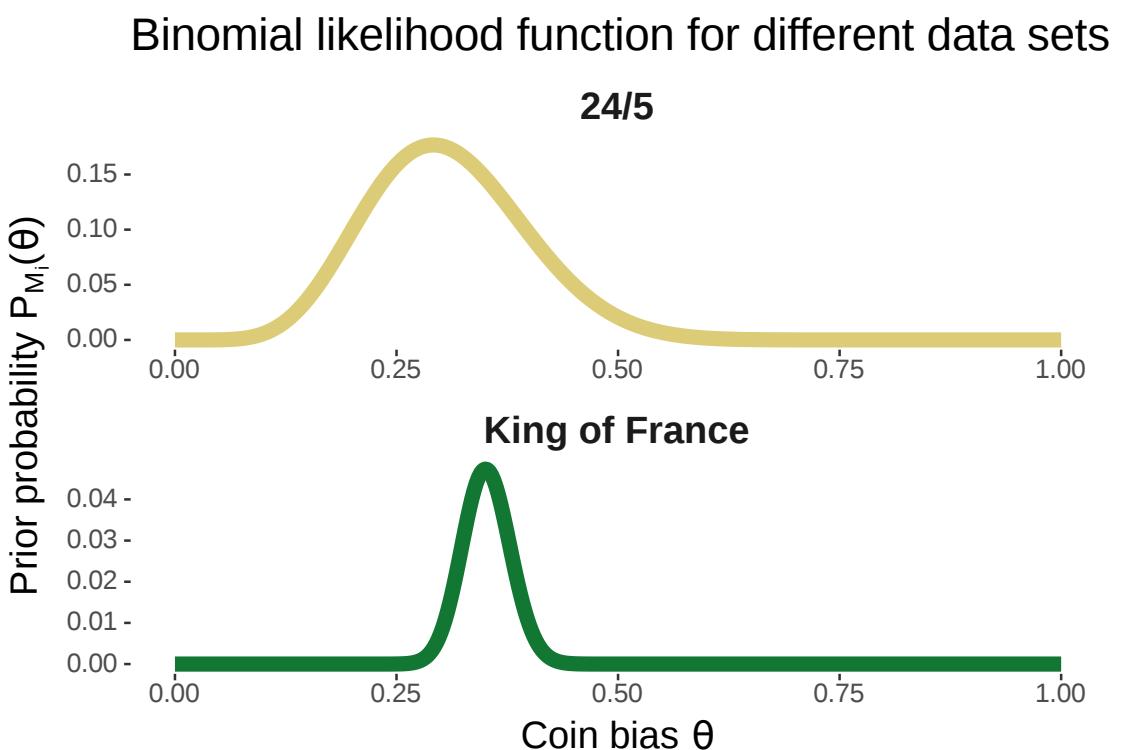


Figure 9.1.: Likelihood for the two examples of bimoial data.

Different kinds of priors over bias θ (Binomial Model)

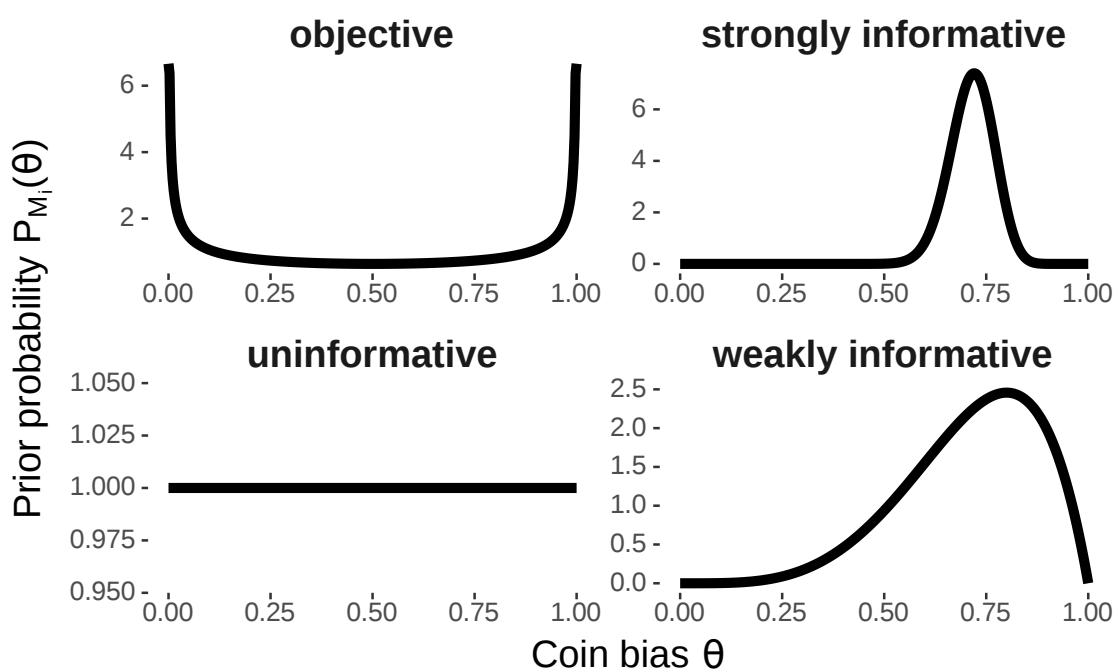


Figure 9.2.: Examples of different kinds of Bayesian priors for the Binomial Model.

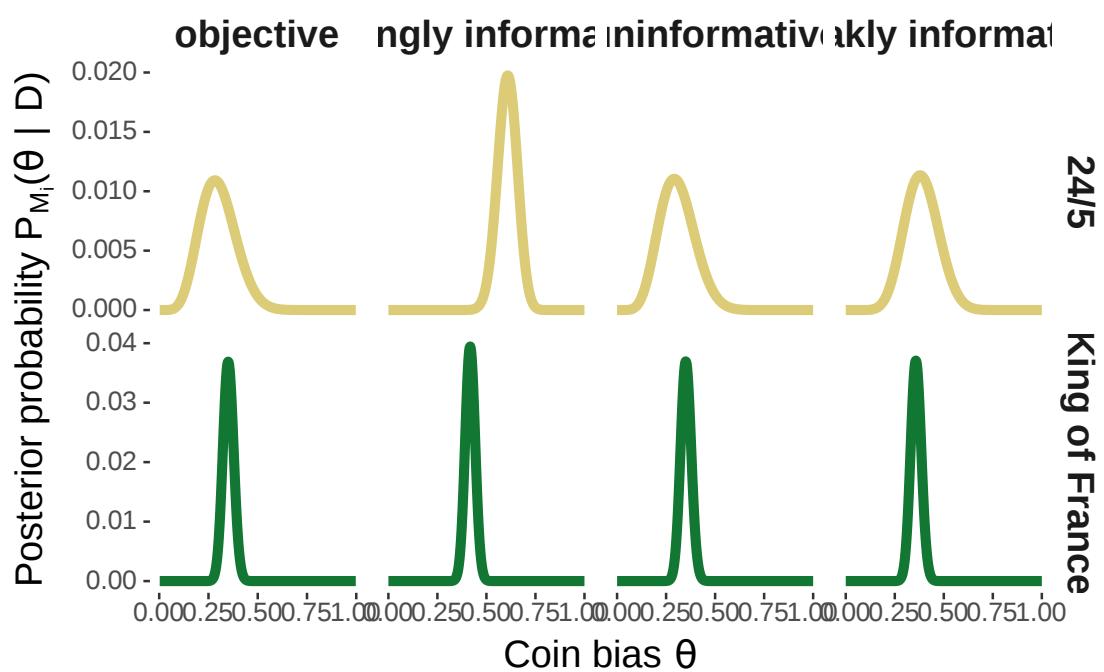
Posterior beliefs in θ for different priors and data

Figure 9.3.: Posterior beliefs over coin biases under different priors and different data sets.

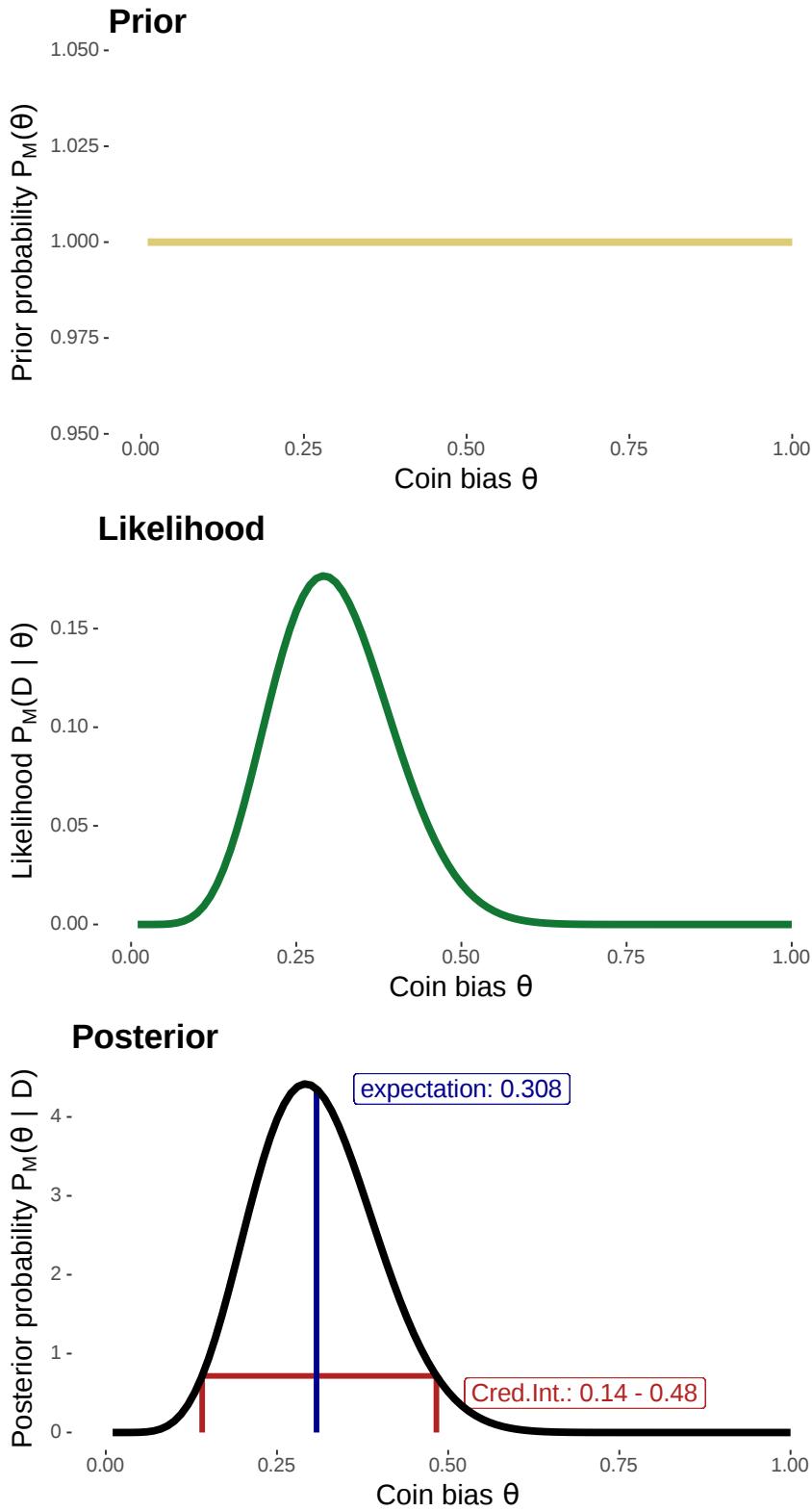


Figure 9.4.: Prior (uninformative), likelihood and posterior for the 24/7 example.

9. Parameter estimation

A common Bayesian point estimate of coin bias parameter θ is **the mean of the posterior distribution**. It gives the value of θ which we would expect to see, when basing out expectations on the posterior distribution:

$$\mathbb{E}_{P(\theta|D)} = \int \theta P(\theta | D) d\theta$$

If we start with flat beliefs, the expected value of θ after k successes in N flips can be calculated rather easily as $\frac{k+1}{n+2}$.² For our example case, we calculate the expected value of θ as $\frac{8}{26} \approx 0.308$ (see also Figure 9.4).

Remark: Maximum a posteriori Another salient point-estimate to summarize a Bayesian posterior distribution is the MAP (maximum *a posteriori*). The map is the parameter value (tuple) that maximizes the posterior distribution:

$$\text{MAP}(P(\theta | D)) = \arg \max_{\theta} P(\theta | D)$$

While the mean of the posterior is “holistic” in the sense that it depends on the whole distribution, the MAP does not. The mean is therefore more faithful to the Bayesian ideal of taking the full posterior distribution into account. Moreover, depending on how Bayesian posteriors are computed / approximated, the estimation of a mean can be more reliable than that of a MAP.

A common Bayesian interval estimate of coin bias parameter θ is a **credible interval**.³ An interval $[l; u]$ is a $\gamma\%$ credible interval for a random variable X if two conditions hold, namely

$$P(l \leq X \leq u) = \frac{\gamma}{100}$$

and, secondly, for every $x \in [l; u]$ and $x' \notin [l; u]$ we have $P(X = x) > P(X = x')$. Intuitively, a 95% credible interval gives the range of values in which we believe with relatively high certainty that the true value resides. Figure 9.4 indicates the 95% credible interval, based on the posterior distribution $P(\theta | D)$ of θ , for the 24/7 example.⁴

9.1.3. Computing Bayesian posteriors with conjugate priors

Bayesian posterior distributions can be hard to compute. Usually, the prior $P(\theta)$ is easy to compute (otherwise we might choose a different one for practicality). Usually, the likelihood function $P(D | \theta)$ is also fast to compute. Everything seems innocuous when we just write:

²This is also known as *Laplace's rule*, or the *rule of succession*.

³Also frequently called “highest-density intervals”, even when we are dealing not with density but probability mass.

⁴Not all random variables have a credible interval for a given γ , according to this definition. A bimodal distribution might not, for example. A bi-modal distribution has two regions of high probability. We can therefore generalize the concept to a finite set of disjoint convex *credible regions*, all of which have the second property of the definition above and all of which conjointly are realized with $\gamma\%$ probability. Unfortunately, common parlance uses the term “credible interval” to refer to credible regions as well. The same disaster occurs with alternative terms, such as “ $\gamma\%$ highest-density intervals”, which also often refers to what should better be called “highest-density regions”.

$$\underbrace{P(\theta | D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \underbrace{P(D | \theta)}_{\text{likelihood}}$$

But the real pain is the normalizing constant, i.e., the marginalize likelihood a.k.a. the “integral of doom”, which to compute can be intractable, especially if the parameter space is large and not well-behaved:

$$P(D) = \int P(D | \theta) P(\theta) d\theta$$

Section 9.5 will therefore enlarge on methods to compute or approximate the posterior distribution efficiently.

Foortunately, computing Bayesian posterior distributions need not always be intractable. If the prior and the likelihood function cooperate, so to speak, the computation of the posterior can be as simple as sleep. The nature of the data often prescribes which likelihood function is plausible. But we have more wiggle room in the choice of the priors. If prior $P(\theta)$ and posterior $P(\theta | D)$ are of the same family, i.e., if they are the same kind of distribution albeit possibly with different parameterizations, we say that they **conjugate**. In that case, the prior $P(\theta)$ is called **conjugate prior** for the likelihood function $P(D | \theta)$. from which the posterior $P(\theta | D)$ is derived.

Theorem 9.1. *The Beta distribution is the conjugate prior of binomial likelihood. For $\theta \sim \text{Beta}(a, b)$ as prior and data k and N , the posterior is $\theta \sim \text{Beta}(a + k, b + N - k)$.*

Proof. By construction, the posterior is:

$$P(\theta | \langle k, n \rangle) \propto \text{Binomial}(k; n, \theta) \text{Beta}(\theta | a, b)$$

We extend the RHS by definitions:

$$\begin{aligned} \text{Binomial}(k; n, \theta) \text{Beta}(\theta | a, b) &= \theta^k (1 - \theta)^{n-k} \theta^{a-1} (1 - \theta)^{b-1} \\ &= \theta^{k+a-1} (1 - \theta)^{n-k+b-1} \end{aligned}$$

This latter expression is the non-normalized Beta-distribution for parameters $a + k$ and $b + N - k$, so that we conclude with what was to be shown:

$$P(\theta | \langle k, n \rangle) = \text{Beta}(\theta | a + k, b + N - k)$$

□

9.1.4. Sequential updating

Ancient wisdom has coined the widely popular proverb: “Today’s posterior is tomorrow’s prior.” Suppose we collected data from an experiment, like $k = 7$ in $N = 24$. Using uninformative priors at the outset, our posterior belief after the experiment is $\theta \sim \text{Beta}(8, 18)$. But now consider what happened at half-time. After half the experiment, we had $k = 2$ and $N = 12$, so our beliefs followed $\theta \sim \text{Beta}(3, 10)$ at this moment in time. But using these beliefs as priors, and observing the rest of the data would consequently result in updating the prior $\theta \sim \text{Beta}(3, 10)$ with another set of observations $k = 5$ and $N = 12$, giving us the same posterior belief as what we would have gotten if we updated in one swoop. Figure 9.5 shows steps through belief space, starting uninformed, and observing one piece of data at a time (going right for each outcome of heads, down for each outcome of tails).

This sequential updating is not a peculiarity of the Beta-Binomial case or of conjugacy. It holds in general for Bayesian inference. Sequential updating is a very intuitive property, but it is not shared by all other forms of inference from data. That Bayesian inference is sequential and commutative follows from commutativity of multiplication of likelihoods (and the definition of Bayes rule).

Theorem 9.2. *Bayesian posterior inference is sequential and commutative in the sense that for a data set D which is comprised of two mutually exclusive subsets D_1 and D_2 such that $D_1 \cup D_2 = D$, we have:*

$$P(\theta | D) \propto P(\theta | D_1) P(D_2 | \theta)$$

Proof.

$$\begin{aligned} P(\theta | D) &= \frac{P(\theta) P(D | \theta)}{\int P(\theta') P(D | \theta') d\theta'} \\ &= \frac{P(\theta) P(D_1 | \theta) P(D_2 | \theta)}{\int P(\theta') P(D_1 | \theta') P(D_2 | \theta') d\theta'} && [\text{from multiplicativity of likelihood}] \\ &= \frac{P(\theta) P(D_1 | \theta) P(D_2 | \theta)}{\frac{k}{k} \int P(\theta') P(D_1 | \theta') P(D_2 | \theta') d\theta'} && [\text{for random positive } k] \\ &= \frac{\frac{P(\theta) P(D_1 | \theta)}{k} P(D_2 | \theta)}{\int \frac{P(\theta') P(D_1 | \theta')}{k} P(D_2 | \theta') d\theta'} && [\text{rules of integration; basic calculus}] \\ &= \frac{P(\theta | D_1) P(D_2 | \theta)}{\int P(\theta' | D_1) P(D_2 | \theta') d\theta'} && [\text{Bayes rule with } k = \int P(\theta) P(D_1 | \theta) d\theta] \end{aligned}$$

□

9.2. A frequentist approach to parameter estimation

We pick up the “24/7” example again. As before, the goal is to draw inferences about the latent bias $\theta \in [0; 1]$. Being a frequentist, we do that based only on likelihood function $P(k | \theta)$, given here by the binomial

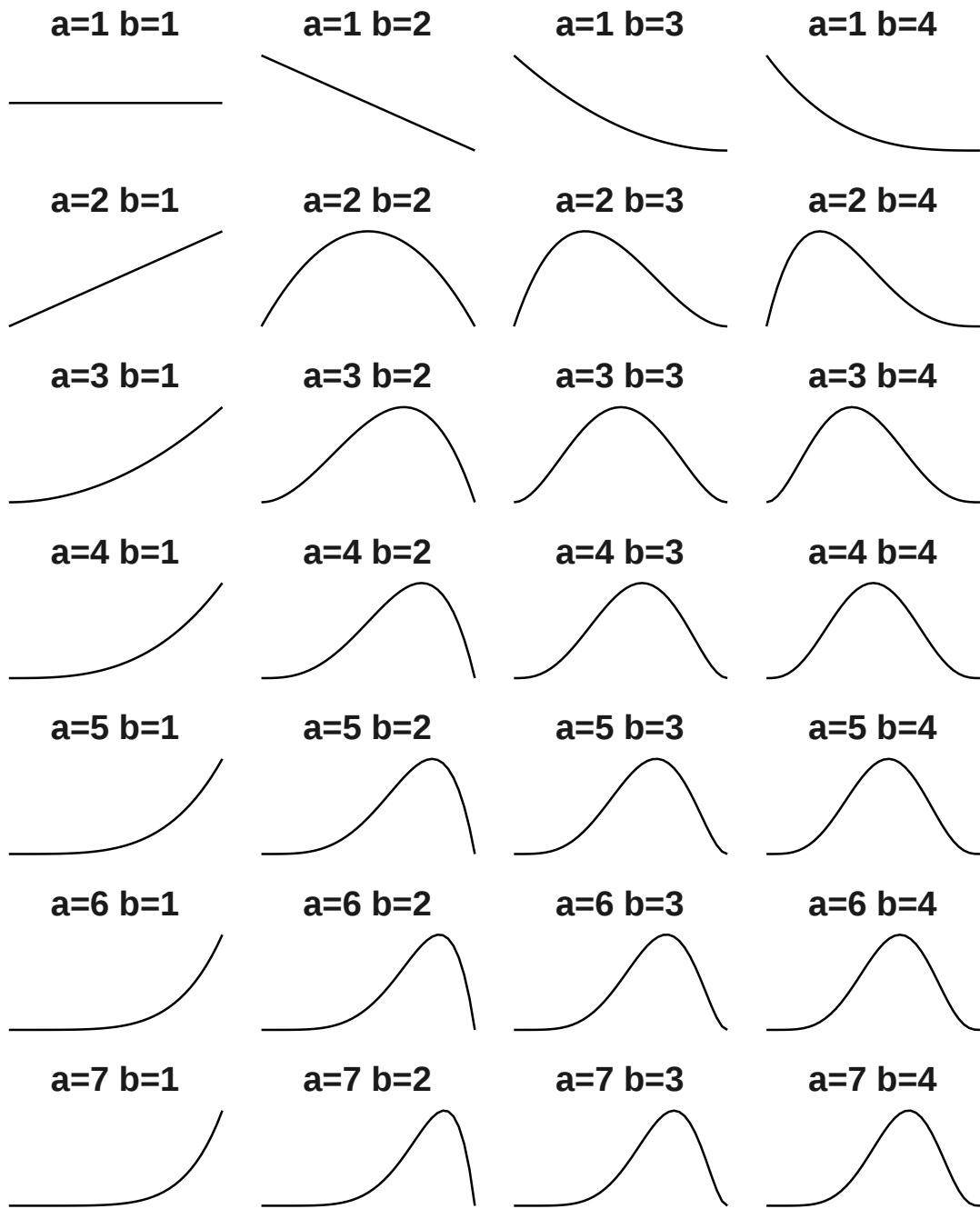


Figure 9.5.: Beta distributions for different parameters. Starting from an uninformative prior (top left), we arrive at the posterior distribution in the bottom left, in any sequence of sequentially updating with the data.

9. Parameter estimation

distribution as before. We will derive, from $P(k \mid \theta)$ alone, plausible point and interval estimates. There are several constructions for both point and interval estimates. We here only look at what seem to be the single most prominent exemplars in each category.

9.2.1. Maximum likelihood estimate

The **maximum likelihood estimate (MLE)** is a point estimate based on the likelihood function alone. It specifies the value of θ for which the observed data is most likely. We often use the notation $\hat{\theta}$ to denote the MLE of θ :

$$\hat{\theta} = \arg \max_{\theta} P(d \mid \theta)$$

For the binomial likelihood function, the maximum likelihood estimate is easy to calculate as $\frac{k}{N}$, yielding $\frac{7}{24} \approx 0.292$ for the running example. Figure 9.6 shows a graph of the non-normalized likelihood function and indicates the maximum likelihood estimate (the value that maximizes the curve).

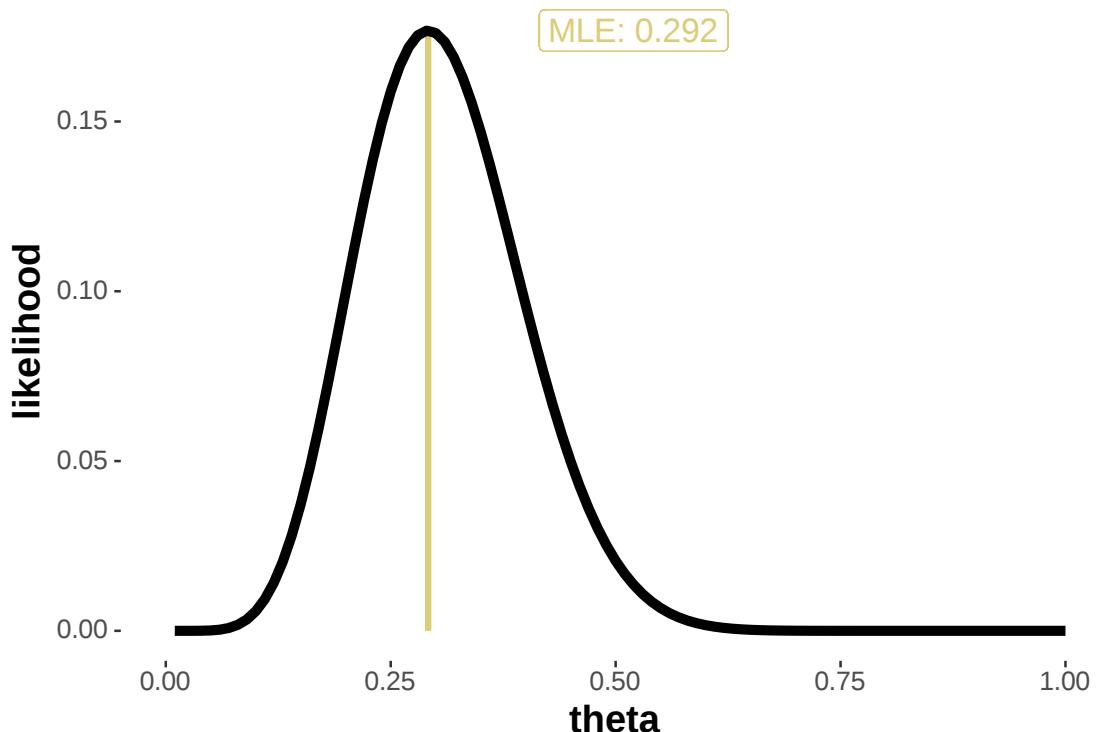


Figure 9.6.: Non-normalized likelihood function for the observation of $k = 7$ successes in $N = 24$ flips, including maximum likelihood estimate.

9.2.2. Confidence intervals

The most commonly used interval estimate in frequentist analyses are *confidence intervals*. Although (frequentist) confidence intervals can coincide with (subjectivist) credible intervals in specific cases, they generally do not. And even when confidence and credible values yield the same numerical results, these notions are fundamentally different and ought not to be confused.

Let's look at credible intervals to establish the proper contrast. Recall that part of the definition of a credible interval for a posterior distribution over θ , captured here notationally in terms a random variable Θ , was the probability $P(l \leq \Theta \leq u)$ that the value realized by random variable Θ lies in the interval $[l; u]$. This statement makes no sense to the frequentist. There cannot be any non-trivial value for $P(l \leq \Theta \leq u)$. The true value of θ is either in the interval $[l; u]$ or it is not. To speak of a probability that θ is in $[l; u]$ is to appeal to an ill-formed concept of probability which the frequentist denies.

In order to give an interval estimate nonetheless, the frequentist appeals to probabilities that she can accept: probabilities derived from (hypothetical) repetitions of a genuine random event with objectively observable outcomes. Let \mathcal{D} be the random variable that captures the probability with which data $\mathcal{D} = D$ is realized. We obtain a pair of derived random variables X_l and X_u from a pair of functions $g_{l,u}: d \mapsto \mathbb{R}$. A $\gamma\%$ **confidence interval** for observed data D_{obs} is the interval $[g_l(D_{\text{obs}}), g_u(D_{\text{obs}})]$ whenever functions $g_{l,u}$ are constructed in such a way that

$$P(X_l \leq \theta_{\text{true}} \leq X_u) = \frac{\gamma}{100}$$

where θ_{true} is the unknown but fixed true value of θ . In more intuitive words, a confidence interval is the outcome of a special construction (functions $g_{l,u}$) such that, when applying this procedure repeatedly to outcomes of the assumed data-generating process, the true value of parameter θ will lie inside of the computed confidence interval in exactly $\gamma\%$ of the cases.

It is easier to think of the definition of a confidence interval in terms of computer code and sampling (see Figure 9.7). Suppose Grandma gives you computer code, a `magic_function` which takes as input data observations, and returns an interval estimate for the parameter of interest. We sample a value for the parameter of interest repeatedly, and consider it the “true parameter” for the time being. For each sampled “true parameter”, we generate data repeatedly. We apply Grandma’s `magic_function`, obtain an interval estimate and check if the true value that triggered the whole process is included in the interval. Grandma’s `magic_function` is a $\gamma\%$ confidence interval if the proportion of inclusions (the checkmarks in Figure 9.7) is $\gamma\%$.

In some complex cases, the frequentist analyst relies on functions g_l and g_u that are easy to compute but only approximately satisfy the condition $P(X_l \leq \theta_{\text{true}} \leq X_u) = \frac{\gamma}{100}$. For example, we might use an asymptotically correct calculation, based on the observation that, if n grows to infinity, the binomial distribution approximates a normal distribution. We can then calculate a confidence interval, as if our binomial distribution actually was a normal distribution. If n is not large enough, this will be increasingly imprecise. Rules of thumb are used to decide how big n has to be to involve at best a tolerable amount of imprecision (see the Info Box below).

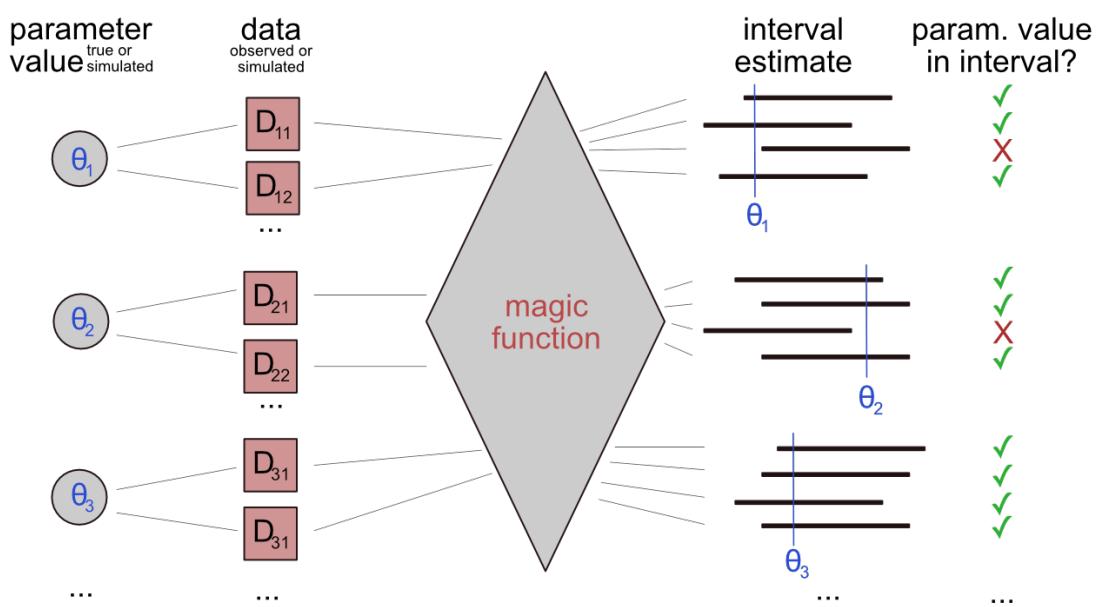


Figure 9.7.: Schematic representation of what a confidence interval does: think of it as a magic function which returns intervals that contain the true value in γ percent of the cases.

For our running example ($k = 7, n = 24$), the rule of thumb mentioned in the Info Box below recommends *not* using the asymptotic calculation. If we did nonetheless, we would get a confidence interval of $[0.110; 0.474]$. For the binomial distribution also a more reliable calculation exist, which yields $[0.126; 0.511]$ for the running example. (We can use numeric simulation to explore how good/bad a particular approximate calculation is, as shown in the next section.) The more reliable construction, the so-called *exact method*, implemented in the function `binom.confint` of R package `binom`, revolves around the close relationship between confidence intervals and p -values. (To foreshadow a later discussion: the exact $\gamma\%$ confidence interval is the set of all parameter values for which an exact (binomial) test does not yield a significant test result as the level of $\alpha = 1 - \frac{\gamma}{100}$.)

Aymptotic approximation of a binomial confidence interval using a normal distribution.

Let X be the random variable that determines the binomial distribution, i.e., the probability of seeing k successes in n flips. For large n , X approximates a normal distribution with a mean $\mu = n\theta$ and a standard deviation of $\sigma = \sqrt{n\theta(1-\theta)}$. The random variable U :

$$U = \frac{X - \mu}{\sigma} = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

Let \hat{P} be the random variable that captures the distribution of our maximum likelihood estimates for an observed outcome k :

$$\hat{P} = \frac{X}{n}$$

Since $X = \hat{P}n$ we obtain:

$$U = \frac{\hat{P}n - n\theta}{\sqrt{n\theta(1-\theta)}}$$

We now look at the probability that U is realized to lie in a symmetric interval $[-c, c]$, centered around zero – a probability which we require to match our confidence level:

$$P(-c \leq U \leq c) = \frac{\gamma}{100}$$

We now expand the definition of U in terms of \hat{P} , equate \hat{P} with the current best estimate $\hat{p} = \frac{k}{n}$ based on the observed k and rearrange terms, yielding the asymptotic approximation of a binomial confidence interval:

$$\left[\hat{p} - \frac{c}{n} \sqrt{n\hat{p}(1-\hat{p})}; \hat{p} + \frac{c}{n} \sqrt{n\hat{p}(1-\hat{p})} \right]$$

This approximation is conventionally considered precise enough when the following *rule of thumb* is met:

$$n\hat{p}(1-\hat{p}) > 9$$

9.3. Addressing point-valued hypotheses with parameter estimation

Using interval-based estimates, we can address research questions formulated as point-valued hypotheses about a parameter of interest. Let Θ be the parameter space of a model M . We are interested in some component Θ_i and our hypothesis is $\Theta_i = \theta_i^*$ for some specific value θ_i^* . A simple (but crude and controversial) way of addressing this point-valued hypothesis based on observed data D is to look at whether θ_i^* lies inside the interval-estimate for parameter Θ_i based on observed data D . This would apply for frequentist confidence intervals and Bayesian credible intervals alike.

Kruschke (2015) extends this approach to addressing point-valued hypothesis. He argues that we should *not* be concerned with point-valued hypotheses, but rather with intervals constructed around the point-value of interest. Kruschke therefore suggests to look at a **region of practical equivalence** (ROPE), usually defined by some ϵ -region around θ_i^* :

$$\text{ROPE}(\theta_i^*) = [\theta_i^* - \epsilon, \theta_i^* + \epsilon]$$

The choice of ϵ is context-dependent and requires an understanding of the scale at which parameter values Θ_i differ. If the parameter of interest is, for example, the difference δ in the means of reaction times, like in the Mental Chronometry example, this parameter is intuitively interpretable. We can say, for instance, that an ϵ -region of $\pm 5\text{ms}$ is really so short that any value in $[-5\text{ms}; 5\text{ms}]$ would be regarded as identical to 0 for all practical purposes, because of what we know about reaction times and their potential differences. However, with parameters that are less clearly anchored to a concrete physical measurement about which we have solid distributional knowledge and/or reliable intuitions, fixing the size of the ROPE can be more difficult. For the bias of a coin flip, for instance, which we want to test at the point value $\theta^* = 0.5$ (testing the coin for fairness), we might want to consider a ROPE like $[0.49; 0.51]$, although this choice may be less objectively defensible without previous experimental evidence from similar situations.

Kruschke (2015) advances the ROPE approach for Bayesian hypothesis testing, based on parameter estimation and credible intervals. The rationale for using a ROPE could, in principle, also be extended to frequentist approaches, using confidence intervals instead of credible intervals. It would, however, undermine the frequentist rationale for confidence intervals to use a ROPE, because we would lose the tight hand on error control, which is built into frequentist testing, and also confidence intervals. (More on this in Chapter 10). That is why the frequentist approach does not use ROPEs, or equivalently only considers $\epsilon = 0$.

In Kruschke's ROPE-based approach where $\epsilon \geq 0$, the decision about a point-valued hypothesis becomes ternary. If $[l; u]$ is an interval-based estimate of parameter Θ_i and $[\theta_i^* - \epsilon; \theta_i^* + \epsilon]$ is the ROPE around the point-value of interest, then we would:

- **accept** the point-valued hypothesis iff $[l; u]$ is contained entirely in $[\theta_i^* - \epsilon; \theta_i^* + \epsilon]$;
- **reject** the point-valued hypothesis iff $[l; u]$ and $[\theta_i^* - \epsilon; \theta_i^* + \epsilon]$ have no overlap; and
- **withhold judgement** otherwise.

Consider the 24/7 example, where the point-valued hypothesis of interest is $\theta^* = 0.5$ (testing the coin for fairness) and the ROPE is $[0.49; 0.51]$ ($\epsilon = 0.1$, arbitrarily set here). The point- and interval-estimates for Bayesian and frequentist approaches are as follows:

```

estimates_24_7 <- tibble(
  `lower_Bayes` = HDInterval::hdi(function(x) qbeta(x, 8, 18))[1],
  `point_Bayes` = 8/25,
  `upper_Bayes` = HDInterval::hdi(function(x) qbeta(x, 8, 18))[2],
  `lower_frequentist` = binom::binom.exact(7, 24)$lower,
  `point_frequentist` = 7/24,
  `upper_frequentist` = binom::binom.exact(7, 24)$upper
) %>%
  pivot_longer(
    everything(),
    names_pattern = "(.*)(.*)",
    names_to = c(".value", "approach")
  )
estimates_24_7

## # A tibble: 2 x 4
##   approach   lower  point  upper
##   <chr>     <dbl> <dbl> <dbl>
## 1 Bayes      0.141  0.32  0.483
## 2 frequentist 0.126  0.292 0.511

```

Figure 9.8 shows these estimates next to the ROPE. We see that the Bayesian 95% credible interval has no overlap with the ROPE, so that we would *reject* the null-hypothesis of $\theta^* = 0.5$ by the ROPE+estimation logic of statistical decision making. If we were to illegitimately (!) apply this approach to frequentist confidence intervals, we would *accept* this point-valued hypothesis instead. However, as stressed above, this is *not* an accepted move in frequentist statistics. The frequentists would *not* consider a ROPE with $\epsilon > 0$. The frequentist would also not accept the point-valued hypothesis in case the critical value is included in the confidence interval. They would merely withhold judgement, i.e., *not reject* the point hypothesis. (More on this in Chapter 10.)

9.4. Comparing Bayesian and frequentist estimates

For Bayesians point-valued and interval-based estimates are just summary statistics to efficiently communicate about or reason with the main thing: the full posterior distribution. For the frequentist, the point-valued and interval-based estimates might be all there is. Computing a full posterior can be very hard. Computing point-estimates is usually much simpler. Yet, all the trouble of having to specify priors, and having to calculate a much more complex mathematical object, can pay off. An example which is intuitive enough is that of a likelihood function in a multi-dimensional parameter space where there is an infinite collection of parameter values that maximize the likelihood function (think of plateau). Asking a godly oracle for the (actually: an) MLE can be disastrously misleading. The full posterior will show the quirkiness.⁵

⁵We will see such an example later for the case of linear regression with correlated independent variables, so-called collinearity.

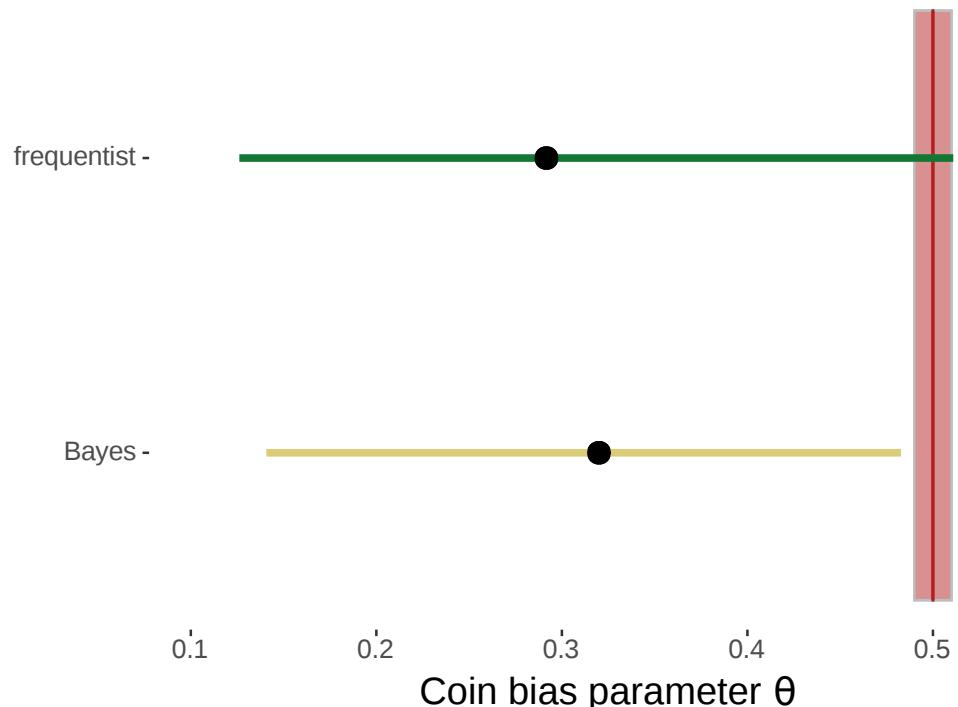


Figure 9.8.: Comparing point- and interval-estimates (Bayesian credible intervals and frequentist confidence intervals) against a ROPE of [0.49; 0.51] (red shaded area) around the point-valued hypothesis of interest is $\theta^* = 0.5$.

Practical issues aside, there are also conceptual arguments that can be pinned against each other. Suppose you do not know the bias of a coin, you flip it once and it lands heads. The case in mathematical notation: $k = 1, N = 1$. As a frequentist, your “best” estimate of the coin’s bias is that it is 100% rigged: it will never land tails. As a Bayesian, with uninformed priors, your “best” estimate is, following Laplace rule, $\frac{k+1}{N+2} = \frac{2}{3}$. Notice that, apparently, there might be different notions of what counts as “best” in place. Still, the frequentist “best” estimate seems rather extreme.

What about interval-ranged estimates? Which is the better tool, confidence intervals or credible intervals?

- This is harder to answer. Numerical simulations can help answer these questions.⁶ The idea is simple but immensely powerful. We simulate, repeatedly, a ground-truth and synthetic results for fictitious experiments, and then we apply the statistical tests/procedures to these fictitious data sets. Since we know the ground-truth, we can check which tests/procedures got it right.

Let’s look at a simulation, comparing credible intervals to confidence intervals, the latter of which calculated by asymptotic approximation or the so-called exact method. To do so, we repeatedly sample a ground-truth (e.g., a known coin bias θ_{true}) from a flat distribution over $[0; 1]$.⁷ We then simulate an experiment in a synthetic world with θ_{true} , using a fixed value of n , here taken from the set $n \in \{10, 25, 100, 1000\}$. We then construct a confidence interval (either approximately or precisely) and a 95% credible interval; for each of the three interval estimates. We check whether the ground-truth θ_{true} is *not* included in any given interval estimate. We calculate the mean number of times such non-inclusion (errors!) happen for each kind of interval estimate. The code below implements this and the figure below shows the results, based on 10,000 samples of θ_{true} .

```
# how many "true" thetas to sample
n_samples <- 10000
# sample a "true" theta
theta_true <- runif(n=n_samples)
# create data frame to store results in
results <- expand_grid(
  theta_true = theta_true,
  n_flips = c(10, 25, 100, 1000)
) %>%
  as_tibble() %>%
  mutate(
    outcome = 0,
    norm_approx = 0,
    exact = 0,
    Bayes_HDI = 0
  )
for (i in 1:nrow(results)) {
```

⁶Even if the math seems daunting, this method is much more tangible and applicable and requires only basic programming experience.

⁷This is already not innocuous. We are fixing, as it were, an assumption about how likely ground-truths should actually occur in the real world.

9. Parameter estimation

```
# sample fictitious experimental outcome for current true theta
results$outcome[i] <- rbinom(
  n = 1,
  size = results$n_flips[i],
  prob = results$theta_true[i]
)

# get CI based on asymptotic Gaussian
norm_approx_CI <- binom::binom.confint(
  results$outcome[i],
  results$n_flips[i],
  method = "asymptotic"
)
results$norm_approx[i] <- !( 
  norm_approx_CI$lower <= results$theta_true[i] &&
  norm_approx_CI$upper >= results$theta_true[i]
)

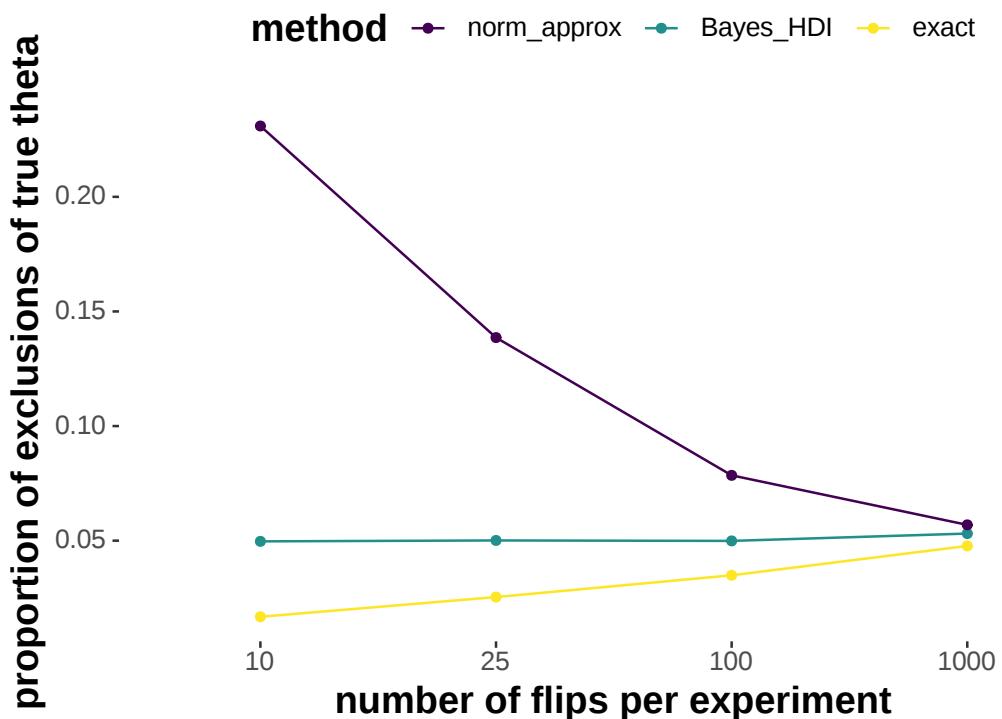
# get CI based on exact method
exact_CI <- binom::binom.confint(
  results$outcome[i],
  results$n_flips[i],
  method = "exact"
)
results$exact[i] <- !( 
  exact_CI$lower <= results$theta_true[i] &&
  exact_CI$upper >= results$theta_true[i]
)

# get 95% HDI (flat priors)
Bayes_HDI <- binom::binom.bayes(
  results$outcome[i],
  results$n_flips[i],
  type = "highest",
  prior.shape1 = 1,
  prior.shape2 = 1
)
results$Bayes_HDI[i] <- !( 
  Bayes_HDI$lower <= results$theta_true[i] &&
  Bayes_HDI$upper >= results$theta_true[i]
)
}
```

```

results %>%
  gather(key = "method", "Type_1", norm_approx, exact, Bayes_HDI) %>%
  group_by(method, n_flips) %>%
  dplyr::summarize(avg_type_1 = mean(Type_1)) %>%
  ungroup() %>%
  mutate(
    method = factor(
      method,
      ordered = T,
      levels = c("norm_approx", "Bayes_HDI", "exact")
    )
  ) %>%
  ggplot(aes(x = as.factor(n_flips), y = avg_type_1, color = method)) +
  geom_point() + geom_line(aes(group = method)) +
  xlab("number of flips per experiment") +
  ylab("proportion of exclusions of true theta")

```



9.5. Algorithms for parameter estimation

Statistical methods are also dictated by practicality. If there is no computer at hand, a statistical model or procedure that is difficult or impossible to compute by hand is not useful. We must understand classic notions and procedures in statistics also from this perspective of computability. In reverse, we must understand recent changes in statistical practice likewise as a natural reaction to advances in computability.

9.5.1. Optimizing functions

Computing the maximum or minimum of a function, such as an MLE or MAP estimate, is a common problem. R has a built-in function `optim` that is useful for finding the the minimum of a function. (If a maximum is needed, just multiply by -1 and search the minimum with `optim`.)

We can use the `optim` function to retrieve an MLE for the three free parameters of a simple linear regression of `average_price` based on `total_volume_sold` in the `avocado` data, like so:

```
# function for the negative log-likelihood of the given
# data and fixed parameter values
nll = function(y, x, beta_0, beta_1, sd) {
  # negative sigma is logically impossible
  if (sd <= 0) {return( Inf )}
  # predicted values
  yPred = beta_0 + x * beta_1
  # negative log-likelihood of each data point
  nll = -dnorm(y, mean=yPred, sd=sd, log = T)
  # sum over all observations
  sum(nll)
}
fit_lh = optim(
  # initial parameter values
  par = c(1.5, 0, 0.5),
  # function to optimize
  fn = function(par) {
    with(avocado_data,
      nll(average_price, total_volume_sold,
          par[1], par[2], par[3]))
  }
)
fit_lh$par

## [1] 1.425080e+00 -2.247373e-08 3.950978e-01
```

This result tells us that the best fitting parameter triple has an intercept of $\beta_0 \approx 1.42$, a slope $\beta_1 \approx -2.47$ and a standard deviation $\sigma \approx 0.39$. We can compare these values with a built-in function for linear regression models (which, however, does not return an estimate of σ (see Chapter 13 for more information)):

```
lm(average_price ~ total_volume_sold, avocado_data)$coef

##           (Intercept) total_volume_sold
##     1.425096e+00      -2.247455e-08
```

9.5.2. Approximating posterior distributions

There are several methods of computing approximations of Bayesian posteriors. **Variational inference**, for example, hinges on the fact that under very general conditions Bayesian posterior distributions are well approximated by (multi-variate) normal distributions. The more data, the better the approximation. We can then reduce approximation of a Bayesian posterior to a problem of optimizing parameter values: we simply look for the parameter values that yield the “best” parametric approximation to the Bayesian posterior. (Here, “best” is usually expressed in terms of minimizing a measure of divergence between probability distributions, such as Kullback-Leibler divergence.) Another prominent method of approximating Bayesian posteriors is rejection sampling.

The most prominent class of methods to approximate Bayesian posteriors are Markov Chain Monte Carlo methods. We will describe the most basic version of such MCMC algorithms below. For the purposes of this class it suffices to accept that there are black boxes (with some knobs for fine-tuning) that, if you supply a model description, priors and data, will return samples from the posterior distribution.

9.5.2.1. Of apples and trees: Markov Chain Monte Carlo sampling

Beginning of each summer, Nature sends out the Children to distribute the apples among the trees. It is custom that bigger trees ought to receive more apples. Indeed, every tree is supposed to receive apples in proportion to how many leaves it has. If Giant George (an apple tree!) has twice as many leaves as Thin Finn (another apple tree!), Giant George is to receive twice as many apples as Thin Finn. This means that if there are n_a apples to distribute in total, and $L(t)$ is the number of leaves of tree t , every tree should receive $A(t)$ apples, where:

$$A(t) = \frac{L(t)}{\sum_{t'} L(t')} n_a$$

The trouble is that not even Nature knows the number of leaves of all the trees. Nature does not care about numbers. The Children, however, can count. But they cannot keep in mind the number of leaves for many trees for a long time. And no single Child could ever visit all the trees before the winter. This is why the Children distribute apples in a way that approximates Nature’s will. The more apples to distribute, the better the approximation. Nature is generally fine with approximate but practical solutions.

9. Parameter estimation

When a Child visits a tree, it affectionately hangs an apple into its branches. It also writes down the name of the tree in a list next to the number of the apple it has just delivered. It then looks around and selects a random tree in the neighborhood. If the current tree t_c where the Child is at present has fewer leaves than this other tree t_o , i.e., if $L(t_c) < L(t_o)$, the Child visits t_o . If instead $L(t_c) \geq L(t_o)$ the child flips a coin and visits t_o with a probability proportional to $\frac{L(t_o)}{L(t_c)}$. In other words, the Child will always visit a tree with more leaves, and it will visit a tree with fewer leaves depending on the proportion of leaves.

When a large number of apples are distributed, and Nature looks at the list of trees each Child has visited, this list of tree names is a set of **representative samples** from the probability distribution:

$$P(t) \propto L(t)$$

These samples were obtained without knowledge of the normalizing constant. The Children only had $L(t)$ at their disposal. When trees are parameter tuples θ and the number of leaves is the product $P(D | \theta) P(\theta)$, the Children would deliver samples from the posterior distribution *without* knowledge of the normalizing constant (a.k.a. the integral of doom).

The sequence of trees visited by a single Child is a **sample chain**. Usually, Nature sends out at least 2-4 Children. The first tree a Child visits is the **initialization of the chain**. Sometimes Nature select initial trees strategically for each Child. Sometimes Nature lets randomness rule. In any case, a Child might be quite far away from the meadow with lush apple trees, the so-called **critical region** (where to dwell makes most sense). It might take many tree hops before a Child reaches this meadow. Nature therefore allows each Child to hop from tree to tree for a certain time, the **warm-up period**, before the Children start distributing apples and taking notes. If each Child only records every k -th tree it visits, Nature calls k a **thinning factor**. Thinning generally reduces **autocorrelation** (think: the amount to which subsequent samples do not carry independent information about the distribution). Since every next hop depends on the current tree (and only on the current tree), the whole process is a **Markov process**. It is light on memory and parallelisable but also affected by autocorrelation. Since we are using samples, a so-called **Monte Carlo method**, the whole affair is a **Markov Chain Monte Carlo** algorithm. It is one of many. It's called **Metropolis Hastings**. More complex MCMC algorithms exist. One class of such MCMC algorithms is called **Hamiltonian Monte Carlo** and these approaches use gradients to optimize the **proposal function**, i.e., the choice of the next tree to consider going to. They use the warm-up period to initialize certain tuning parameters, making them much faster and more reliable (at least if the distribution of leaves among neighboring trees is well-behaved).

How could Nature be sure that the plan succeeded? If not even Nature knows the distribution $P(t)$, how can we be sure that the Children's list gives representative samples to work with? - Certainty is petty. Reduction of uncertainty is key. Since we send out several Children in parallel, and since each Child distributed many apples, we can compare the list of trees delivered by each Child (= the set of samples in each chain). We can use statistics and ask: is it plausible that the set of samples in each chain has been generated from the same probability distribution? - The answer to this question can help reduce uncertainty about the quality of the sampling process.

9.6. Probabilistic modeling with *greta*

There are a number of software solutions for Bayesian posterior approximation, all of which implement a form of MCMC sampling, and most of which also realize at least one other form of parameter estimation. Many of these use a special language to define the model, and rely on a different programming language (like R, python, Julia etc.) to communicate with the program that does the sampling. Some options are:

- WinBUGS: a classic which has grown out of use a bit
- JAGS: another classic
- Stan: strongly developed current workhorse
- WebPPL: light-weight, browser-based full probabilistic programming language
- pyro: for probabilistic (deep) machine learning, based on pytorch
- *greta*: R-only probabilistic modeling package, based on Python and tensorflow

We will be using *greta* to look at explicitly formulated models. Later, when focusing on regression models, we will use an R package called `brms`, which relies on Stan, but we will not actively engage with Stan in this course.

9.6.1. Basics of *greta*

In order to approximate a posterior distribution over parameters for a model, given some data, using an MCMC algorithm, we need to specify the model for the sampler. In particular we must tell it about (i) the parameters, (ii) their priors, (iii) the likelihood function. The R-package *greta* allows us to specify this model inside of an R script in a syntax that looks like we are using regular R functions, even if in fact we are not.

For more information on *greta*, see the *greta* starting guide.

9.6.2. Binomial Model

Figure 9.9 shows the Binomial model for coin flips, as discussed before. We are going to implement it in *greta*.

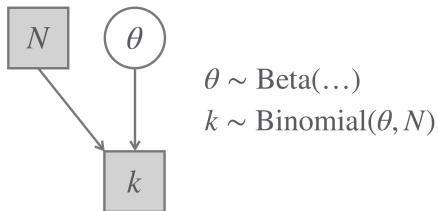


Figure 9.9.: The Binomial Model (repeated from before).

We data from the King of France example, where we are interested in the number $k = 109$ of “true” responses to sentences with a false presupposition over all $N = 311$ relevant observations. The research

9. Parameter estimation

question of interest is whether the idea that the parameter θ which models to overall disposition to answer “true” for these sentences could plausibly be fixed to $\theta = 0.5$. We declare the relevant counts using the `greta` function `as_data`. In this way all subsequent calculations with `k` and `N` will make it clear that we are defining `greta`-objects as part of a (so far) static model definition.

```
# greta data
k <- as_data(109)
N <- as_data(311)
```

Next, we need to tell the model what the prior for the latent parameter `theta` is. We use an uninformative Beta distribution here, $\theta \sim \text{Beta}(1, 1)$. This is realized with the function `beta` from the `greta` package. Notice that `beta` is *not* a function defined in R!

```
# coin bias & prior (here: uninformative)
theta <- beta(1, 1)
```

Finally, we tell `greta` about the likelihood of the data. This uses the `greta` functions `distribution` and `binomial`. By using `distribution` for the already defined and given data in variable `k` we inform `greta` that this line is (part of) the definition of the likelihood function.

```
# likelihood of data given theta
distribution(k) <- binomial(N, theta)
```

It remains to tell `greta` that the model definition is done and which parameters the MCMC sampler should return information about (here: `theta`).

```
# declare the greta model
m <- model(theta)
```

We can then call an the `greta` function `mcmc` which draws samples using an MCMC algorithm (the default is Hamiltonian Monte Carlo). We can specify additional parameters to tweak the sampler, such as the length of the warm-up period, the number of samples, chains, etc.

```
# take 4 chains of 1000 samples
draws <- greta::mcmc(
  model = m,
  n_samples = 1000,
  warmup = 1000,
  chains = 4
)
```

The resulting `draws` object is a special kind of object, an `MCMC.list` as defined in the `coda` package. This is not very important for us, however, we simply transform the samples to a tidy representation, using the function `ggs` from the `ggmcmc` package:

```
# cast results (type 'mcmc.list') into tidy tibble
tidy_draws = ggcmc::ggs(draws)
tidy_draws

## # A tibble: 4,000 x 4
##   Iteration Chain Parameter value
##       <int>   <int> <fct>     <dbl>
## 1         1       1 theta     0.343
## 2         2       2 theta     0.323
## 3         3       3 theta     0.352
## 4         4       4 theta     0.356
## 5         5       5 theta     0.356
## 6         6       6 theta     0.398
## 7         7       7 theta     0.398
## 8         8       8 theta     0.346
## 9         9       9 theta     0.405
## 10      10      10 theta    0.308
## # ... with 3,990 more rows
```

We can use these samples to compute Bayesian point- and interval estimates, for example:

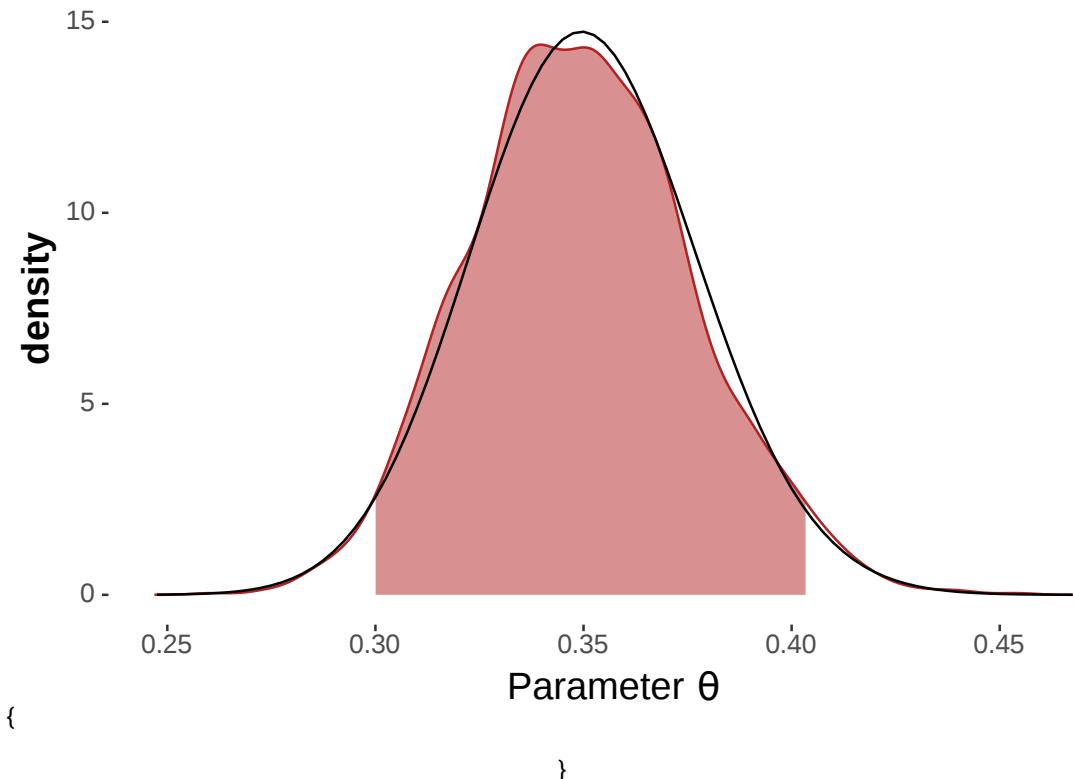
```
# obtain Bayesian point and interval estimates
Bayes_estimates <- tidy_draws %>%
  group_by(Parameter) %>%
  summarise(
    '|95%' = HDInterval::hdi(value)[1],
    mean = mean(value),
    '95|%' = HDInterval::hdi(value)[2]
  )
Bayes_estimates

## # A tibble: 1 x 4
##   Parameter `|95%`  mean `95|%` 
##   <fct>     <dbl> <dbl> <dbl>
## 1 theta      0.300 0.350 0.403
```

Using the (controversial) method of inspecting posterior estimates, we would conclude that $\theta = 0$ is not an *a posteriori* credible value for the inclination to judge the truth of sentences with a false presupposition.

Figure 9.6.2 moreover shows a density plot derived from the MCMC samples, together with the estimated 95% HDI and the true posterior distribution (in back), as derived by conjugacy.

\begin{figure}



\caption{Posterior over coin bias θ given $k = 109$ and $N = 311$ approximated by samples from `greta`, with estimated 95% credible interval (red area). The black curve shows the true posterior, derived through conjugacy.} \end{figure}

9.6.3. T-Test Model for Mental Chronometry

We will use the Mental Chronometry data to compare the reaction times in the “go/No-go” condition to the reaction times in the “discrimination” condition. To do this, we implement a T-Test model by hand in `greta`.

First we read in the data and get some handy summary statistics:

```
mc_data_cleaned <- read_csv('data_sets/mental-chromo-data_cleaned.csv',
                           col_types = cols(
                             submission_id = col_double(),
                             trial_number = col_double(),
                             block = col_character(),
                             stimulus = col_character(),
                             RT = col_double(),
                             handedness = col_character(),
```

```

        gender = col_character(),
        total_time_spent = col_double(),
        comments = col_character(),
        mean_C = col_double(),
        sd_C = col_double(),
        trial_type = col_character(),
        trial = col_double()
    ))
means_and_diffs <- mc_data_cleaned %>%
  filter(block != "reaction") %>%
  group_by(block) %>%
  summarise(
    mean_RT = mean(RT)
  ) %>%
  pivot_wider(
    names_from = block,
    values_from = mean_RT
  ) %>%
  mutate(
    `discr - gng` = discrimination - goNoGo
  )
means_and_diffs

## # A tibble: 1 x 3
##   discrimination goNoGo `discr - gng`
##       <dbl>     <dbl>      <dbl>
## 1          488.     427.      60.4

```

The model we will use for this situation is the T-Test model shown in Figure 9.10, repeated from the previous Chapter. We use the model which explicitly codes the difference between means (the variable δ) to directly address the question of whether $\delta = 0$ is a plausible point-value for this parameter.

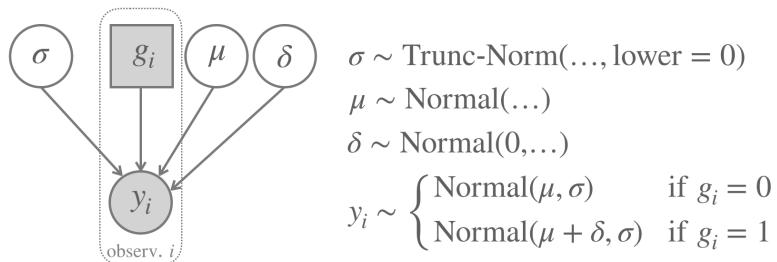


Figure 9.10.: A T-Test Model where one group is the default and the difference between group means is explicitly coded as a parameter.

9. Parameter estimation

We extract the relevant data and declare it as a `greta` object:

```
# isolate data vectors
RT_goNoGo <- mc_data_cleaned %>% filter(block == "goNoGo") %>% pull(RT)
RT_discrm <- mc_data_cleaned %>% filter(block == "discrimination") %>% pull(RT)
# declare as greta data arrays
y0 <- as_data(RT_goNoGo)
y1 <- as_data(RT_discrm)
```

We then define the model, using weakly informative, partly regularizing priors, i.e., priors that are informed by the data to ensure swift convergence (expecting value of `mean_0` to lie in plausible region), but that are not very biased to allow a large impact of the likelihood function (using relatively large standard deviations).

```
# priors
mean_0    <- normal(430, 50)
delta      <- normal(0, 100)
sigma      <- normal(100, 10, truncation = c(0, Inf))
# derived parameters
mean_1    <- mean_0 + delta
# likelihood
distribution(y0) <- normal(mean_0, sigma)
distribution(y1) <- normal(mean_1, sigma)
# model
m <- model(mean_0, mean_1, delta, sigma)## --- sampling ---
draws <- greta::mcmc(m, warmup = 4000, n_samples = 6000, thin = 2)
```

Bayesian point- and interval-estimates can be calculated from the posterior samples, including

```
tidy_draws = ggmcmc::ggs(draws)
Bayes_estimates <- tidy_draws %>%
  group_by(Parameter) %>%
  summarise(
    '|95%' = HDInterval::hdi(value)[1],
    mean = mean(value),
    '95|%' = HDInterval::hdi(value)[2]
  )
Bayes_estimates

## # A tibble: 4 x 4
##   Parameter `|95%`  mean `95|%`
##   <fct>     <dbl> <dbl> <dbl>
## 1 delta       49.6  60.1  71.2
```

```
## 2 mean_0      419.   427.   436.
## 3 mean_1      481.   488.   494.
## 4 sigma       101.   105.   109.
```

The Bayesian point-estimates for means and the difference correspond closely to the summary statistics we derived previously:

`means_and_diffs`

```
## # A tibble: 1 x 3
##   discrimination goNoGo `discr - gng`
##   <dbl>     <dbl>     <dbl>
## 1 488.     427.     60.4
```

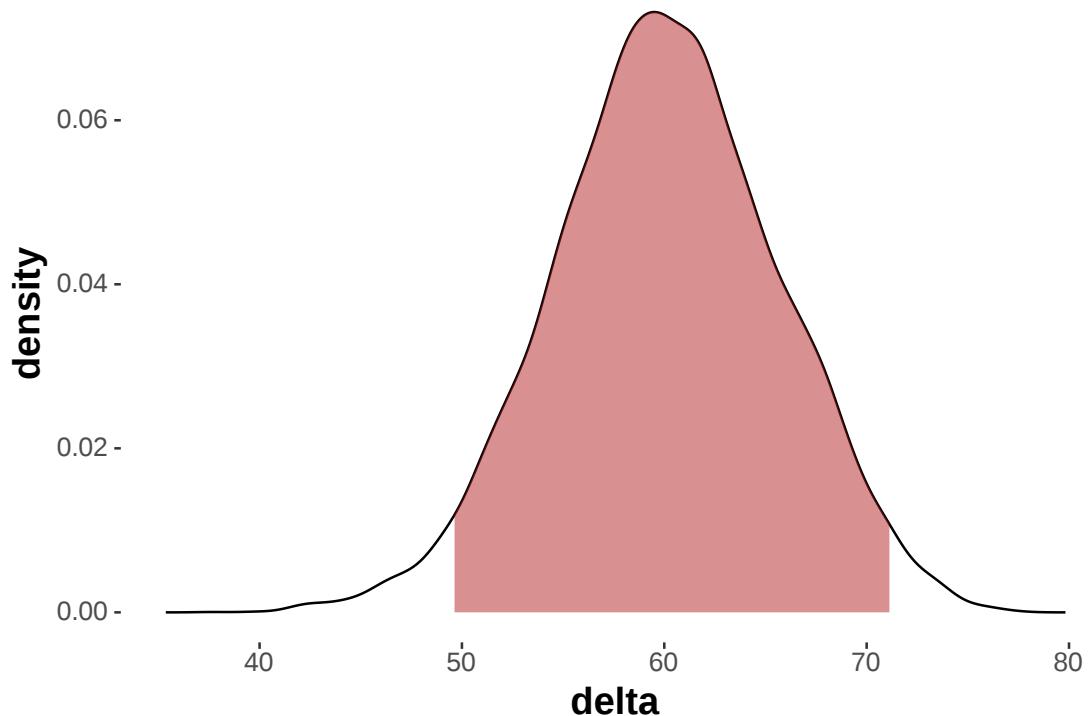
But we also now get indications of credible ranges of parameter values. Most interestingly, we obtain a 95% credible interval for the δ parameter, the difference between the means, which quite clearly does not include the case $\delta = 0$. The lower bound of the estimated 95% credible interval is more than 40 ms. We could conclude from this that, given this data set and the model used here, it is plausible that the difference in mean reaction times between the “discrimination” condition and the “go/no-go” condition is at least 40ms.

The plot below shows the density estimated from the posterior samples of δ , together with the estimated 95% credible interval.

```
dens <- filter(tidy_draws, Parameter == "delta") %>% pull(value) %>%
  density()

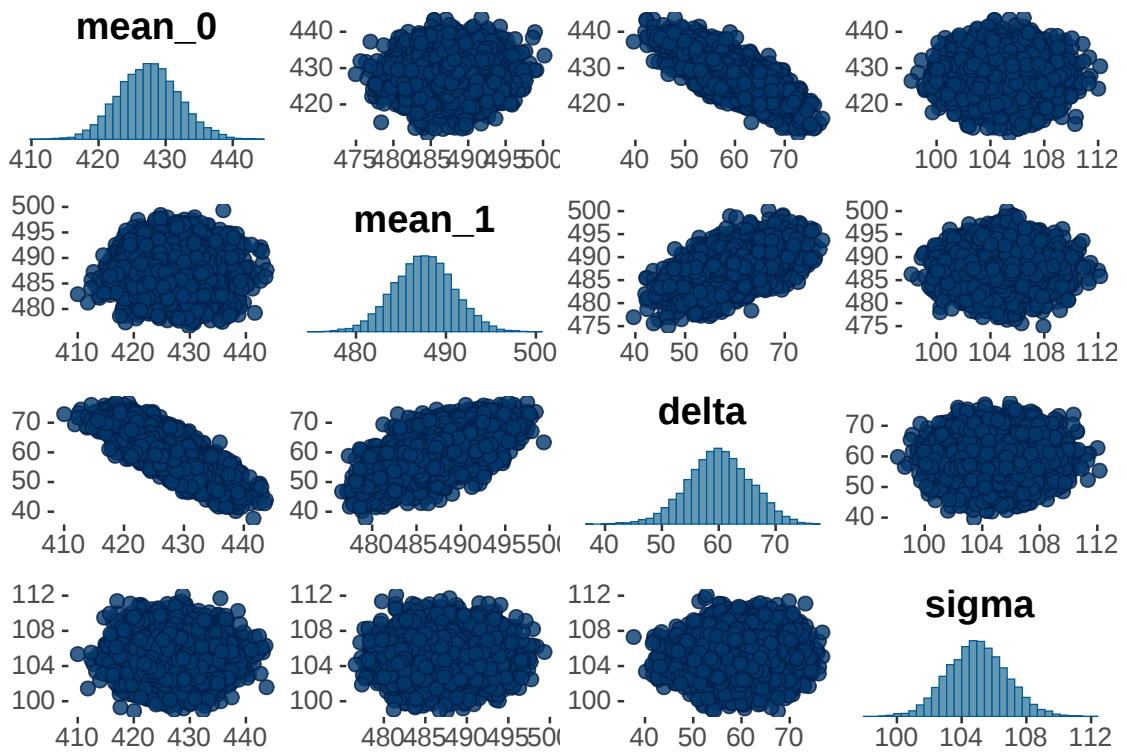
tibble(
  delta = dens$x,
  density = dens$y
) %>%
  ggplot(aes(x = delta, y = density)) +
  geom_line() +
  geom_area(aes(x = ifelse(
    delta > Bayes_estimates[1,2] %>% as.numeric &
    delta < Bayes_estimates[1,4] %>% as.numeric ,
    delta, 0)),
    fill = "firebrick", alpha = 0.5) +
  ylim(0, max(dens$y)) +
  xlim(min(dens$x), max(dens$x))
```

9. Parameter estimation



The actual posterior is multi-dimensional, and it always pays to inspect the full joint-posterior distribution so as not to miss any unexpected dependencies that might indicate sub-optimal inference or modeling. The `mcmc_pairs` function from the `bayesplot` package plots samples individually for each pair of parameters. Doing this we see the obvious (and perfectly fine) linear relation between estimates of `mean_0` and `delta`: the lower `mean_0` is estimated, the higher `delta` needs to be to yield a value of `mean_0` that explains the data well.

```
bayesplot::mcmc_pairs(draws)
```



10. Hypothesis Testing

Hypothesis testing is the workhorse of much of frequentist statistics. Researchers usually have a **research hypothesis**, such as:

- “truth-value judgements are at chance level for sentences with false presuppositions” (a claim about data from the King of France experiment); or:
- “discrimination takes longer than go/no-go decisions” (a claim about data from the mental chronometry experiment).

However, the hypothesis that is tested is not necessarily the research hypothesis. What is tested is rather a so-called **null hypothesis** H_0 .¹ For technical reasons, the null hypothesis is usually a *point-valued hypothesis* in that it assumes that some model parameter takes a single fixed value (see the earlier chapter “Expressing hypotheses with models”). Classical hypothesis testing then looks at whether the null hypothesis is plausible, given some observed data.

Of course, the null hypothesis is chosen in such a way as to give us information on the research question in the background. For example, depending on the framework we work in (see below), the null-hypothesis can be the research hypothesis, as in the case of the King of France example above. Alternatively, the research hypothesis is a statement about the existence of a difference between groups or an effect of an experimental manipulation, and in this case the null-hypothesis is opposed to the research hypothesis. E.g., the null-hypothesis could state that there is *no* difference in reaction times between discrimination and the go/no-go task.

Unfortunately, there is no single uncontroversially accepted and universally practiced recipe for frequentist hypothesis testing.² But there are three main approaches: Fisher’s approach, the Neyman-Pearson approach and the (modern) hybrid approach, often referred to simply as NHST (= null-hypothesis significance testing). These approaches differ in the way they relate the null hypothesis to the research question. They also differ in several technical details. We will cover differences and commonalities of these three major approaches in Section 10.4 at the end of this chapter, after we have covered basic technical notions and seen some applications in terms of some common tests.

More concretely, the chapter is structured as follows. Section 10.1) will elaborate on the notion of a *p*-value. We will then pay a visit to a very important mathematical result, the Central Limit Theorem, in Section 10.2, which allows us to derive approximations of *p*-values for complex cases. Section 10.3 covers a select sample of important tests. Section 10.4 discusses the three major approaches to

¹The term “null hypothesis” does not want to imply that the hypothesis is that some value of interest is equal to zero (although in practice that is frequently the case). The term rather implicates that this hypothesis is put out there in order to be possibly refuted, i.e., nullified, by the data (Gigerenzer 2004).

²There is no universally accepted Bayesian treatment either. This is why it is important to apply rational deliberation in each case, and not to succumb to the temptation of following easy recipes too quickly.

10. Hypothesis Testing

significance testing. Finally, Section 10.5 takes a step back and shows how we can think of hypothesis testing also as something more general, namely a method of **model checking**.

and after we have seen some applications of these basic ingredients (in Section 10.3).

The learning goals for this chapter are:

- become familiar with frequentist hypothesis testing
 - see the differences between different approaches
- understand key statistical notions such as:
 - sampling distribution
 - p -value
 - α - and β -error
- understand and be able to exploit the relation between p -values and confidence intervals
- understand and become able to apply and interpret basic tests:
 - binomial test, t-tests, ANOVA, linear regression, χ^2 -test

10.1. p -values

All prominent frequentist approaches to statistical hypothesis testing (see Section 10.4) agree that if empirical observations are sufficiently *unlikely* from the point of view of the null-hypothesis H_0 , this should be treated (in some way or other) as evidence *against* the null-hypothesis.³ A measure, perhaps approximate, of how unlikely (some aspect of) the data is in the light of H_0 is the p -value. To preview the main definition and intuition (to be worked out in detail hereafter), let's first consider a verbal and then a mathematical formulation.

Definition p -value. The p -value associated with observed data D_{obs} gives the probability, derived from the assumption that H_0 is true, of observing an outcome for the chosen test statistic that is at least as extreme evidence against H_0 as the observed outcome.

Formally, the p -value of observed data D_{obs} is:

$$p(D_{\text{obs}}) = P(T|H_0 \geq^{H_0,a} t(D_{\text{obs}}))$$

where $t: \mathcal{D} \rightarrow \mathbb{R}$ is a **test statistic** which picks out a relevant summary statistic of each potential data observation, $T|H_0$ is the **sampling distribution**, namely the random variable derived from test statistic t and

³To preview later material (see Section 10.4), the Neyman-Pearson approach goes further and also looks at evidence in favor of the null-hypothesis. It also tries to quantify something like evidence in favor of the research hypothesis. But Fisher's approach and some flavors of the hybrid approach only consider how much the data speaks against the null hypothesis.

the assumption that H_0 is true, and $\geq^{H_0,a}$ is a linear order on the image of t such that $t(D_1) \geq^{H_0,a} t(D_2)$ expresses that test value $t(D_1)$ is at least as extreme evidence *against* H_0 as test value $t(D_2)$.⁴

A few aspects of this definition are particularly important (and subsequent text is dedicated to making these aspects more comprehensible):

1. this is a frequentist approach in the sense that probabilities are entirely based on (hypothetical) repetitions of the assumed data-generating process, which assumes that H_0 is true;
2. the test statistic t plays a fundamental role and should be chosen such that:
 - it must necessarily select exactly those aspects of the data that matter to our research question,
 - it should optimally make it possible to derive a closed form (approximation) of T , and⁵
 - it would be desirable (but not necessary) to formulate t in such a way that the comparison relation $\geq^{H_0,a}$ coincides with a simple comparison of numbers: $t(D_1) \geq^{H_0,a} t(D_2)$ iff $t(D_1) \geq t(D_2)$;
3. there is an assumed data-generating model buried inside notation $T|H_0$; and
4. the notion of “more extreme evidence against H_0 ”, captured in comparison relation $\geq^{H_0,a}$ depends on our epistemic purposes, i.e., what research question we are ultimately interested in.⁶

The following sections will elaborate on all of these points. It is important to mention that especially the third aspect (that there is an implicit data-generating model “inside of” classical hypothesis tests) is not something that receives a lot of emphasis in traditional statistics textbooks. Bad textbooks do not even mention the assumptions implicit in a given test. Better textbooks mention these assumptions, good ones stress them. Using a model-centric approach, as we do here, tries to go even a bit further. We will not only stress key assumptions behind a test but present all of the assumptions behind classical tests in a graphical model, similar to what we did for Bayesian models. This arguably makes all implicit assumptions maximally transparent in a concise and lucid representation. It will also help see parallels between Bayesian and frequentist approaches, thereby helping to see both as more of the same rather than as something completely different. In order to cash in this model-based approach, the following sections will therefore introduce new graphical tools to communicate the data-generating model implicit in the classical tests we cover.

⁴This formulation in terms of a context-dependent (i.e., H_0 -dependent ordering) is not usual. However, the interpretation is de facto context-dependent in this way, and so it makes sense to highlight this aspect of the use of *p*-values also formally. Notice, however, that we can get rid of the context-dependence by using different test-statistics. But this is also not how it is done in practice. Essentially, this definition aims for maximal generality so as to cover all cases of use. Since the class of use cases is fuzzy, the definition needs this flexibility. Alternative mathematical definitions that appear to be simpler just do not capture all the use cases.

⁵This latter aspect has been particularly important historically. Given more readily available computing power, alternative approaches based on Monte Carlo simulation of *p*-values can also be used.

⁶It is admittedly a bit of a notational overkill to write this comparison relation as a function of H_0 and H_a (the alternative hypothesis). Other definitions of the *p*-value do not. But the comparison is context dependent, and you deserve to see this clearly. To see it clearly, a certain heaviness of notation is the price to pay.

10.1.1. Binomial Model - frequentist version

We start with the Binomial Model because it is the simplest and perhaps most intuitive case. We work out what a *p*-value is for data for this model and introduce the new graphical language to communicate “frequentist models” in the following. We also introduce the notions of *test statistic* and *sampling distribution* based on a case that should be very intuitive, if not familiar.

The Binomial Model was covered before from a Bayesian point of view, where we represented it using graphical notation like in Figure 10.1 (repeated from above). Remember that this is a model to draw inferences about a coin’s bias θ based on observations of outcomes of flips of that coin. The Bayesian modeling approach treated the number of observed heads k and the number of flips in total N as given, and the coin’s bias parameter θ as latent.

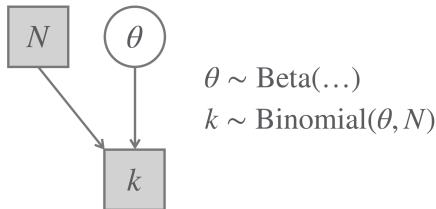


Figure 10.1.: The Binomial Model (repeated from before) for a Bayesian approach to parameter inference/testing.

Actually, this way of writing the Binomial Model is rather a shortcut. It glosses over each individual data observation (whether the i -the coin flip was heads or tails) and jumps directly to the most relevant summary statistic of how many of the N flips were heads. This might, of course, be just the relevant level of analysis. If our assumption is true that the outcome of each coin flip is independent of any other flip, and given our goal to learn something about θ , all that really matters is k . But to prepare ourselves for subsequent frequentist approaches, and in order to appreciate (later!) how powerful a tool a test statistic can be, we can also rewrite the Bayesian model from Figure 10.1 as the equivalent extended model in Figure 10.2. In the latter representation, the individual outcomes of each flip are represented as $x_i \in \{0, 1\}$. Each individual outcome is sampled from a Bernoulli distribution. Based on the whole vector of x_i -s, together with knowledge of N , we derive the **test statistic** k , which maps each observation (a vector x or zeros and ones) to a single number k (the number of heads in the vector). Notice that the node for k has a solid double edge, indicating that it follows deterministically from its parent nodes. This is why we can think of k as a sample from a random variable constructed from “raw data” observations x .

Compare this latter representation in Figure 10.2 with the frequentist Binomial Model in Figure 10.3. The frequentist model treats the number of observations N as observed, just like the Bayesian model. But it also fixes a specific value for the coin’s bias θ . This is where the (point-valued) null hypothesis comes in. For purposes of analysis, we fix the value of the relevant unobservable latent parameter to a specific value (because we do not want to assign probabilities to latent parameters, but we still like to talk about probabilities somehow). In our graphical model in Figure 10.3 the node for the coin’s bias is shaded (=treated as known) but also has a dotted second edge to indicate that this is where our null-hypothesis assumption kicks in. We then treat the data vector x and with it the associated test statistic k as

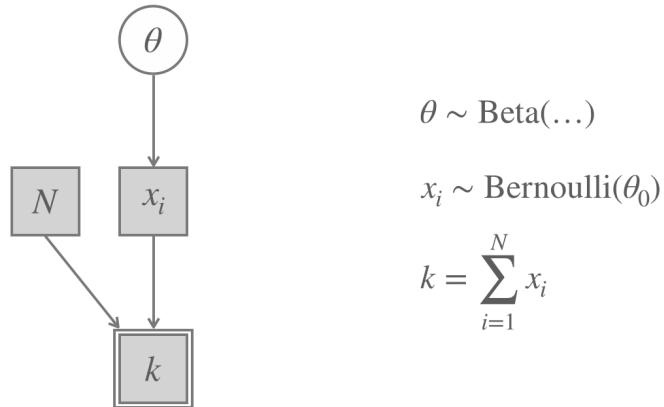


Figure 10.2.: The Binomial Model for a Bayesian approach, extended to show ‘raw observations’ and the ‘summary statistic’ implicitly used.

unobserved. The data we actually observed will, of course, come in at some point. But the frequentist model leaves the observed data out at first in order to bring in the kinds of probabilities frequentist approaches feel comfortable with: probabilities derived from (hypothetical) repetitions of chance events. So, the frequentist model can now make statements about the likelihood of (raw) data x and values of the derived summary statistic k based on the assumption that the null hypothesis is true. Indeed, for the case at hand, we already know that the **sampling distribution**, i.e., the distribution of values for k given θ_0 is the Binomial distribution.

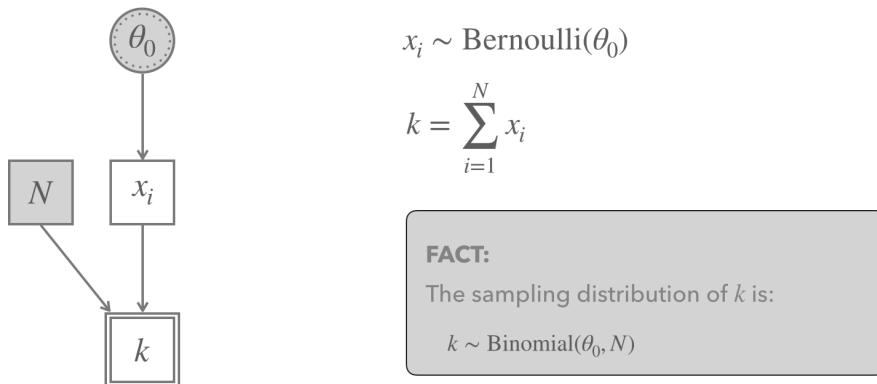


Figure 10.3.: The Binomial Model for a frequentist binomial test.

Let’s take a step back. The frequentist model for the binomial case considers (“raw”) data of the form $\langle x_1, \dots, x_N \rangle$ where each $x_i \in \{0, 1\}$ indicates whether the i -th flip was a success (= heads, =1) or a failure (=tails, =0). We identify the set of all binary vectors of length N as the set of hypothetical data which we could, in principle, observe in fictitious repetition of this data-generating process. \mathcal{D}^{H_0} is then the random variable that assigns each potential observation $D = \langle x_1, \dots, x_N \rangle$ the probability with which it would occur if H_0 (=a specific value of θ) is true. In our case, that is:

$$P(\mathcal{D}^{H_0} = \langle x_1, \dots, x_N \rangle) = \prod_{i=1}^N \text{Bernoulli}(x_i, \theta_0)$$

The model does not work with this raw data and its implied distribution (represented by random variable \mathcal{D}^{H_0}), it uses a (very natural!) **test statistic** t : $\langle x_1, \dots, x_N \rangle \mapsto \sum_{i=1}^N x_i$ instead. The **sampling distribution** for this model is therefore the distribution of values for the derived measure k - a distribution which follows from the distribution of the raw data (\mathcal{D}^{H_0}) and this particular test statistic t . In its most general form, we write the sampling distribution as $T^{H_0} = t(\mathcal{D}^{H_0})$.⁷ It just so happens (what a relief!) that we know how to express T^{H_0} in a mathematically very concise fashion. It's just the Binomial distribution, so that $k \sim \text{Binomial}(\theta_0, N)$. (Notice how the sampling distribution is really a function of θ_0 (the null-hypothesis) and also of N .)

10.1.2. *p*-values for the Binomial Model

After seeing a frequentist model and learning about test statistic and sampling distribution, let's explore what a *p*-value is based on the frequentist Binomial Model. Our running example will be the 24/7 case, where $N = 24$ and $k = 7$. Notice that we are glossing over the "raw" data immediately and work with the value of the test statistic of the observed data directly: $t(D_{\text{obs}}) = 7$.

Remember that, by the definition given above, $p(D_{\text{obs}})$ is the probability of observing a value of the test statistic that is at least as extreme evidence against H_0 as $t(D_{\text{obs}})$, under the assumption that H_0 is true:

$$p(D_{\text{obs}}) = P(T^{H_0} \geq^{H_0,a} t(D_{\text{obs}}))$$

To fill this with life, we need to set a null hypothesis, i.e., a value θ_0 of coin bias θ , that we would like to collect evidence *against*. A fixed H_0 will directly fix T^{H_0} but we will have to put extra thought into how to conceptualize $\geq^{H_0,a}$ for any given H_0 . To make exactly this clearer is the job of this section. Specifically, we will look at what is standardly called a **two-sided p-value** and a **one-sided p-value**.

As stated in the introduction to this chapter, since this testing routine is geared to give us evidence against H_0 , we should choose H_0 in such a way as to give us information about the research question that really matter to us. So, let's suppose that our research question is either one of the following:

- Is the coin fair ($\theta = 0.5$)?
- Is the coin biased towards heads ($\theta > 0.5$)?

As we will see below, in both cases the null hypothesis will be the same: we are going to assume that $\theta_0 = 0.5$. But given our research question, the **alternative hypothesis** H_a to the null-hypothesis will be different. In the case of testing for fairness ($\theta = 0.5$), the pair of null hypothesis and alternative hypothesis are:

⁷Most often the random variable capturing the sampling distribution is just written as T , but it does make sense to stress also notationally that T depends crucially on H_0 .

$$H_0: \theta = 0.5 \quad H_a: \theta \neq 0.5$$

Notice that here the alternative hypothesis H_a is two-sided in the sense that it departs from H_0 left and right, so to speak.

But in the case of testing whether there is a bias towards heads ($\theta > 0.5$), the null hypothesis is the same but the alternative hypothesis is one-sided:⁸[An alternative way of looking at this case is to say that, at first, we test the interval-range null hypothesis $\theta > 0.5$, so that research and null hypothesis coincide. To gather evidence against this interval-range null hypothesis, we will compare it to the alternative hypothesis $\theta < 0.5$. Since we need a point-value to easily generate the sampling distribution, the question becomes which point for the null-hypothesis to take. We then choose the value that if we gather evidence *against* this value, this is most disastrous to the null hypothesis in question.]

$$H_0: \theta = 0.5 \quad H_a: \theta < 0.5$$

Research question $\theta = 0.5$. To begin with, assume that we want to address the question of whether the coin is fair, i.e., whether $\theta = 0.5$. In this case, we identify the research question with the null hypothesis and set $\theta_0 = 0.5$. Figure 10.4 shows the sampling distribution of the test statistic k . The probability of the observed value of the sampling statistic is shown in red.

The question we need to settle to obtain a p -value is which alternative values of k would count as more extreme evidence against the chosen null hypothesis, i.e., how to interpret $\geq^{H_{0,a}}$ for this case. The obvious approach is to use the probability of any value of the test statistic k directly and say that observing D_1 counts as at least as extreme evidence against H_0 as observing D_2 , $t(D_1) \geq^{H_{0,a}} t(D_2)$, iff the probability of observing the test statistic associated with D_1 is at least as unlikely as observing D_2 : $P(T|H_0 = t(D_1)) \leq P(T|H_0 = t(D_2))$. To calculate the p -value in this way, we therefore need to sum up the probabilities of all values k under the Binomial distribution (with parameters $N = 24$ and $\theta = \theta_0 = 0.5$) that are no larger than the value of the observed $k = 7$. In mathematical language:⁸

$$p(k) = \sum_{k'=0}^N [\text{Binomial}(k', N, \theta_0) \leq \text{Binomial}(k, N, \theta_0)] \text{Binomial}(k', N, \theta_0)$$

In code, we calculate the this p -value as follows:

```
# exact p-value for k=7 with N=24 and null-hypothesis theta = 0.5
k_obs <- 7
N <- 24
theta_0 <- 0.5
tibble( lh = dbinom(0:N, N, theta_0) ) %>%
  filter( lh <= dbinom(k_obs, N, theta_0) ) %>%
  pull(lh) %>% sum %>% round(5)
```

⁸Here, the bracket notation [Boolean] is the Iverson bracket, evaluation to 1 if the Boolean expression is true and to 0 otherwise.

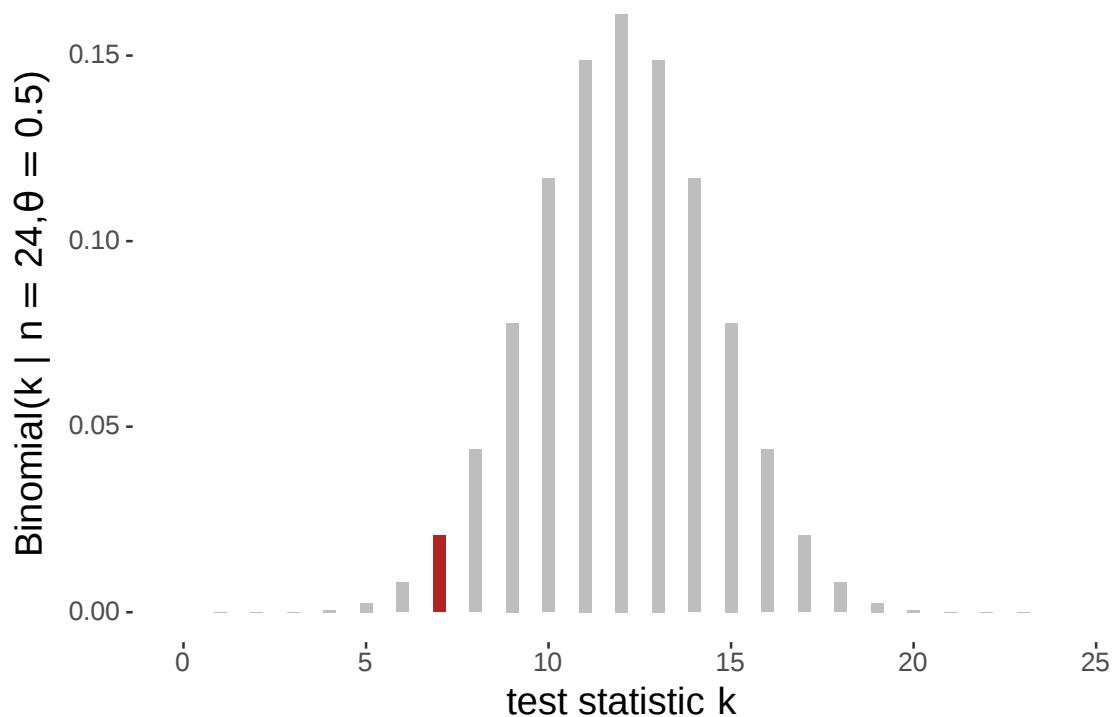


Figure 10.4.: Sampling distribution (here: Binomial distribution) and probability associated with observed data $k = 7$ highlighted in red, for $N = 24$ coin flips, under the assumption of a null-hypothesis $\theta = 0.5$.

```
## [1] 0.06391
```

Figure 10.5 shows the values that need to be summed over in red.

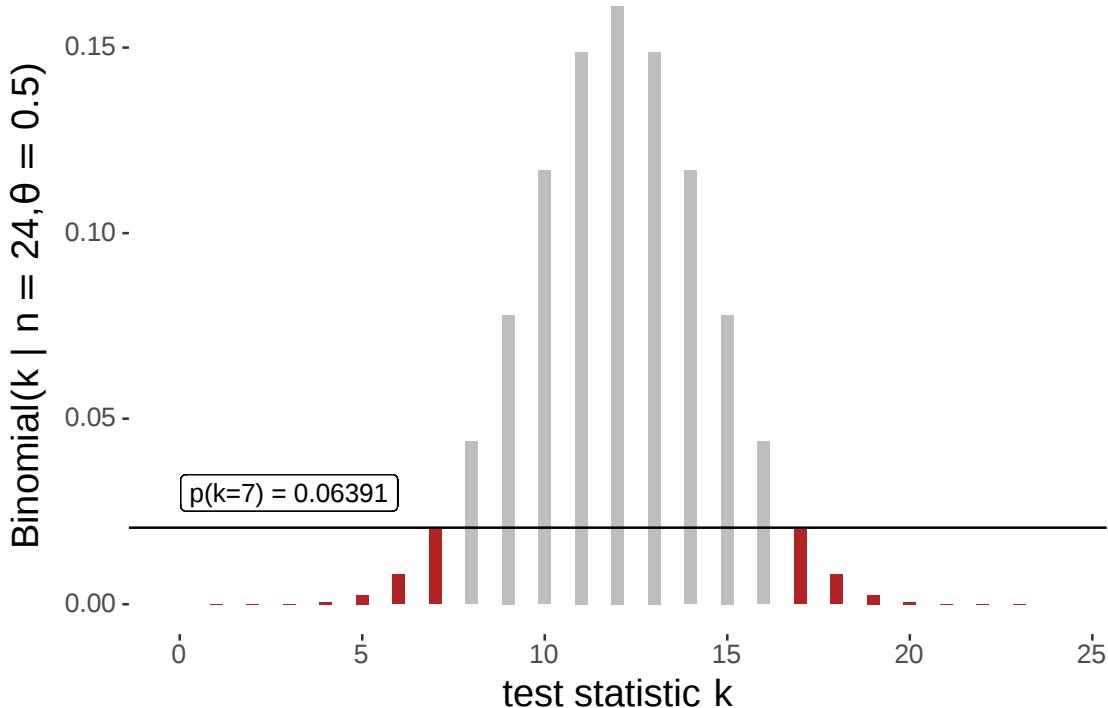


Figure 10.5.: Sampling distribution (Binomial likelihood function) and p -value for the observation of $k = 7$ successes in $N = 24$ coin flips, under the assumption of a null-hypothesis $\theta = 0.5$.

Of course, R also has a built-in function for a Binomial test. We can use it to verify that we get the same result for the p -value:

```
binom.test(
  x = 7,      # observed successes
  n = 24,      # total nr. observations
  p = 0.5      # null hypothesis
)

##
## Exact binomial test
##
## data: 7 and 24
## number of successes = 7, number of trials = 24, p-value = 0.06391
```

10. Hypothesis Testing

```
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1261521 0.5109478
## sample estimates:
## probability of success
##                      0.2916667
```

Research question $\theta < 0.5$. Let's now look at the case where our research hypothesis is that $\theta > 0.5$, i.e., that there is a bias towards heads. Again we need to settle what the null hypothesis should be and how $\geq^{H_{0,a}}$ should be interpreted. In this case, we could consider an interval-valued null hypothesis like $\theta_0 > 0.5$ (making the research hypothesis the null). But even so, we would still resort to testing a point-valued null hypothesis $\theta_0 = 0.5$ eventually. This is because point-valued null hypotheses are easy to work with.⁹ It has therefore become customary to pick the value from the relevant interval which is most favorable of the interval-valued null-hypothesis. In case we find strong evidence *against* even the most favorable value from the interval, than this does constitute the strongest possible case *against* the whole interval-based null hypothesis.

But even though we use the same null-value of $\theta_0 = 0.5$, the calculation of the p -value will be different from the case we looked at previously. The reason lies in a change to what we should consider more extreme evidence against this interval-valued null hypothesis, i.e., the interpretation of $\geq^{H_{0,a}}$. In the case at hand, observing values of k larger than 12, even if they are unlikely for the point-valued hypothesis $\theta_0 = 0.5$ do not constitute evidence against the interval-valued hypothesis we are interested in. So, therefore, we disregard the contribution of the right hand side in Figure 10.5 to arrive at a picture like in Figure 10.6. The associated p -value with this, so-called **one-sided test**, is consequently:

```
k_obs <- 7
N <- 24
theta_0 <- 0.5
# exact p-value for k=7 with N=24 and null-hypothesis theta > 0.5
dbinom(0:k_obs, N, theta_0) %>% sum %>% round(5)

## [1] 0.03196
```

We can double-check against the built-in function `binom.test` when we ask for a one-sided test:

```
binom.test(
  x = 7,      # observed successes
  n = 24,     # total nr. observations
  p = 0.5,    # null hypothesis
  alternative = "less" # the alternative to compare against is theta < 0.5
)
```

⁹The problem for interval-valued null hypothesis is that we would need to specify some more information about how to rank parameters, if only to say that they are all ranked equally (*plausible, a priori*), which is something that we try to avoid in frequentist approaches.

```
##  
## Exact binomial test  
##  
## data: 7 and 24  
## number of successes = 7, number of trials = 24, p-value = 0.03196  
## alternative hypothesis: true probability of success is less than 0.5  
## 95 percent confidence interval:  
## 0.0000000 0.4787279  
## sample estimates:  
## probability of success  
## 0.2916667
```

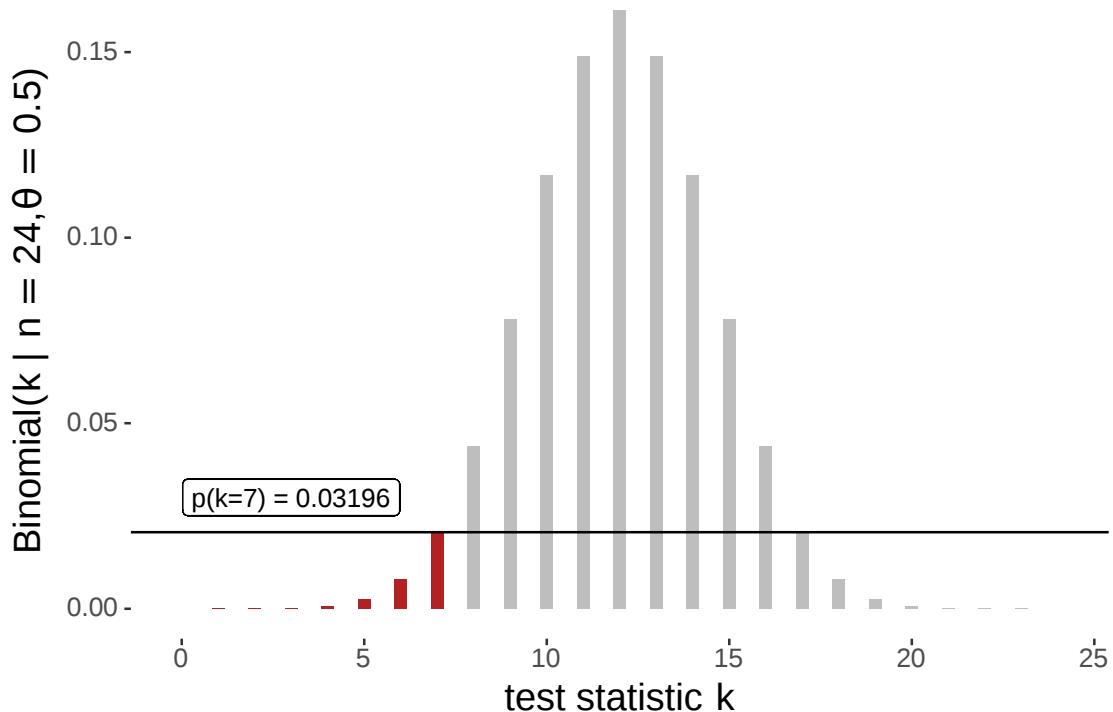


Figure 10.6.: Sampling distribution (Binomial likelihood function) and p -value for the observation of $k = 7$ successes in $N = 24$ coin flips, under the assumption of a null-hypothesis $\theta > 0.5$.

10.1.3. Statistical significance

Fisher's early writing suggests that he considered p -values as quantitative measures of strength of evidence against the null hypothesis. What would need to be done or concluded from such a quantitative measure would need to depend on further careful case-by-base deliberation. In contrast, present practice

10. Hypothesis Testing

often uses p -values to check whether a test result is noteworthy in a categorical, not quantitative way. Fixing an α -level of significance (with common values $\alpha \in \{0.05, 0.01, 0.001\}$), we say that a test result is significant if the p -value of the observed data is lower than the specified α .

Significance of a test result, as a categorical measure, can then be further interpreted as a trigger for decision making. Commonly, a significant test results is interpreted as the signal to reject the null hypothesis, i.e., to speak and act as if it was false.

10.1.4. p -values and α -errors

Some blends of frequentist statistics focus on establishing a tight regime of error control (see Section 10.4 below): we want to keep a cap on the long-run amount of errors that we make in statistical decision making. The p -value is then related to the α -error, or Type-I error, when decisions to reject the null hypothesis are made based on whether the p -value crosses a particular threshold.

An α -error occurs when we falsely reject the null hypothesis, i.e., we reject the null hypothesis (as implausible) when it is actually true. Suppose we decide to reject the null hypothesis that the coin is fair in a two-sided test exactly when the p -value of our test is smaller than α , e.g., $\alpha = 0.05$. In this case, α is an upper bound on the α -error.

10.1.5. Relation of p -values to confidence intervals

There is a close relation between p -values and confidence intervals.¹⁰ For a two sided test of null hypothesis $\theta = \theta_0$, with alternative $H_a: \theta \neq \theta_0$, it holds for all possible data observations D that

$$p(D) < \alpha \text{ iff } \theta_0 \notin \text{CI}(D)$$

where $\text{CI}(D)$ is the $(1 - \alpha) \cdot 100\%$ confidence interval constructed for data D .

This connection is intuitive when we think about long-term error. Decisions to reject the null hypothesis are false in exactly $(\alpha \cdot 100)\%$ of the cases when the null hypothesis is true. The definition of a confidence interval was exactly the same: the true value should like outside a $(1 - \alpha) \cdot 100\%$ confidence interval in exactly $(\alpha \cdot 100)\%$ of the cases. (Of course, this is only a vague and intuitively appealing argument based on the overall rate, not any particular case.)

10.1.6. Distribution of p -values

A result that might seem surprising at first is that if the null hypothesis is true, the distribution of p -values is uniform. This, however, is intuitive on second thought. Mathematically it is a direct consequence of the **Probability Integral Transform Theorem**.

¹⁰An important caveat applies here. There can be different (approximate) ways of defining p -values and confidence intervals. The relation described here does not hold, when the (approximate) way of computing the p -value does not match the (approximate) way of computing the confidence interval.

Theorem 10.1 (Probability Integral Transform). *If X be a continuous random variable with cumulative distribution function F_X , the random variable $Y = F_X(X)$ is a uniformly distributed over interval $[0; 1]$, i.e., $y \sim \text{Uniform}(0, 1)$.*

Proof. Notice that the cumulative density function of a standard uniform distribution $y \sim \text{Uniform}(0, 1)$ is linear line with intercept 0 and slope 1. It therefore suffices to show that $F_Y(y) = y$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && [\text{def. of cumulative distribution}] \\ &= P(F_X(X) \leq y) && [\text{by construction / assumption}] \\ &= P(X \leq F_X^{-1}(y)) && [\text{applying inverse cumulative function}] \\ &= F_X(F_X^{-1}(y)) && [\text{def. of cumulative distribution}] \\ &= y && [\text{inverses cancel out}] \end{aligned}$$

□

Seeing the uniform distribution of p -values (under a true null hypothesis) helps appreciate how the α -level of significance is related to long-term error control. If the null hypothesis is true, the probability of a significant test result is exactly the significance level.

10.1.7. How (not) to interpret p -values

Though central to much of frequentist statistics, p -values are frequently misinterpreted, even by seasoned scientists (Haller and Krauss 2002). To repeat, the p -value measures the probability of observing, if the null-hypothesis is correct, a value of the test statistic that is (in a specific, contextually specified sense) more extreme than the value of the test statistic that we assign to the observed data. We can therefore treat p -values as a measure of evidence *against* the null-hypothesis. And if we want to be even more precise, we interpret this as evidence against the whole assumed data-generating process, a central part of which is the null-hypothesis.

The p -value is *not* a statement about the probability of the null hypothesis given the data. So, it is *not* something like $P(H_0 \mid D)$. The latter is a very appealing notion, but it is one that the frequentist denies herself access to. It can also only be computed based on some consideration of prior plausibility of H_0 in relation to some alternative hypothesis. Indeed, to calculate $P(H_0 \mid D)$ is the topic of the chapter on Model Comparison.

10.2. Central Limit Theorem

The previous section expanded of the notion of a p -value and it showed how to calculate p -values for different kinds of research questions for data from repeated Bernoulli trials (= coin flips). We saw that a natural test statistic is the Binomial distribution. The Binomial distribution described the sampling distribution precisely, i.e., the sampling distribution for the frequentist Binomial Model as we set it up is

10. Hypothesis Testing

the Binomial distribution. Unfortunately, there are models and types of data for which the sampling distribution is not known precisely. In these cases, frequentist statistics works with approximations to the true sampling distribution. These approximations get better the more data was observed, i.e., these are limit-approximations that hold in the limit when the amount of data observed goes towards infinity. For small samples, the error might be substantial. Rules of thumb have become conventional guides for judging when (not) to use a given approximation. Which (approximation for a) sampling distribution to use needs to be decided on a case-by-case basis.

To establish that a particular distribution is a good approximation of the true sampling distribution, the most important formal result is the *Central Limit Theorem* (CLT). In rough terms, the CLT says that, under certain conditions, we can use a normal distribution as an approximation of the sampling distribution.

To appreciate the CLT, let's start with another seminal results, the **Law of Large Numbers**, which we have already relied on when we discussed a sample-based approach to representing probability distributions.

For example, the Law of Large Numbers justifies why taking (large) samples from a random variable sufficiently approximates a mean (the most prominent Bayesian point-estimator of, e.g., a posterior approximated by samples from MCMC algorithms).

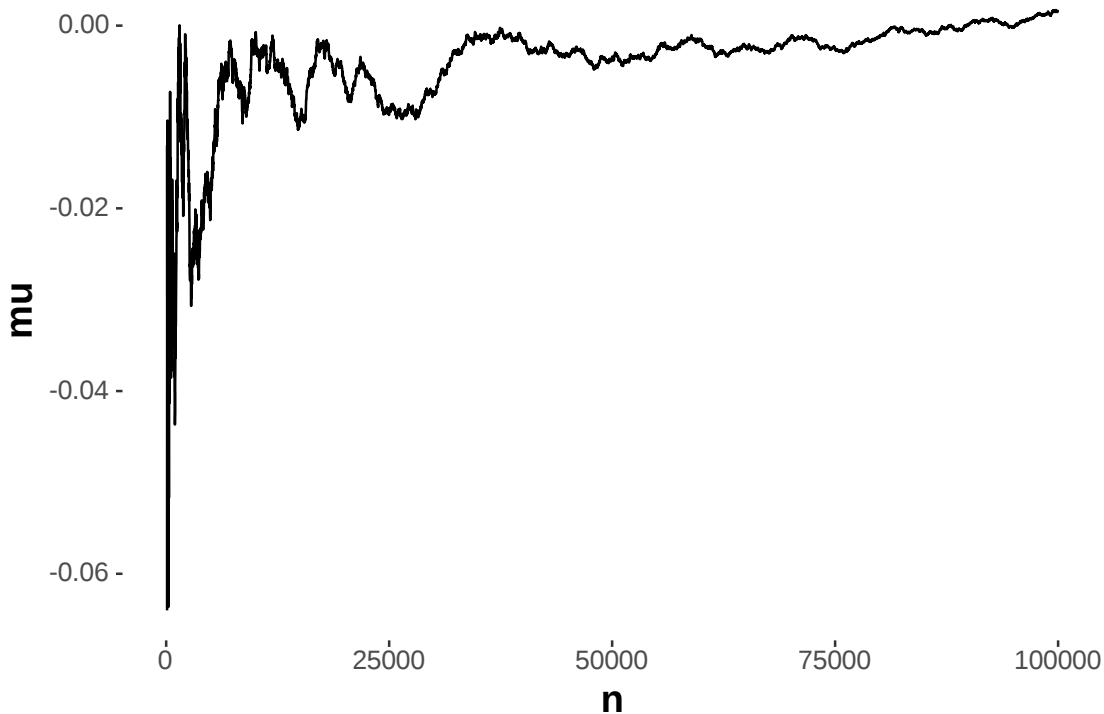
Theorem 10.2 (Law of Large Numbers). *Let X_1, \dots, X_n be a sequence of n differentiable random variables with equal mean, such that $\mathbb{E}_{X_i} = \mu_X$ for all $1 \leq i \leq n$.¹¹ As the number of samples n goes to infinity the mean of any tuple of samples, one from each X_i , converges almost surely to μ_X :*

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu_X\right) = 1$$

Computer simulation makes the point and usefulness of this fact easier to appreciate:

```
# sample from a standard normal distribution (mean = 0, sd = 1)
samples <- rnorm(100000)
# collect the mean after each 10 samples & plot
tibble(
  n = seq(100, length(samples), by = 10)
) %>%
  group_by(n) %>%
  mutate(
    mu = mean(samples[1:n])
) %>%
  ggplot(aes(x = n, y = mu)) +
  geom_line()
```

¹¹Though the result is more general, it is convenient to think of a natural application as the case where all X_i are samples from the exact same distribution.



For practical purposes, think of the Central Limit Theorem as an extension of the Law of Large Numbers.

While the latter tells us that, as $n \rightarrow \infty$, the mean of repeated samples from a random variable X converges to the mean of X , the Central Limit Theorem tells us something about the distribution of our estimate of X 's mean. The Central Limit Theorem tells us that the sampling distribution of the mean approximates a normal distribution for large enough sample size.

Theorem 10.3 (Central Limit Theorem). *Let X_1, \dots, X_n be a sequence of n differentiable random variables with equal mean $\mathbb{E}_{X_i} = \mu_X$ and equal finite variance $\text{Var}(X_i) = \sigma_X^2$ for all $1 \leq i \leq n$.¹² The random variable S_n which captures the distribution of the sample mean for any n is:*

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i$$

As the number of samples n goes to infinity the random variable $\sqrt{n}(S_n - \mu_X)$ converges in distribution to a normal distribution with mean 0 and standard deviation σ_X .

A proof of the CLT is not trivial, and we will omit it here. We will only point to the CLT when justifying approximations of sampling distributions, just as exemplified in the next section, which deals with Pearson's χ^2 -test.

¹²As with the Law of Large Numbers, the most common application is the case where all X_i are samples from the exact same distribution.

10.3. Selected tests

10.3.1. Pearson's χ^2 -tests

There are many tests that use the χ^2 -distribution as an (approximate) sampling distribution. But given relevance and historical prominence, the name “ χ^2 -test” is usually interpreted to refer to one of several flavor's of what we could specifically call “Pearson's χ^2 -test”.

We will look at two flavors here. Pearson's χ^2 -test for **goodness of fit** tests whether an observed vector of counts is well explained by a given vector of predicted proportion. Pearson's χ^2 -test for **independence** tests whether a (two-dimensional) table of counts could plausibly have been generated by a process of independently selecting the column and the row category. We will explain how both of these work based on an application to the BLJM data, which we load as usual:

```
data_BLJM_processed <- read_csv(url('https://processed.githubusercontent.com/michaelm...'))
```

The focus is on the counts of music-subject choices:

```
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM, LB)
BLJM_associated_counts

## # A tibble: 4 x 3
##   JM     LB     n
##   <chr> <chr> <int>
## 1 Jazz   Biology    38
## 2 Jazz   Logic      26
## 3 Metal  Biology    20
## 4 Metal  Logic      18
```

Remember that the lecturer's bold conjecture was that a preference for Logic over Biology goes together with a preference for Metal over Jazz. Visualization suggests that there might be such a trend but that the (statistical) jury is still out as to whether this conjecture has empirical support.

10.3.1.1. Pearson's χ^2 -test for goodness of fit

“Goodness of fit” is a term used in model checking (a.k.a. model criticism, model validation, ...). In such a context, tests for goodness-of-fit investigate whether a model's predictions compatible with the observed data. Pearson's χ^2 -test for goodness of fit does exactly this for categorical data.

Categorical data is data where each data observation falls into one of several unordered categories. If we have k such categories, a **prediction vector** $\vec{p} = \langle p_1, \dots, p_k \rangle$ is a probability vector of length k such that p_i gives the probability with which a single data observation falls into the i -th category. The likelihood of a single data observation is given by the Categorical distribution, and the likelihood of N data observations is given by the Multinomial distribution. These are generalizations of the Bernoulli and Binomial distributions, which cover the case of two unordered categories, to the case of more than two unordered categories.

The BLJM data supplies us with categorical data. Here is the vector of counts of how many participants selected a given music+subject pair:

```
# add category names
BLJM_associated_counts <- BLJM_associated_counts %>%
  mutate(
    category = str_c(
      BLJM_associated_counts %>% pull(LB),
      "-",
      BLJM_associated_counts %>% pull(JM)
    )
  )
counts_BLJM_choice_pairs_vector <- BLJM_associated_counts %>% pull(n)
names(counts_BLJM_choice_pairs_vector) <- BLJM_associated_counts %>% pull(category)
counts_BLJM_choice_pairs_vector

##  Biology-Jazz      Logic-Jazz  Biology-Metal   Logic-Metal
##        38             26          20            18
```

Figure 10.7 shows a crude plot of these counts, together with a baseline prediction of equal proportion in each category.

Pearson's χ^2 -test for goodness of fit allows us to test whether this data could plausibly have been generated by (a model whose predictions are given by) a prediction vector $\vec{p} = \langle p_1, \dots, p_4 \rangle$, where p_1 would be the predicted probability of a choice pair "Biology-Jazz" occurring for a single participant etc. Frequently, this test is used to check whether a baseline equal distribution could have generated the data. We do that here, too. We form the null-hypothesis that $\vec{p} = \vec{p}_0$ with $p_{0i} = \frac{1}{4}$ for all categories i .

Figure 10.8 shows a graphical representation of the model implicitly assumed in the background for a Pearson's χ^2 -test for goodness of fit. The model assumes that the observed vector of counts (like our `counts_BLJM_choice_pairs_vector` from above) is follows a Multinomial distribution.¹³ Each vector of (hypothetical) data is associated with a test statistic, called χ^2 , which sums over the

¹³Notice that for economy or presentation we now (again) gloss over the "raw" data of individual choices, and present the summarized count data instead. In the previous case of the Binomial Test it made good pedagogical sense to tease apart the "raw" observations from the summarized counts because this helped to show what the test statistic is for a case where the choice of test statistic was very, very obvious; so much so, that we would normally not even bother to make it explicit. Now that we understand what a test statistic is in principle, we can gloss over some steps of data summarizing.

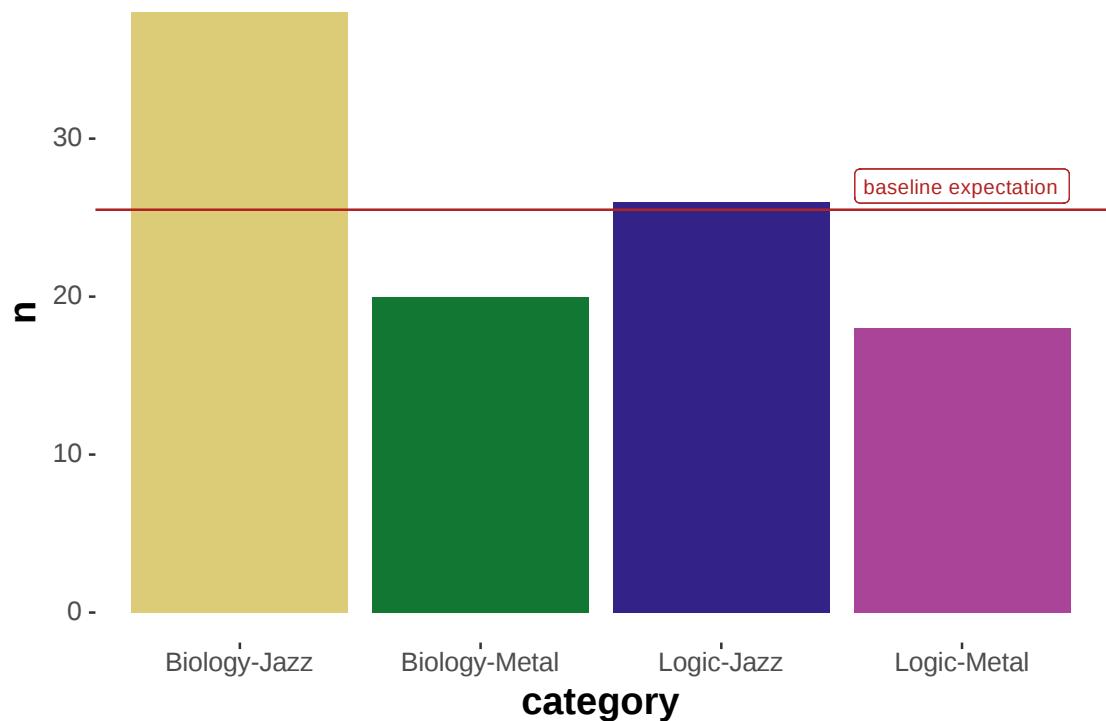


Figure 10.7.: Observed counts of choice pairs of music+subject preference in the BLJM data.

standardized squared deviation of the observed counts from the predicted baseline in each cell. It can be shown that, if the number of observations N is large enough, the sampling distribution of the χ^2 test statistic is approximated well enough by the χ^2 distribution with $k - 1$ degrees of freedom (where k is the number of categories).¹⁴ Notice that the approximation by a χ^2 -distribution hinges on an approximation, which is only met when there are enough samples (just as we needed in the CLT). A rule-of-thumb is that at most 20% of all cells should have expected frequencies below 5 in order for the test to be applicable, i.e., $np_i < 5$ for all i in Figure 10.8.

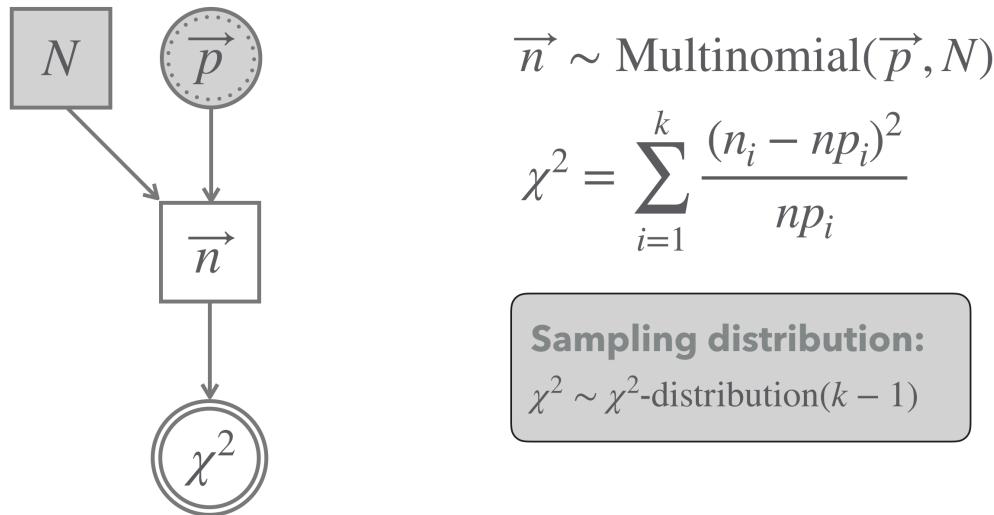


Figure 10.8.: Graphical representation of Pearson's χ^2 test for goodness of fit (testing a vector of predicted proportion).

We can compute the χ^2 -value associated with the observed data $t(D_{obs})$ as follows:

```

# observed counts
n <- counts_BLJM_choice_pairs_vector
# proportion predicted
p <- rep(1/4, 4)
# expected number in each cell
e <- sum(n)*p
# chi-squared for observed data
chi2_observed <- sum((n-e)^2 * 1/e)
chi2_observed

## [1] 9.529412

```

¹⁴A proof of this fact is non-trivial, but an intuition why this might be so is available, if we think of each cell independently first. In each cell, with more and more samples, the distribution of counts will approximate a normal distribution by the CLT. The χ^2 -distribution rests (by construction) on a sum of squared samples from a standard normal distribution.

10. Hypothesis Testing

We can then compare this value to the sampling distribution, which is a χ^2 -distribution with $k - 1 = 3$ degrees of freedom. We compute the p -value associated with our data as the tail of the sampling distribution, as shown also in Figure 10.9:¹⁵

```
p_value_BLJM <- 1 - pchisq(chi2_observed, df = 3)
```

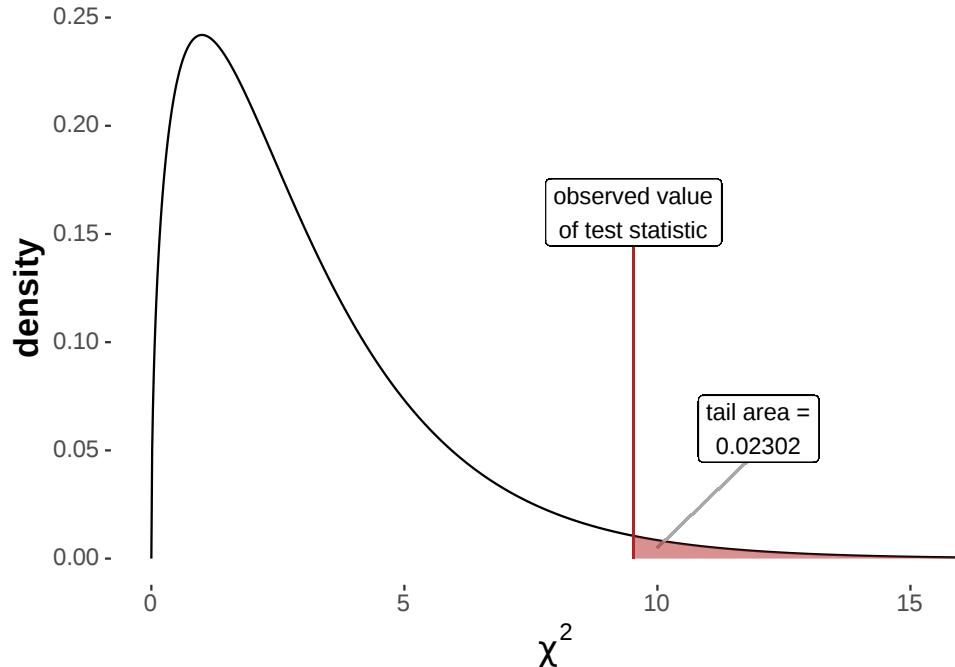


Figure 10.9.: Sampling distribution for a Pearson's χ^2 test of goodness of fit (χ^2 -distribution with $k - 1 = 3$ degrees of freedom), testing a flat baseline null hypothesis based on the BLJM data. .

Of course, these calculations can be performed also by using a built-in R function, namely `chisq.test`:

```
counts_BLJM_choice_pairs_vector <- BLJM_associated_counts %>% pull(n)
chisq.test(counts_BLJM_choice_pairs_vector)

##
## Chi-squared test for given probabilities
##
## data: counts_BLJM_choice_pairs_vector
## X-squared = 9.5294, df = 3, p-value = 0.02302
```

¹⁵Notice that this is a one-sided test due to the nature of the test statistic, which measures squared deviation from the baseline and not deviation in any particular direction (because it is hard to say what a "direction" would be in this case anyway).

The common interpretation of our calculations would be to say that the test yielded a significant result, at least at the significance level of $\alpha = 0.5$. In a research paper we might report this results roughly as follows:

Observed counts deviated significantly from what is expected if each category (here: pair of music+subject choice) was equally likely (χ^2 -test, with $\chi^2 \approx 9.53$, $df = 3$ and $p \approx 0.023$).

Notice that this test is an “omnibus test of difference”. We can conclude from a significant test result that the whole vector of observations is unlikely to have been generated from chance, but we cannot conclude from this result (without anything doing else) why, where or how the observations deviated from the assumed prediction vector. Looking at the plot of the data in Figure 10.7 above, it seems intuitive to think that Metal is disproportionately disfavored and that the combination of Biology and Jazz looks particularly outliery, when compared to baseline expectations.

10.3.1.2. Pearson's χ^2 -test for independence

The previous test of goodness of fit does not allow us to address the lecturer's conjecture that a preference of Metal over Jazz goes with a preference of Logic over Biology. A slightly different kind of χ^2 -test is better suited for this. In Pearson's χ^2 -test of independence, we look at a two-dimensional table of correlated data observations, like this one:

```
BLJM_table <- BLJM_associated_counts %>%
  select(-category) %>%
  pivot_wider(names_from = LB, values_from = n)
BLJM_table

## # A tibble: 2 x 3
##   JM     Biology Logic
##   <chr>    <int> <int>
## 1 Jazz      38    26
## 2 Metal     20    18
```

For easier computation and compatibility with the function `chisq.test` we handle the same data but stored as a matrix:

```
counts_BLJM_choice_pairs_matrix <- matrix(
  counts_BLJM_choice_pairs_vector,
  nrow = 2,
  byrow = T
)
rownames(counts_BLJM_choice_pairs_matrix) <- c("Jazz", "Metal")
colnames(counts_BLJM_choice_pairs_matrix) <- c("Biology", "Logic")
counts_BLJM_choice_pairs_matrix
```

10. Hypothesis Testing

```
##      Biology Logic
## Jazz      38    26
## Metal     20    18
```

Pearson's χ^2 -test of independence addresses the question of whether two-dimensional tabular count data like the above could plausibly have been generated by a prediction vector \vec{p} which results from the assumption that the realizations of row- and column-choices are stochastically independent. If row- and column-choices are independent, the probability of seeing an outcome result in cell ij is the probability of realizing row i times the probability of realizing column j . So, under an independence assumption, we expect a matrix, and a resulting vector of choice proportions like this:

```
# number of observations in total
N <- sum(counts_BLJM_choice_pairs_matrix)
# marginal proportions observed in the data
# the following is the vector r in the model graph
row_prob <- counts_BLJM_choice_pairs_matrix %>% rowSums() / N
# the following is the vector c in the model graph
col_prob <- counts_BLJM_choice_pairs_matrix %>% colSums() / N
# table of expected observation under independence assumption
# NB: %o% is the outer product of vectors
BLJM_expectation_matrix <- (row_prob %o% col_prob) * N
BLJM_expectation_matrix

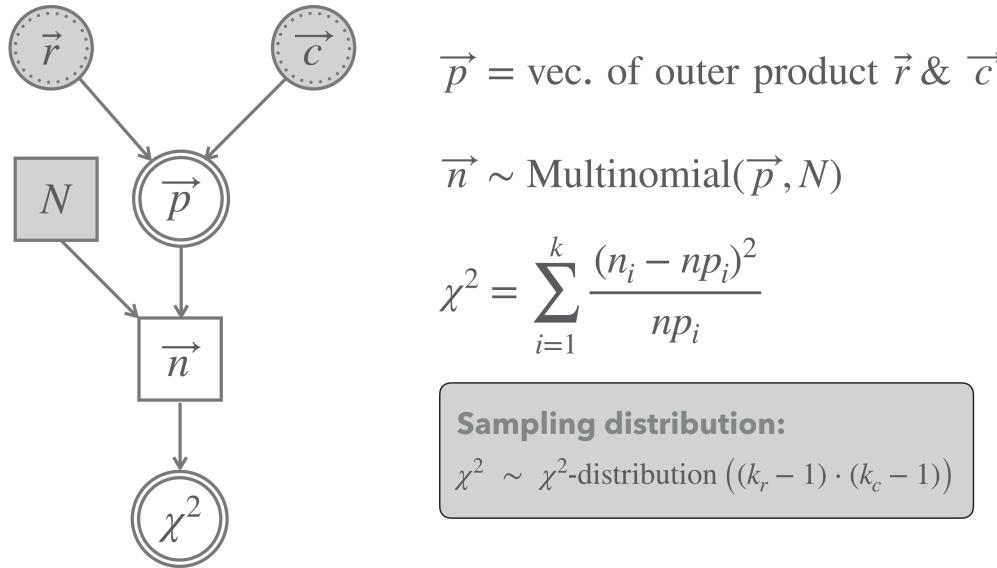
##      Biology Logic
## Jazz  36.39216 27.60784
## Metal 21.60784 16.39216

# the following is the vector p in the model graph
BLJM_expectation_vector <- as.vector(BLJM_expectation_matrix)
BLJM_expectation_vector

## [1] 36.39216 21.60784 27.60784 16.39216
```

Figure 10.10 shows a graphical representation of the χ^2 -test of independence. The main difference to the previous test of goodness of fit is that we do no longer just fix any-old prediciton vector \vec{p} , but consider \vec{p} the deterministic results of independence *and* the best estimates (based on the data at hand) of the row- and column probabilities.

We can compute the observed χ^2 -test statistic as follows:

Figure 10.10.: Graphical representation of Pearson's χ^2 test for independence.

```

chi2_observed <- sum(
  (counts_BLJM_choice_pairs_matrix - BLJM_expectation_matrix)^2 /
  BLJM_expectation_matrix
)
p_value_BLJM <- 1-pchisq(q = chi2_observed, df = 1)
round(p_value_BLJM, 5)

## [1] 0.50615

```

Figure 10.11 shows the sampling distribution, the value of the test statistic for the observed data and the p -value.

We can also use the built-in function `chisq.test` in R to obtain this result more efficiently:

```

chisq.test(
  # supply data as a matrix, not as a vector, for test of independence
  counts_BLJM_choice_pairs_matrix,
  # do not use the default correction (because we didn't introduce it)
  correct = FALSE
)

##
## Pearson's Chi-squared test

```

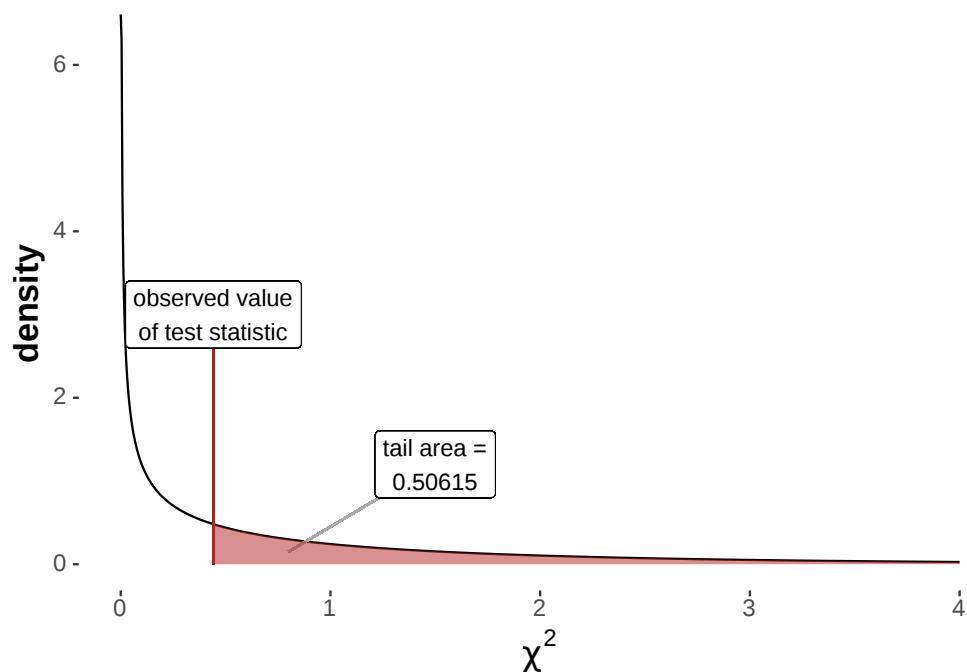


Figure 10.11.: Sampling distribution for a Pearson's χ^2 test of independence (χ^2 -distribution with 1 degree of freedom), testing a flat baseline null hypothesis based on the BLJM data. .

```
##  
## data: counts_BLJM_choice_pairs_matrix  
## X-squared = 0.44202, df = 1, p-value = 0.5061
```

With p -value of about 0.50615 we should conclude that there is no indication of strong evidence *against* the assumption of independence. Consequently, there is no evidence *in favor* of the lecturer's conjecture of dependence of musical and academic preferences. In a research paper we might report this result as follows:

A χ^2 -test of independence did not yield a significant test result (χ^2 -test, with $\chi^2 \approx 0.44$, $df = 1$ and $p \approx 0.5$). Therefore, we cannot claim to have found any evidence for the research hypothesis of dependence.

10.3.2. z-test

The Central Limit Theorem tells us that, given enough data, we can treat means of repeated samples from any arbitrary probability distribution as approximately normally distributed. If we notice in addition that if X and Y are random variables following a normal distribution, then so is $Z = X - Y$ (see also the chapter on the normal distribution), it becomes clear how research questions about means and about differences between means (e.g., in the Mental Chronometry experiment) can be addressed, at least approximately, by using tests that hinge on a sampling distribution which is a normal distribution (usually a standard normal distribution).

The z -test is perhaps the simplest of a family of tests that rely on normality of the sampling distribution. Unfortunately, what makes it so simple is also what makes it inapplicable in a wide range of cases. The z -test assumes that a quantity which is normally distributed has an unknown mean (to be inferred by testing) but it also assumes that the *variance is known*. Since we do not know the variance in most cases of practical relevance, the z -test needs to be replaced by a more adequate test, usually a test from the t -test family, to be discussed below.

We start with the z -test nonetheless because of the added benefit to our understanding. Figure 10.12 shows the model which lies implicitly underneath a z -test for checking whether data \vec{x} , which are assumed to be normally distributed with known σ , could have been generated by a hypothesized mean $\mu = \mu_0$. The sampling distribution of the derived test statistic z is a standard normal distribution.

We know that IQ test results are normally distributed around a mean of 100 with a standard deviation of 15. This holds when the sample is representative of the whole population. But suppose we have reason to believe that the sample is from CogSci students. The standard deviation in a sample from CogSci students might still plausibly be fixed to 15, but we'd like to test whether the assumption that *this* sample was generated by a mean $\mu = 100$, out null hypothesis.

For illustration, suppose we observed this data set of IQ test results:

```
# fictitious IQ-data  
IQ_data <- c(87, 91, 93, 97, 100, 101, 103, 104,
```

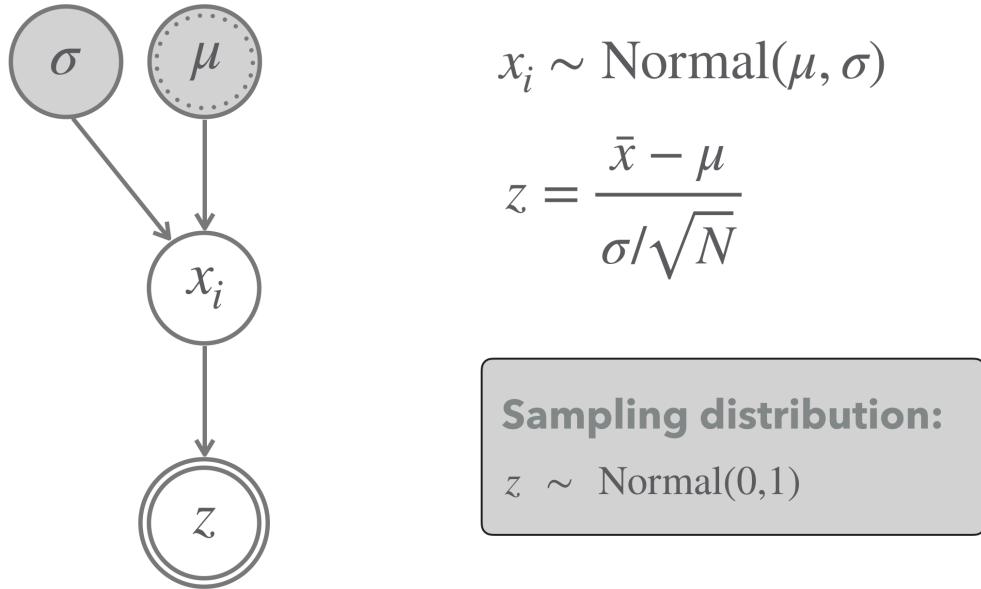


Figure 10.12.: Graphical representation of a z -test.

```
104, 105, 105, 106, 108, 110, 111,
112, 114, 115, 119, 121)
mean(IQ_data)

## [1] 105.3
```

The mean of this data set is 105.3. Suspicious!

Following the model in Figure 10.12 we calculate the value of the test statistic for the observed data.

```
# number of observations
N <- length(IQ_data)
# null hypothesis to test
mu_0 <- 100
# standard deviation (known/assumed as true)
sd <- 15
z_observed <- (mean(IQ_data) - mu_0) / (sd / sqrt(N))
z_observed %>% round(4)

## [1] 1.5802
```

We focus on a one-sided p -value because our “research” hypothesis is that CogSci students have, on average, a higher IQ. Since we observed a mean of 105.3 in the data, which is higher than the critical value

of 100, we test the null hypothesis $\mu = 100$ against an alternative hypothesis which assumes that the data was generated by mean *bigger* than 100 (which is exactly our research hypothesis).

We can then compute the p -value, as before, by checking the area under the sampling distribution, here a standard normal, in the appropriate way. Figure 10.13 shows this result graphically.

```
p_value_IQ_data_ztest <- 1 - pnorm(z_observed)
p_value_IQ_data_ztest %>% round(6)
```

```
## [1] 0.057036
```

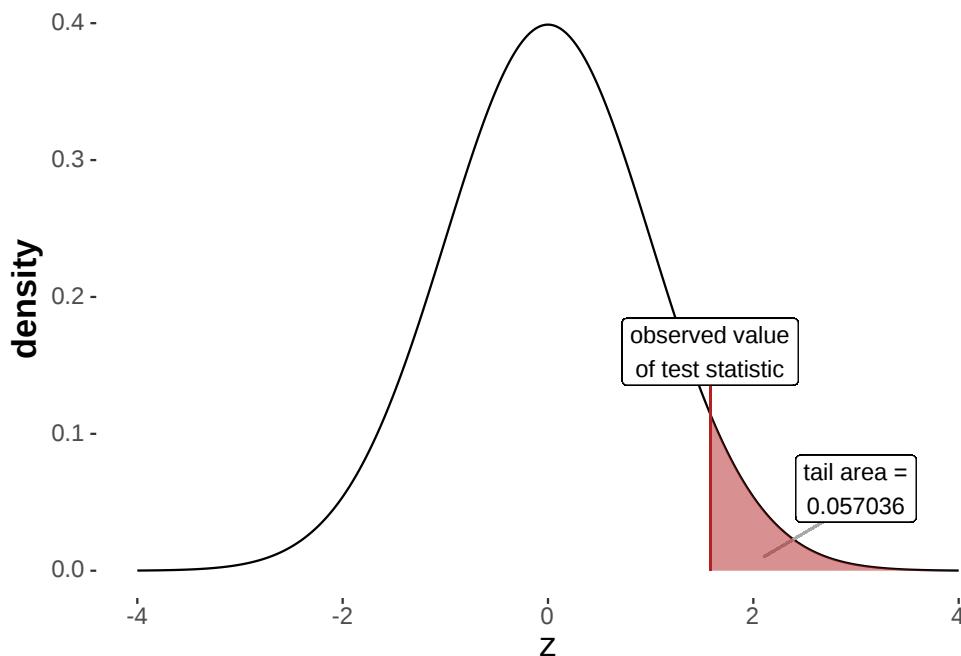


Figure 10.13.: Sampling distribution for a z -test, testing the null hypothesis based that the IQ-data was generated by $\mu = 100$ (with assumed/known σ).

We can also use a ready-made function for the z -test. However, as the z -test is so uncommon, it is not built into core R. We need to rely on the BSDA package to find the function `z.test`.

```
BSDA::z.test(x = IQ_data, mu = 100, sigma.x = 15, alternative = "greater")
##
## One-sample z-Test
##
## data: IQ_data
```

10. Hypothesis Testing

```
## z = 1.5802, p-value = 0.05704
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
## 99.78299      NA
## sample estimates:
## mean of x
##          105.3
```

The conclusion to be drawn from this test could be formulated in a research report as follows:

We tested the null-hypothesis of a mean equal to 100, assuming a known standard deviation of 15, in a one-sided z -test against the alternative hypothesis that the data was generated by a mean greater than 100 (our research hypothesis). The test was not significant ($N = 20$, $z \approx 1.5802$, $p \approx 0.05704$), giving us no indication of strong evidence that the assumption that the mean is at most 100 if false.

10.3.3. t-tests

In most practical applications where a z -test might be useful the standard deviation is not known. If unknown, it should also not lightly be fixed by clever guess-work. This is where the family of t -tests comes in. We look at two examples of these, the one-sample t -test, which compares one set of samples to a fixed mean, and a two-sample t -test, which compares the means of two sets of samples.

10.3.3.1. One-sample t -test

The simplest example of this family, a t -test for one metric vector \vec{x} of normally distributed observations, tests the null hypothesis, just like the z -test, that \vec{x} was generated by some $\mu = \mu_0$. Unlike the z -test, a one-sample t -test does not, however, assume that the standard deviation is known. It rather uses the observed data to obtain an estimate for this parameter. More concretely, a one-sample t -test for \vec{x} estimates the standard deviation in the usual way (see Chapter 5):

$$\hat{\sigma}_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_{\vec{x}})^2}$$

Figure 10.14 shows a graphical representation of a one-sample t -test model. The light shading of the node for the standard deviation is meant to indicate that this parameter is estimated from the observed data. Importantly, the distribution of the test statistic t is no longer well approximated by a normal distribution when the sample size is low. It is better captured by a Student's t distribution.

Let's revisit our IQ-data set from above to calculate a t -test. Using a t -test implies that we are assuming now that the standard deviation is actually unknown. We can calculate the value of the test statistic for the observed data and use this to compute a p -value, much like in the case of the z -test before.

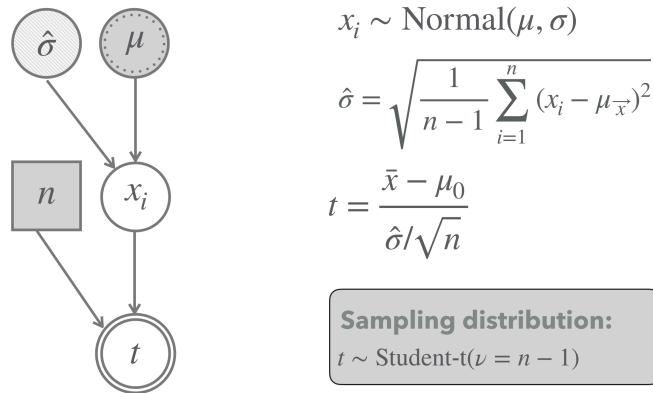


Figure 10.14.: Graphical representation of the model underlying a frequentist one-sample t -test. Notice that the lightly shaded node for the standard deviation represents that the value for this parameter is estimated from the data.

```
N <- length(IQ_data)
# fix the null hypothesis
mean_0 <- 100
# unlike in a z-test we use the sample to estimate SD
sigma_hat <- sd(IQ_data)
t_observed <- (mean(IQ_data) - mean_0) / sigma_hat * sqrt(N)
t_observed %>% round(4)

## [1] 2.6446
```

We calculate the relevant one-sided p -value using the cumulative distribution function `pt` of the t -distribution.

```
p_value_t_test_IQ <- 1 - pt(t_observed, df = N-1)
p_value_t_test_IQ %>% round(6)
```

```
## [1] 0.007992
```

Compare these calculations against the built-in function `t.test`:

```
t.test(x = IQ_data, mu = 100, alternative = "greater")
```

```
##
##  One Sample t-test
##
```

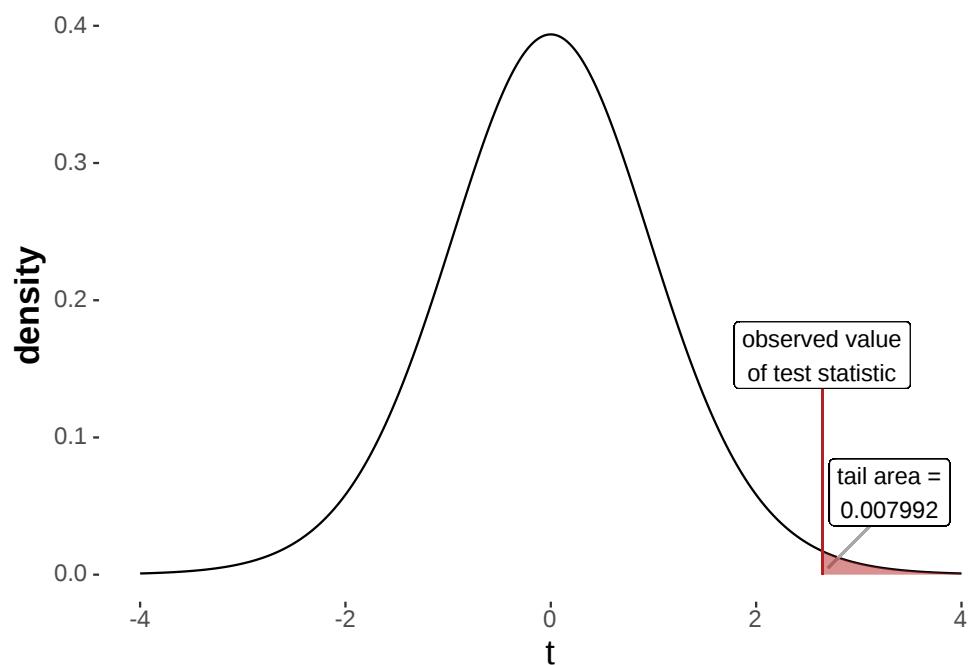


Figure 10.15.: Sampling distribution for a t -test, testing the null hypothesis based that the IQ-data was generated by $\mu = 100$ (with unknown σ).

```
## data: IQ_data
## t = 2.6446, df = 19, p-value = 0.007992
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
## 101.8347      Inf
## sample estimates:
## mean of x
##      105.3
```

These results could be stated in a research report much like so:

We tested the null-hypothesis of a mean equal to 100, assuming an unknown standard deviation, using a one-sided, one-sample t -test against the alternative hypothesis that the data was generated by a mean greater than 100 (our research hypothesis). The significant test result ($N = 20, t \approx 2.6446, p \approx 0.007992$) suggests that the data provides strong evidence against the assumption that the mean is not bigger than 100.

Notice that the conclusions we draw from the previous z -test and this one-sample t -test are quite different. Why is this so? Well, it is because we (cheekily) chose a data set `IQ_data` which was actually *not* generated by a normal distribution with standard deviation of 15, contrary to what we said about IQ-scores normally having this standard deviation. The assumption about σ fed into the z -test was (deliberately!) wrong. The result of the t -test, at least for this example, is better. The data in `IQ_data` are actually samples from $\text{Normal}(105, 10)$. This demonstrates why the one-sample t -test is usually preferred over a z -test: unshakable, true knowledge of σ is very rare.

10.3.3.2. Two-sample t -test (for unpaired data with equal variance and unequal sample sizes)

The “mother of all experimental designs” compares two groups of measurements. We give a drug to one group of patients; a placebo to another. We take a metric measure (say, blood sugar level) and ask whether there is a difference between these two groups. Section 8.5 introduced the T -Test Model for a Bayesian approach. Here, we look at a corresponding model for a frequentist approach, a so-called two-sample t -test. There are different kinds of such two-sample t -tests. The differences lie, e.g., in whether we assume that both groups have equal variance, in whether the sample sizes are the same in both groups, or in whether observations are paired (e.g., as in a within-subjects design, where we get two measurements from each participant, one from each condition / group). Here, we focus on unpaired data (as from a between-subjects design), assume equal variance but (possibly) unequal sample sizes. The case we look at is the avocado data, and we want to specifically investigate whether the weekly average price of organically grown avocados is higher than that of conventionally grown avocados.¹⁶

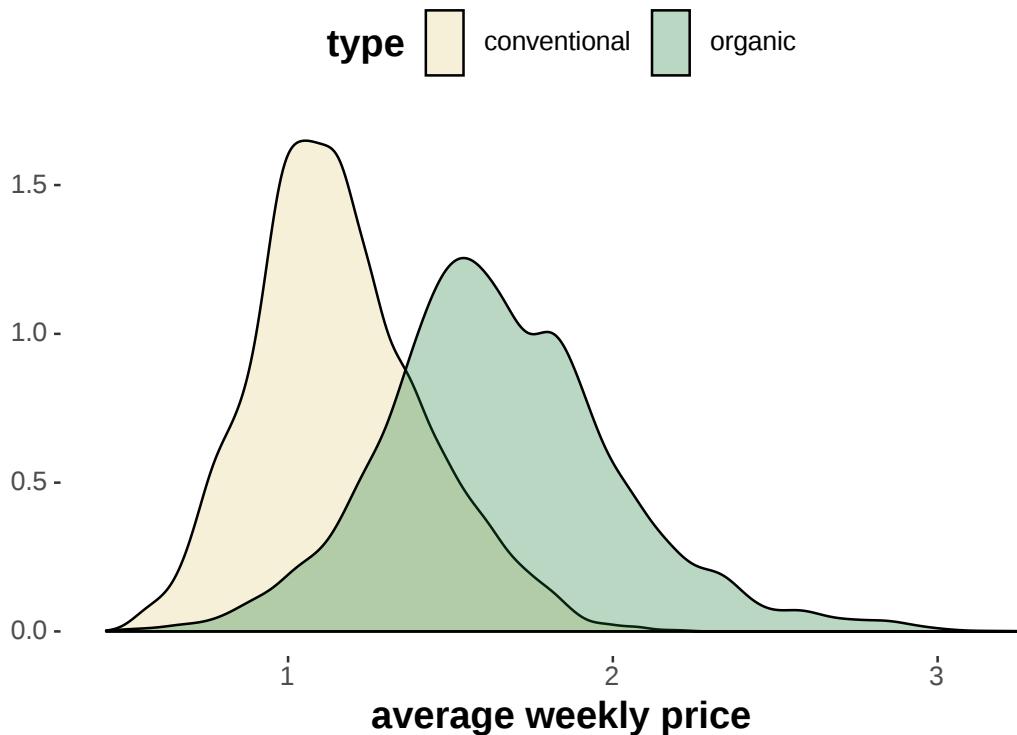
The data to consider is this:

¹⁶Notice that the original avocado data set contains information also about the place of measurement, which would in principle allow us to treat the price measurements as paired samples (one pair for each week and place). For simplicity, but with a note of care that this makes us lose possibly relevant structural information, we here treat the avocado data as if it contained unpaired samples.

10. Hypothesis Testing

```
avocado_data <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intr
# remove currently irrelevant columns
select( -X1 , - contains("Bags") , - year , - region) %>%
# rename variables of interest for convenience
rename(
  total_volume_sold = `Total Volume` ,
  average_price = `AveragePrice` ,
  small = '4046' ,
  medium = '4225' ,
  large = '4770' ,
)
)
```

Remember that the distribution of prices looks as follows:



A graphical representation of the two-sample t -test (for unpaired data with equal variance and unequal sample sizes), which we will apply to this case, is shown in Figure 10.16. The model assumes that we have two vectors of metric measurements \vec{x}_A and \vec{x}_B , with length n_A and n_B respectively. These are the price measures for conventionally grown and for organically grown avocados. The model assumes that measures in both \vec{x}_A and \vec{x}_B are iid samples from a normal distribution. The mean of one group (group B in the graph) is assumed to be some unknown μ . Interestingly, this parameter will cancel out eventually:

the approximation of the sampling distribution turns out to be independent of this parameter.¹⁷ The mean of the other group (group A in the graph) is computed as $\mu + \delta$, so with some additive parameter δ indicating the difference between means of these groups. This δ is the main parameter of interest for inferences regarding hypotheses concerning differences between the groups. Finally, the model assumes that both groups have the same standard deviation, an estimate of which is derived from the data (in a rather convoluted looking formula that is not important for our introductory concerns). As indicated in Figure 10.16, the sampling distribution for this model is an instance of Student's t -distribution with mean 0, standard deviation 1 and degrees of freedom ν given as $n_A + n_B - 2$.

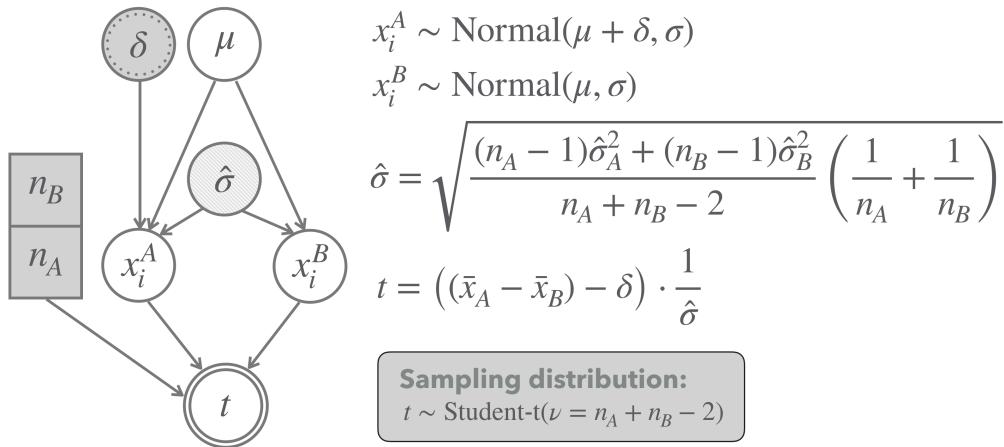


Figure 10.16.: Graphical representation of the model underlying a frequentist two-population t -test (for unpaired data with equal variance and unequal sample sizes). Notice that the light shading of the node for the standard deviation indicates that the values for this parameter is estimated from the data.

Figure 10.16 gives us the template to compute the value of the test statistic for the observed data.

```
# fix the null hypothesis: no difference between groups
delta_0 <- 0
# data (group A)
x_A <- avocado_data %>%
  filter(type == "organic") %>% pull(average_price)
# data (group B)
x_B <- avocado_data %>%
  filter(type == "conventional") %>% pull(average_price)
# sample mean for organic (group A)
mu_A <- mean(x_A)
# sample mean for conventional (group B)
mu_B <- mean(x_B)
# numbers of observations
```

¹⁷This is intuitively so because the test statistic is concerned only with the difference between sample means.

10. Hypothesis Testing

```
n_A <- length(x_A)
n_B <- length(x_B)
# variance estimate
sigma_AB <- sqrt(
  ((n_A -1) * sd(x_A)^2 + (n_B -1) * sd(x_B)^2) /
  (n_A + n_B -2) ) * (1/n_A + 1/n_B)
)
t_observed <- (mu_A - mu_B - delta_0) / sigma_AB
t_observed

## [1] 105.5878
```

We can use the value of the test statistic for the observed data to compute a one-sided p -value, as before.

Notice that we use a one-sided test because we hypothesize that organically grown avocados are more expensive, not just that they have a different price (more expensive or cheaper).

```
p_value_t_test_avocado <- 1 - pt(q = t_observed, df = n_A + n_B - 1)
p_value_t_test_avocado
```

```
## [1] 0
```

Owing to number imprecision, the calculated p -value comes up as a flat zero. We have a lot of data and the task of defending that conventionally grown avocados are not less expensive than organically grown is very tough. This also shows in the corresponding picture in Figure 10.17.

We can also, of course, calculate this test result with the built-in function `t.test`:

```
t.test(
  x = x_A,           # first vector of data measurements
  y = x_B,           # sec vector of data measurements
  paired = FALSE,    # measurements are to be treated as unpaired
  var.equal = TRUE,  # we assum equal variance in both groups
  mu = 0            # NH is delta = 0 (name 'mu' is misleading!)
)

##
## Two Sample t-test
##
## data: x_A and x_B
## t = 105.59, df = 18247, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

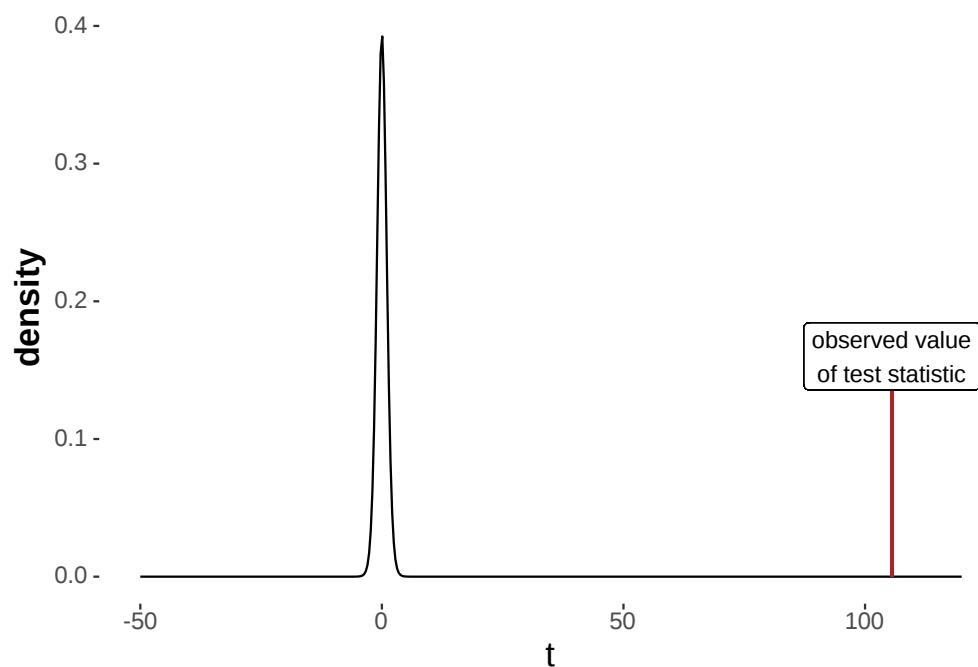


Figure 10.17.: Sampling distribution for a two-sample t -test, testing the null hypothesis of no difference between groups, based that the avocado data.

10. Hypothesis Testing

```
## 0.4867522 0.5051658
## sample estimates:
## mean of x mean of y
## 1.653999 1.158040
```

The result could be reported as follows:

We used a two-sample t -test of differences of means (unpaired samples, equal variance, unequal sample sizes) to compare the average weekly price of conventionally-grown avocados to that of organically grown avocados. The test result indicates a significant difference for the null hypothesis that conventionally-grown avocados are not cheaper ($N_A = 9123$, $N_B = 9126$, $t \approx 105.587848$, $p \approx 0$).

10.3.4. ANOVA

We have k groups of metric observations. For group $1 \leq j \leq k$, there are n_j observations. Let x_{ij} be observation $1 \leq i \leq n_j$ for group $1 \leq j \leq k$. Let $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$ be the mean of group j and let $\bar{x} = \frac{1}{k} \sum_{j=1}^k \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$ be the grand mean of all data points. We would like to show that the total sum of squares can be decomposed into two summands: the within-group sum of squares and the between-group sum of squares:

$$\underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2}_{\text{Total SS}} = \underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}_{\text{Within-Group SS}} + \underbrace{\sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2}_{\text{Between-Group SS}}$$

To show this, we first establish a lemma, which will also be useful later:

Lemma 10.1 (Sum of squares cancellation). *Let \vec{x} be a vector of n real-valued numbers, and let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ be its mean. The sum of squares around the mean is zero:*

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Proof.

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}) &= \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{j=1}^n x_j) \quad [\text{by def. of mean}] \\ &= \sum_{i=1}^n x_i - \frac{n}{n} \sum_{j=1}^n x_j \quad [\text{second summand independent of } i] \\ &= 0 \end{aligned}$$

□

Proposition 10.1 (Sum of squares decomposition (ANOVA)). *If x_{ij} is observation $1 \leq i \leq n_j$ for group $1 \leq j \leq k$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n_j} x_{ij}$ the mean of group j and $\bar{\bar{x}} = \frac{1}{k} \sum_{j=1}^k \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$ be the grand mean of all data points, then:*

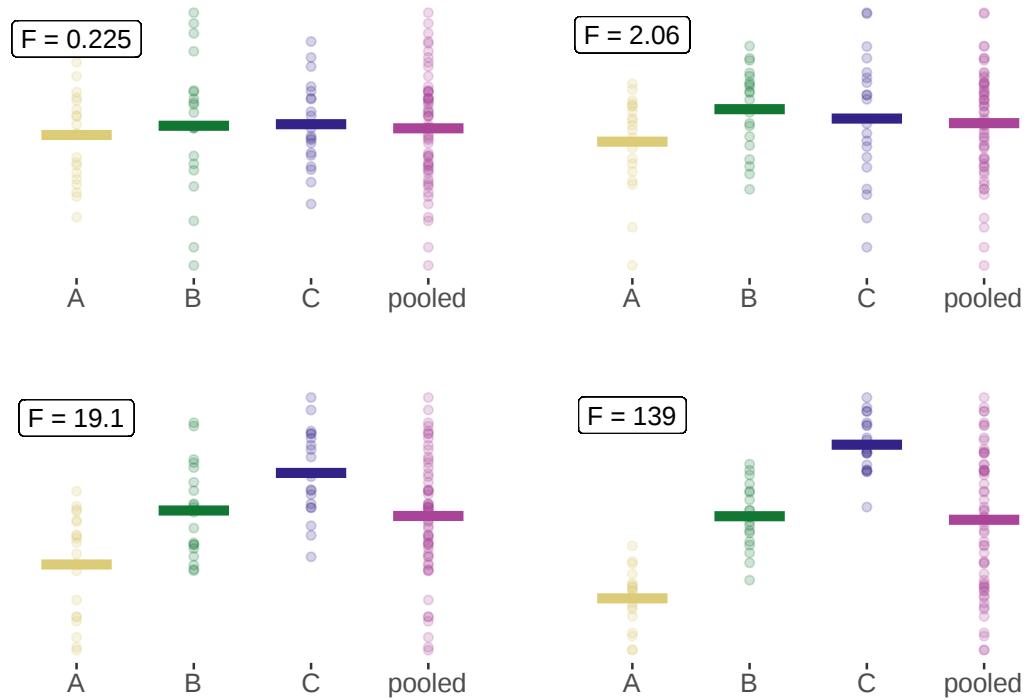
$$\underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2}_{\text{Total SS}} = \underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}_{\text{Within-Group SS}} + \underbrace{\sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2}_{\text{Between-Group SS}}$$

Proof.

$$\begin{aligned}
 & \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \\
 &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j + \bar{x}_j - \bar{\bar{x}})^2 && [-\bar{x}_j + \bar{x}_j = 0] \\
 &= \sum_{j=1}^k \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_j)^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}})^2] && [\text{binomial theorem}] \\
 &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{\bar{x}}) && [\text{rearranging}] \\
 &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2 + 2 \sum_{j=1}^k (\bar{x}_j - \bar{\bar{x}}) \underbrace{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}_{=0 \text{ by lemma}} && [\text{independences}] \\
 &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2 && [\text{by lemma}]
 \end{aligned}$$

□

10. Hypothesis Testing

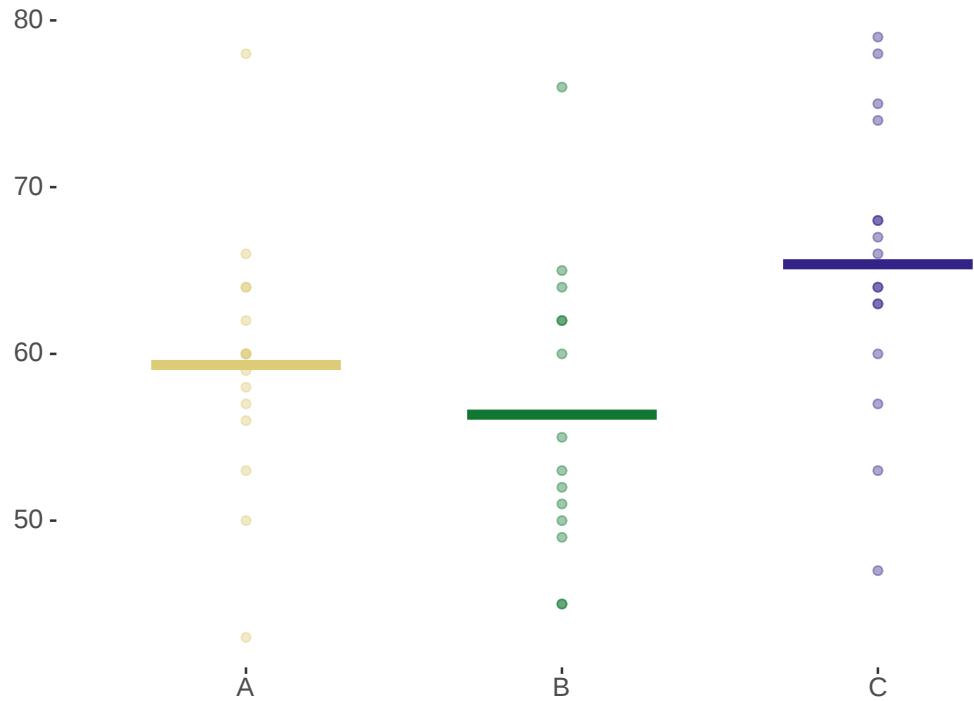


```
# fictitious data
x_A <- c(78, 43, 60, 60, 60, 50, 57, 58, 64, 64, 56, 62, 66, 53, 59)
x_B <- c(52, 53, 51, 49, 64, 60, 45, 50, 55, 65, 76, 62, 62, 45)
x_C <- c(78, 66, 74, 57, 75, 64, 64, 53, 63, 60, 79, 68, 68, 47, 63, 67)
# number of observations in each group
n_A <- length(x_A)
n_B <- length(x_B)
n_C <- length(x_C)
# in tibble form
anova_data <- tibble(
  condition = c(
    rep("A", n_A),
    rep("B", n_B),
    rep("C", n_C)
  ),
  value = c(x_A, x_B, x_C)
)

anova_data %>%
  ggplot(
    aes(x = condition, y = value, color = condition)
```

```
) +
geom_point(
  fill = "lightgray",
  size = 1.5,
  alpha = 0.4
) +
guides(color = "none") +
geom_segment(size = 2,
  aes(
    x = condition_number - 0.3,
    xend = condition_number + 0.3,
    y = condition_mean,
    yend = condition_mean
),
  data = anova_data %>% group_by(condition) %>%
  summarise(condition_mean = mean(value)) %>%
  mutate(condition_number = 1:3)
) +
labs(
  x = "",
  y = ""
)
```

10. Hypothesis Testing



```

grand_mean <- anova_data %>% pull(value) %>% mean()
df1 <- 2
df2 <- n_A + n_B + n_C -3

# between-group sum-of-squares
between_group_variance <- 1/df1 *
(
  n_A * (mean(x_A) - grand_mean)^2 +
  n_B * (mean(x_B) - grand_mean)^2 +
  n_C * (mean(x_C) - grand_mean)^2
)

# within-group sum-of-squares
within_group_variance <- 1/df2 *
(
  sum((x_A - mean(x_A))^2) +
  sum((x_B - mean(x_B))^2) +
  sum((x_C - mean(x_C))^2)
)
# test statistic of observed data
F_observed <- between_group_variance / within_group_variance

```

```

p_value_anova <- 1 - pf(F_observed, 2, n_A + n_B + n_C -3)
p_value_anova %>% round(4)

## [1] 0.0172

aov(formula = value ~ condition, anova_data) %>% summary()

##          Df Sum Sq Mean Sq F value Pr(>F)
## condition    2   640.8   320.4   4.485 0.0172 *
## Residuals   42  3000.3    71.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Based on a one-way ANOVA, we find evidence against the assumption of equal means across all groups ($F(2, 42) \approx 4.485, p \approx 0.0172$.)

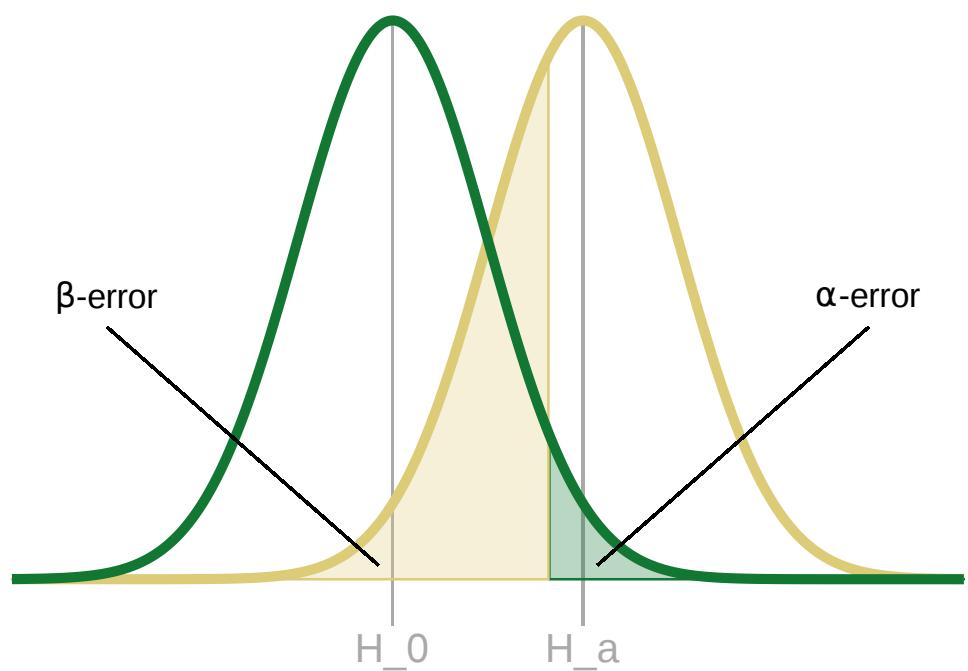
10.4. Three approaches

10.4.1. Fisher

Frequentist hypothesis testing is superficially similar to Popperian falsificationism. It is, however, quite the opposite when looked at more carefully. Popper famously denied that empirical observation could constitute positive evidence in favor of a research hypothesis. Research hypotheses can only be refuted, viz., when their logically consequences are logically incompatible with the observed data. In a Popperian science, what is refuted are research hypotheses; frequentist statistics instead seeks to refute null-hypotheses and counts successful refutation of a null-hypothesis as evidence in favor of a research hypothesis.

10.4.2. Neyman-Pearson

A conventional threshold on p -values may govern a categorical decision whether to reject or not reject the null-hypothesis (in the Neyman-Pearson approach). This threshold is essentially an upper-bound on a particular kind of error, namely the error to falsely reject the null-hypothesis when it is in fact true (a so-called type-1 error or α -error).



10.4.3. Hybrid modern NHST

10.5. Relation to model checking Section

11. Model Comparison

Parameter estimation asks: given a single model and the data, what are good (e.g., credible) values of the model's parameters? Hypothesis testing applies this logic: given a model and fixing (for the sake of construction) a certain parameter value, is it plausible that the data was generated by the assumed model + parameter setting? Finally, model comparison (the topic of this chapter) asks: based on the data at hand, which of several models is better? Or even: *how much* better is this model compared to this, given the data?

Frequentist and Bayesian approaches agree that the pivotal criterion by which to compare models is how well a model explains the observed data. A good explanation of observed data D is one that makes D unsurprising. Intuitively, we long for an explanation for things that puzzle us. A good explanation is a way of looking at the world in which puzzle disappear, in which all observations make sense, in which what we have seen would have been quite expectable after all. Consequently, the pivotal quantity for comparing models is how likely D is given a model M_i : $P(D \mid M_i)$.

But there is more to a good explanation, also intuitively. All else equal, a good explanation is simple. If theories A and B both explain the facts equally well, but A does so with less "mental machinery", most people would choose the more economical explanation A .

In this chapter, we will look at three common methods of comparing models. Two are frequentist (Akaike information criterion & likelihood-ratio test). The third is Bayesian (Bayes factor). There are more approaches and methods (e.g., K -fold cross-validation, other varieties of information criteria, ANOVA-based model comparisons). Our goal is not to be exhaustive, but to introduce the main ideas of model comparison and showcase a reasonable selection of alternative approaches.

The learning goals for this chapter are:

- understand the differences between estimation, testing and model comparison
- understand the idea behind and become able to apply the covered methods:
 - Akaike information criterion
 - likelihood-ratio test
 - Bayes factor
- become familiar with pro's and con's of each of these methods

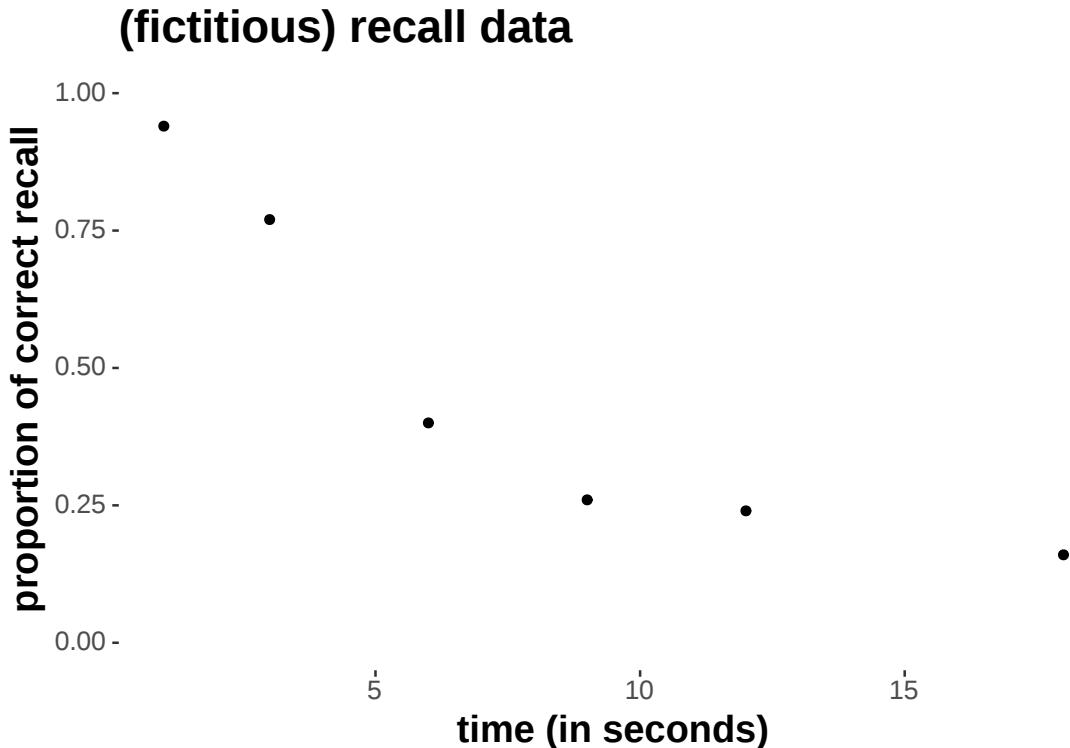
11.1. Case study: recall models

As a running example for this chapter, we borrow from Myung (2003) and consider a fictitious data set of recall rates and two models to explain this data.

As for data, for each time point (in seconds) $t \in \{1, 3, 6, 9, 12, 18\}$, we have 100 (binary) observations of whether a previously memorized item was recalled correctly.

```
# time after memorization (in seconds)
t <- c(1, 3, 6, 9, 12, 18)
# proportion (out of 100) of correct recall
y <- c(.94, .77, .40, .26, .24, .16)
# number of observed correct recalls (out of 100)
obs <- y * 100
```

A visual representation of this data set is here:



We are interested in comparing two theoretically different models for this data. Models differ in their assumption about the functional relationship between recall probability and time. The **exponential model** assumes that the recall probability θ_t at time t is an exponential decay function with parameters a and b :

$$\theta_t(a, b) = a \exp(-bt), \quad \text{where } a, b > 0$$

The resulting (frequentist) exponential model, assuming a Binomial likelihood function of the recall data is therefore:

$$P(D = \langle k, N \rangle \mid \langle a, b \rangle) = \text{Binom}(k, N, a \exp(-bt)), \quad \text{where } a, b > 0$$

In contrast, the **power model** assumes that the relationship is that of a power function:

$$\theta_t(c, d) = ct^{-d}, \quad \text{where } c, d > 0$$

The resulting (frequentist) power model, assuming a Binomial likelihood function of the recall data is therefore:

$$P(D = \langle k, N \rangle \mid \langle c, d \rangle) = \text{Binom}(k, N, c t^{-d}), \quad \text{where } c, d > 0$$

These models therefore make different (parameterized) predictions about the time course of forgetting/recall. Figure 11.1 shows predictions of each model for θ_t for different parameter values:

The research question of relevance is: which of these two models is a better model for the observed data?

11.2. Akaike Information Criterion

A wide-spread approach to model comparison used in the frequentist paradigm is to use the **Akaike information criterion (AIC)**. The AIC is the most common instance of a class of measures for model comparison known as *information criteria*, which all draw on information-theoretic notions to compare how good each model is.

If M_i is a frequentist model, specified by likelihood function $P(D \mid \theta_i, M_i)$, with k model parameters in parameter vector θ_i , and if D_{obs} is the observed data, then the AIC score of model M_i given D_{obs} is defined as:

$$\text{AIC}(M_i, D_{\text{obs}}) = 2k - 2 \log P(D_{\text{obs}} \mid \hat{\theta}_i, M_i)$$

Here, $\hat{\theta}_i = \arg \max_{\theta_i} P(D_{\text{obs}} \mid \theta_i, M_i)$ is the best-fitting parameter vector, i.e., the maximum likelihood estimate (MLE), and k is the number of free parameters in model M_i .

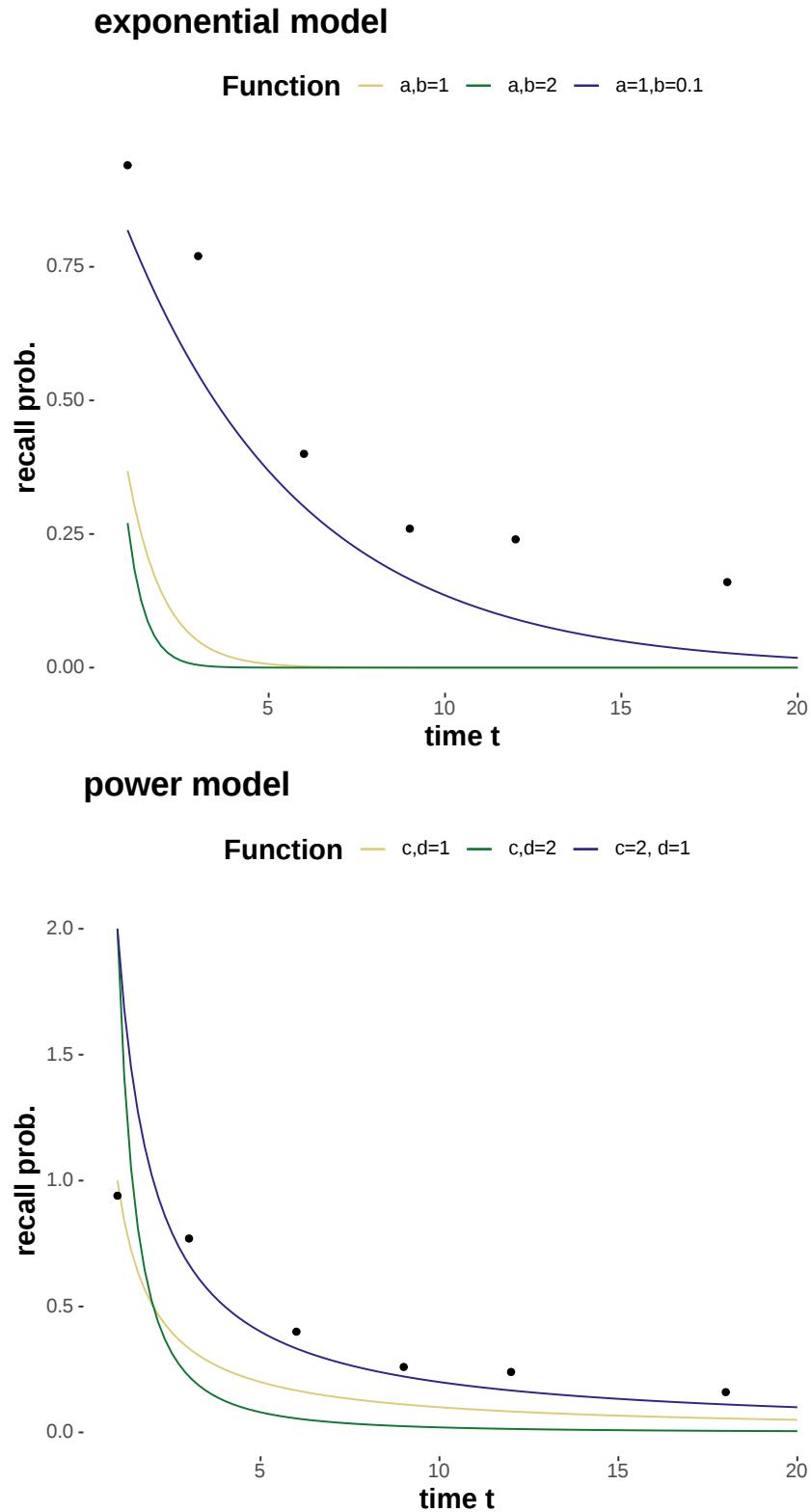


Figure 11.1.: Examples of predictions of the exponential and the power model of forgetting for different values of each model's parameters.

The lower an AIC score the better the model (in comparison to other models for the same data D_{obs}). All else equal, the higher the number of free parameters k the worse the model's AIC score. The first summand in the definition above can, therefore, be conceived of as a measure of **model complexity**. As for the second summand, think of $-\log P(D_{\text{obs}} \mid \hat{\theta}_i)$ as a measure of (information-theoretic) surprisal: how surprising is the observed data D_{obs} from the point of view of model M under the most favorable circumstances (that is, the MLE of θ_i). The higher the probability $P(D_{\text{obs}} \mid \hat{\theta}_i)$, the better the model's AIC score, all else equal.

To apply AIC-based model comparison to the recall models, we first need to compute the MLE of each model (see Chapter 9.2). Here are functions that return the negative log-likelihood of each model, for any (suitable) pair of parameter values:

```
# generic neg-log-LH function (covers both models)
nLL_generic <- function(par, model_name) {
  w1 <- par[1]
  w2 <- par[2]
  # make sure parameters are in acceptable range
  if (w1 < 0 | w2 < 0 | w1 > 20 | w2 > 20) {
    return(NA)
  }
  # calculate predicted recall rates for given parameters
  if (model_name == "exponential") {
    theta <- w1*exp(-w2*t) # exponential model
  } else {
    theta <- w1*t^(-w2)      # power model
  }
  # avoid edge cases of infinite log-likelihood
  theta[theta <= 0.0] <- 1.0e-4
  theta[theta >= 1.0] <- 1-1.0e-4
  # return negative log-likelihood of data
  - sum(dbinom(x = obs, prob = theta, size = 100, log = T))
}
# negative log likelihood of exponential model
nLL_exp <- function(par) {nLL_generic(par, "exponential")}
# negative log likelihood of power model
nLL_pow <- function(par) {nLL_generic(par, "power")}
```

These functions are then optimized with built-in function `optim`. The results are shown in the table below.

```
# getting the best fitting values
bestExpo <- optim(nLL_exp, par = c(1,0.5))
bestPow <- optim(nLL_pow, par = c(0.5,0.2))
MLEstimates = data.frame(model = rep(c("exponential", "power"), each = 2),
                          parameter = c("a", "b", "c", "d"),
```

11. Model Comparison

```
value = c(bestExpo$par, bestPow$par))
knitr::kable(MLEstimates)
```

model	parameter	value
exponential	a	1.0701722
exponential	b	0.1308151
power	c	0.9531330
power	d	0.4979154

The MLE-predictions of each model are shown in Figure 11.2 below, alongside the observed data.

MLE fits

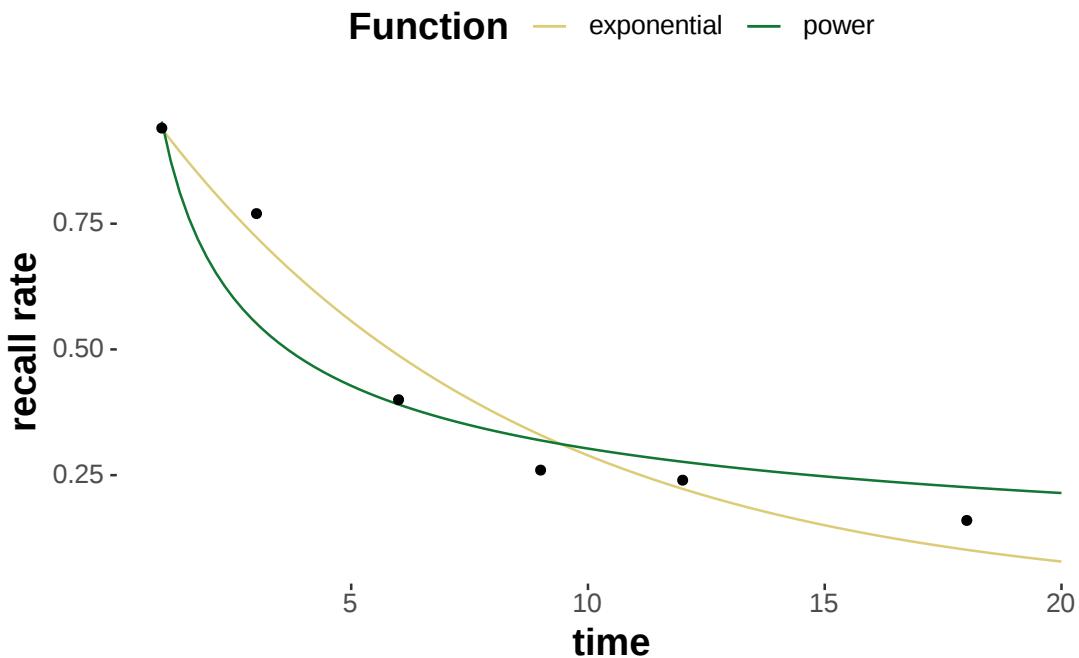


Figure 11.2.: Predictions of the exponential and the power model under best-fitting parameter values.

It is impossible to say by visual inspection of Figure 11.2 which model should be preferred. Numbers might help see more fine-grained differences:

```
predExp <- expo(t,a,b)
predPow <- power(t,c,d)
modelStats <- tibble(
  model = c("expo", "power"),
  `log likelihood` = round(c(-bestExpo$value, -bestPow$value), 3),
  probability = signif(exp(c(-bestExpo$value, -bestPow$value)), 3),
```

```

# sum of squared errors
SS = round(c( sum((predExp-y)^2), sum((predPow-y)^2)),3)
)
modelStats

## # A tibble: 2 x 4
##   model `log likelihood` probability     SS
##   <chr>          <dbl>        <dbl> <dbl>
## 1 expo            -18.7    7.82e- 9  0.019
## 2 power           -26.7    2.47e-12 0.057

```

The exponential model has a higher log-likelihood, a higher probability, and a lower sum of squares. This suggests that the exponential model is better.

The AIC-score of these models is a direct function of the negative log-likelihood. Since both models have the same number of parameters, we also arrive at the same verdict: based on comparison of AIC-scores, the exponential model is the better model.

```

get_AIC <- function(optim_fit) {
  2 * length(optim_fit$par) + 2 * optim_fit$value
}
AIC_scores <- tibble(
  AIC_exponential = get_AIC(bestExpo),
  AIC_power = get_AIC(bestPow)
)
AIC_scores

## # A tibble: 1 x 2
##   AIC_exponential AIC_power
##             <dbl>      <dbl>
## 1              41.3      57.5

```

How should we interpret the difference in AIC-scores? Some suggest that differences in AIC-scores larger than 10 should be treated as implying that the weaker model has practically no empirical support (Burnham and Anderson 2002). Adopting such a criterion, we would therefore favor the exponential model based on the data observed.

But we could also try to walk a more nuanced, a more quantitative road. Indeed, we can look at relative AIC-scores in terms of so-called **Akaike weights** (Wagenmakers and Farrell 2004; Burnham and Anderson 2002). If models M_1, \dots, M_n are on the table and $\text{AIC}(M_i, D)$ is the AIC-score of model M_i for observed data D , then the Akaike weight of model M_i is defined as:

$$w_{\text{AIC}}(M_i, D) = \frac{\exp(-0.5 * \Delta_{\text{AIC}}(M_i, D))}{\sum_{j=1}^k \exp(-0.5 * \Delta_{\text{AIC}}(M_j, D))} \quad \text{where}$$

$$\Delta_{\text{AIC}}(M_i, D) = \text{AIC}(M_i, D) - \min_j \text{AIC}(M_j, D)$$

Akaike weights are relative and normalized measures, and may serve as an approximate measure of a model's posterior probability given the data:

$$P(M_i | D) \approx w_{\text{AIC}}(M_i, D)$$

For the running example at hand, this would mean that we could conclude that the posterior probability of the exponential model is approximately:

```
delta_AIC_power <- AIC_scores$AIC_power - AIC_scores$AIC_exponential
delta_AIC_exponential <- 0
Akaike_weight_exponential <- exp(-0.5 * delta_AIC_exponential) /
  (exp(-0.5 * delta_AIC_exponential) + exp(-0.5 * delta_AIC_power))
Akaike_weight_exponential

## [1] 0.9996841
```

We would conclude from this approximate quantitative assessment that the evidence in favor of the exponential model is rather strong.

Our approximation is better, the more data we have. We will see a method below, the Bayesian method, which computes $P(M_i | D)$ in a non-approximate way.

11.3. Likelihood-Ratio Test

The likelihood-ratio (LR) test is a very popular frequentist method of model comparison. The LR-test assimilates model comparison to frequentist hypothesis testing. It defines a suitable test statistic and supplies an approximation of the sampling distribution. The LR-test first and foremost applies to the comparison of **nested models**, but there are results about how approximate results can be obtained when comparing non-nested models with an LR-test (Vuong 1989).

A frequentist model M_i is **nested** inside another frequentist model M_j iff M_i can be obtained from M_j by fixing at least one of M_j 's free parameters to a specific value. If M_i is nested under M_j , M_i is called the **nested model**, and M_j is called the **nesting model** or the **encompassing model**. Obviously, the nested model is simpler (of lower complexity) than the nesting model.

For example, we had the two-parameter exponential model previously:

$$P(D = \langle k, N \rangle \mid \langle a, b \rangle) = \text{Binom}(k, N, a \exp(-bt)), \quad \text{where } a, b > 0$$

An example of a model that is nested under this two-parameter model is the following one-parameter model, which fixes $a = 1.1$.

$$P(D = \langle k, N \rangle \mid b) = \text{Binom}(k, N, 1.1 \exp(-bt)), \quad \text{where } b > 0$$

Here's an ML-estimation for the nested model:

```
nLL_expo_nested <- function(b) {
  # calculate predicted recall rates for given parameters
  theta <- 1.1*exp(-b*t) # one-param exponential model
  # avoid edge cases of infinite log-likelihood
  theta[theta <= 0.0] <- 1.0e-4
  theta[theta >= 1.0] <- 1-1.0e-4
  # return negative log-likelihood of data
  - sum(dbinom(x = obs, prob = theta, size = 100, log = T))
}

bestExpo_nested <- optim(
  nLL_expo_nested,
  par = 0.5,
  method = "Brent",
  lower = 0,
  upper = 20
)
bestExpo_nested

## $par
## [1] 0.1372445
##
## $value
## [1] 19.21569
##
## $counts
## function gradient
##       NA       NA
##
```

11. Model Comparison

```
## $convergence
## [1] 0
##
## $message
## NULL
```

The LR-test looks at the likelihood ratio of the nested model M_0 over the encompassing model M_1 as its test statistic:

$$\text{LR}(M_1, M_0) = -2 \log \left(\frac{P_{M_0}(D_{\text{obs}} | \hat{\theta}_0)}{P_{M_1}(D_{\text{obs}} | \hat{\theta}_1)} \right)$$

We can calculate the value of this test statistic for the current example as follows:

```
LR_observed <- 2 * bestExpo_nested$value - 2 * bestExpo$value
LR_observed

## [1] 1.098429
```

If the simpler (nested) model is true, the sampling distribution of this test statistic approximates a χ^2 -distribution with d if we have more and more data. The degrees of freedom d is given by the difference in free parameters, i.e., the number of parameters the nested model fixes to specific values, but which are free in the nesting model.

We can therefore calculate the p -value for the LR-test for our current example like so:

```
p_value_LR_test <- 1 - pchisq(LR_observed, 1)
p_value_LR_test

## [1] 0.2946111
```

The p -value of this test quantifies the evidence against the assumption that the data was generated by the simpler model. A significant test result would therefore indicate that it would be surprising if the data was generated by the simpler model. This is standardly taken as evidence in favor of the more complex, nesting model. For the current p -value $p \approx 0.2946$, there is no strong evidence against the simpler model and we would therefore rather favor the nested model, due to its simplicity; the data at hand does not seem to warrant the added complexity of the nesting model; the nested model seems to suffice.

11.4. Bayes factors

Bayes factors are the flagship Bayesian measure of comparing models. Since Bayesians do not hesitate to assign probabilities to models, parameters and hypotheses, we would ideally want to know the *absolute probability* of M_i given the data: $P(M_i | D)$. Unfortunately, to calculate this (by Bayes rule) we would need to normalize by quantifying over *all* models. Alternatively, we look at the relative probability of a small selection of models, or, even more economic, compare only two models, preferably in terms of their odds, as follows.

Take two Bayesian models:

- M_1 has prior $P(\theta_1 | M_1)$ and likelihood $P(D | \theta_1, M_1)$
- M_2 has prior $P(\theta_2 | M_2)$ and likelihood $P(D | \theta_2, M_2)$

Using Bayes rule, we compute the posterior odds of models (given the data) as the product of the likelihood ratio and the prior odds.

$$\underbrace{\frac{P(M_1 | D)}{P(M_2 | D)}}_{\text{posterior odds}} = \underbrace{\frac{P(D | M_1)}{P(D | M_2)}}_{\text{Bayes factor}} \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}}$$

The likelihood ratio is also called the **Bayes factor**. Formally, the Bayes factor is the factor by which a rational agent changes her prior odds in the light of observed data to arrive at updated, posterior odds. More intuitively, the Bayes factor quantifies the strength of evidence given by the data about the models of interest. It expresses this evidence in terms of the models' relative prior predictive accuracy. To see the latter, let's expand the Bayes factor as what it actually is: the ratio of marginal likelihoods.

$$\frac{P(D | M_1)}{P(D | M_2)} = \frac{\int P(\theta_1 | M_1) P(D | \theta_1, M_1) d\theta_1}{\int P(\theta_2 | M_2) P(D | \theta_2, M_2) d\theta_2}$$

Three insights are to be gained from this expansion. Firstly, the Bayes factor is a measure of how well each model would have predicted the data *ex ante*, i.e., before having seen any data. In this way, it is diametrically opposed to a concept like AIC or the LR-test, which both relied on models' maximum likelihood fits (therefore *using the data*, so being *ex post*).

Secondly, the marginal likelihood of a model is exactly the quantity which we identified (in the context of parameter estimation) as being very hard to compute, especially for complex models. The fact that marginal likelihoods are hard to compute was the reason that we looked briefly at MCMC sampling, for example. It follows that Bayes factors are very tough nuts to crack. But, as we will see, there are very clever approaches to computing Bayes factors after all.

Thirdly, Bayes factor model comparison implicitly (and quite vigorously) punishes model complexity, but in a more sophisticated manner than just counting free parameters. To appreciate this intuitively, imagine a

11. Model Comparison

model with a large parameter set a very diffuse prior that spreads its probability over a wide range of parameter values. Since Bayes factors are computed based on *ex ante* predictions, a diffuse model is punished for its imprecision of prior predictions, because we integrate over all parameters (weighted by priors) and their associated likelihood.

As for notation, we write:

$$BF_{12} = \frac{P(D | M_1)}{P(D | M_2)}$$

for the Bayes factor in favor of model M_1 over model M_2 . This quantity can take on positive values, which are often translated into natural language as follows:

BF_{12}	interpretation
1	irrelevant data
1 - 3	hardly worth ink or breath
3 - 6	anecdotal
6 - 10	now we're talking: substantial
10 - 30	strong
30 - 100	very strong
100 +	decisive (bye, bye M_2 !)

As $BF_{12} = BF_{21}^{-1}$ it suffices to give this translation into natural language only for values ≥ 1 .

There are at least two general approaches to calculating or approximating Bayes factors, paired here with a (non-exhaustive) list of instantiations, only some of which will be dealt with here

1. get each model's marginal likelihood

- grid approximation
- by Monte Carlo sampling
- brute force clever math
- bridge sampling

2. get Bayes factor directly

- Savage-Dickey method
- using encompassing priors
- transdimensional MCMC
- supermodels

In the following, we will concern ourselves with grid approximation and MC sampling, based on the running example of memory recall as a function of time.

11.4.1. Grid approximation

Grid approximation for a model's marginal likelihood works for relatively small models with, say, no more than 4-5 free parameters. Grid approximation considers discrete values for each parameter, evenly spaced over the whole range of plausible parameter values, thereby approximating the integral in the definition of marginal likelihoods.

To begin with, we need to define a prior over parameters to obtain Bayesian versions of the exponential and power model of forgetting. Here, we assume flat priors over a reasonable range of parameter values for simplicity. For the exponential model, we choose:

$$\begin{aligned} P(D = \langle k_i, N \rangle \mid \langle a, b \rangle, M_{\text{exp}}) &= \text{Binom}(k, N, a \exp(-bt_i)) \\ P(a \mid M_{\text{exp}}) &= \text{Uniform}(a, 0, 1.5) \\ P(b \mid M_{\text{exp}}) &= \text{Uniform}(b, 0, 1.5) \end{aligned}$$

The (Bayesian) power model is then:

$$\begin{aligned} P(D = \langle k_i, N \rangle \mid \langle c, d \rangle, M_{\text{pow}}) &= \text{Binom}(k, N, c t_i^{-d}) \\ P(c \mid M_{\text{pow}}) &= \text{Uniform}(c, 0, 1.5) \\ P(d \mid M_{\text{pow}}) &= \text{Uniform}(d, 0, 1.5) \end{aligned}$$

We can also express these models in code, like so:

```
# prior exponential model
priorExp <- function(a, b){
  dunif(a, 0, 1.5) * dunif(b, 0, 1.5)
}

# likelihood function exponential model
lhExp <- function(a, b){
  theta <- a*exp(-b*t)
  theta[theta <= 0.0] <- 1.0e-5
  theta[theta >= 1.0] <- 1-1.0e-5
  prod(dbinom(x = obs, prob = theta, size = 100))
}

# prior power model
priorPow <- function(c, d){
  dunif(c, 0, 1.5) * dunif(d, 0, 1.5)
}

# likelihood function power model
lhPow <- function(c, d){}
```

11. Model Comparison

```
theta <- c*t^(-d)
theta[theta <= 0.0] <- 1.0e-5
theta[theta >= 1.0] <- 1-1.0e-5
prod(dbinom(x = obs, prob = theta, size = 100))
}
```

To approximate each model's marginal likelihood via grid approximation, we consider equally spaced values for both parameters (a tightly knit grid), asses the prior and likelihood for each parameter pair and finally take the sum over all of the visited values:

```
# make sure the functions accept vector input
lhExp <- Vectorize(lhExp)
lhPow <- Vectorize(lhPow)

# define the step size of the grid
stepsize <- 0.01
# calculate the "evidence" aka marginal likelihood
evidence <- expand.grid(
  x = seq(0.005, 1.495, by = stepsize),
  y = seq(0.005, 1.495, by = stepsize)
) %>%
  mutate(
    lhExp = lhExp(x,y), priExp = 1 / length(x), # uniform priors!
    lhPow = lhPow(x,y), priPow = 1 / length(x)
  )
# output result
message(
  "BF in favor of exponential model: ",
  with(evidence, sum(priExp*lhExp)/ sum(priPow*lhPow)) %>% round(2)
)
```

Based on this computation, we would be entitled to conclude that the data provides overwhelming evidence in favor of the exponential model. No matter what we believed at the outset about which model is more likely, we should adjust these beliefs by a factor of more than 1000 in favor of the exponential model.

11.4.2. Naive Monte Carlo

For simple models (with maybe 4-5 free parameters), we can also use naive Monte Carlo sampling to approximate Bayes factors. In particular, we can approximate the marginal likelihood by taking samples from the prior, calculating the likelihood of the data for each sampled parameter tuple, and then averaging over all calculated likelihoods:

$$P(D, M_i) = \int P(D | \theta, M_i) P(\theta | M_i) d\theta \approx \frac{1}{n} \sum_{\theta_j \sim P(\theta | M_i)}^n P(D | \theta_j, M_i)$$

Here is a calculationg using one million samples from the prior of each model:

```
nSamples <- 1000000
a <- runif(nSamples, 0, 1.5)
b <- runif(nSamples, 0, 1.5)
lhExpVec <- lhExp(a,b)
lhPowVec <- lhPow(a,b)
message(
  "BF in favor of exponential model: ",
  signif(sum(lhExpVec) / sum(lhPowVec)),6
)
```

We can also check the time course of our MC-estimate by a plot like that in Figure 11.3.

```
BFVec <- map_dbl(
  # start at 10.000 and then inspect every 500 samples
  seq(10000, nSamples, by = 500),
  function(i){
    # what's the BF-estimate at that point in time?
    sum(lhExpVec[1:i]) / sum(lhPowVec[1:i])
  }
)

tibble(
  i = seq(10000, nSamples, by = 500),
  BF = BFVec
) %>%
ggplot(aes(x = i, y = BF)) +
  geom_line() +
  geom_hline(
    yintercept = 1221,
    color = "firebrick"
  ) +
  xlab("number of samples")
```

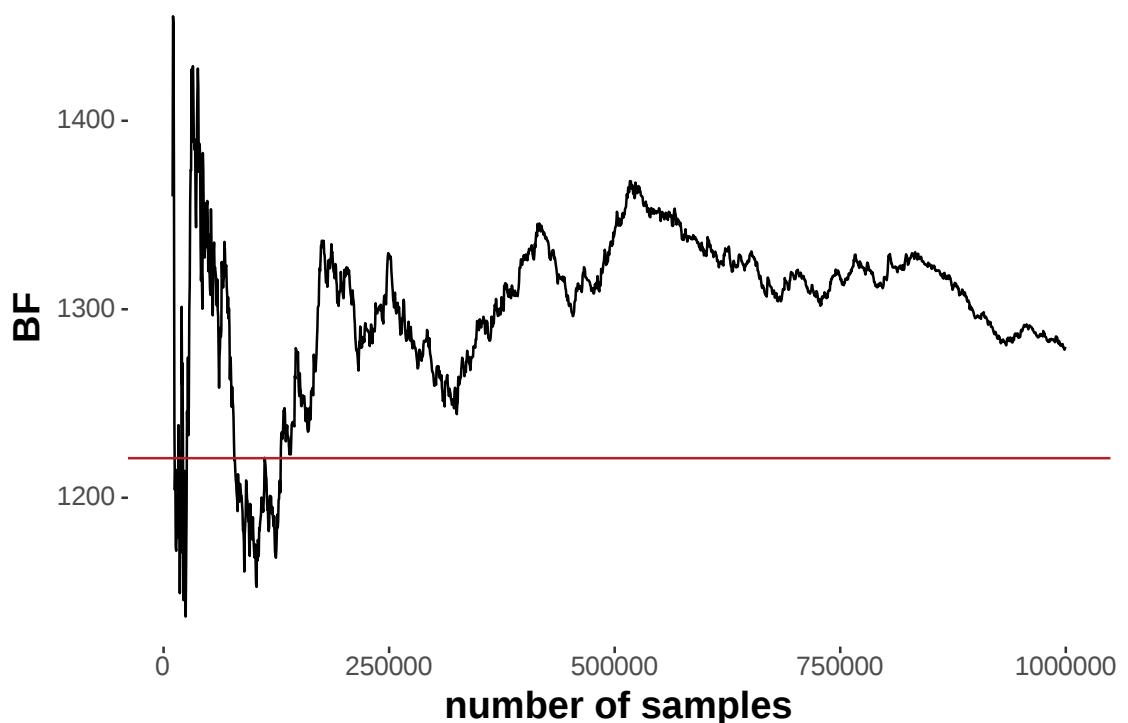


Figure 11.3.: Temporal development (as more samples come in) of the Monte Carlo estimate of the Bayes factor in favor of the exponential model over the power model of forgetting.

11.5. Outlook

For more complex models (e.g., high-dimensional / hierarchical parameter spaces), naive Monte Carlo methods can be highly inefficient. If random sampling of parameter values from the priors is unlikely to deliver values for which the likelihood of the data is reasonably high, most naive MC samples will contribute very little information to the overall estimate of the marginal likelihood. For this reason, there are better sampling-based procedures which preferentially sample *a posteriori* credible parameter values (given the data) and use clever math to compensate for using the wrong distribution to sample from. This is the main idea behind approaches like importance sampling. A very promising approach is in particular **bridge sampling**, which also has its own R package (but we will not be able to deal with this in this course) (Gronau et al. 2017).

For many prominent models (e.g., Bayesian *t*-tests, ANOVA, linear regression), it is possible to calculate Bayes factors analytically if the right kind of priors are specified [Rouder et al. (2009); RouderMorey2012:Default-Bayes-F;@GronauLy2019:Informed-Bayesi].

12. Bayesian hypothesis testing

The goal of this chapter is to compare previously covered frequentist testing of hypotheses to Bayesian methods of testing hypotheses.

We consider two types of hypotheses here. The first is the case where we are interested in testing a precise point-value of a parameter of interest, like $\theta = \theta^*$. The second is that using an interval-region around the parameter of interest. Suppose that instead of addressing the point-valued hypothesis $\theta = \theta^*$, we are able (e.g., through prior research or *a priori* conceptual considerations) to specify a reasonable *region of practical equivalence* (=ROPE) around the parameter value of interest. We could then address what we may call the **ROPE-d hypothesis** $\theta \in [\theta^* - \epsilon ; \theta^* + \epsilon]$, or $\theta = \theta^* \pm \epsilon$ for short.

We consider two types of Bayesian approaches to testing either point-valued or ROPE-d hypotheses of the kind introduced above:

- estimation-based: inspect posterior distribution of the parameter of interest
- comparison-based: compare a model that assumes $\theta = \theta^*$ or $\theta = \theta^* \pm \epsilon$ to model that does not

The learning goals for this chapter are:

- understand the logic of different Bayesian approaches to hypothesis testing
 - qualitative vs quantitative information
 - support for/against null model
- be able to apply these approaches to (simple) case studies
- become aware of the difference with frequentist testing
- understand and be able to apply the Savage-Dickey method (and its extension)

12.1. Data and models for this chapter

12.1.1. 24/7

We will use the same (old) example of binomial data, $k = 7$ heads out of $N = 24$ flips. We are interested in whether the coin is fair, so we address the point-value $\theta = 0.5$. We consider as a ROPE around this value a margin of $\epsilon = 0.01$.

We will use the standard binomial model, just as before, with a flat Beta prior.

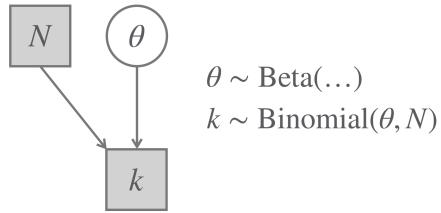


Figure 12.1.: The Binomial Model (repeated from before).

12.1.2. Eco-sensitivity (fictitious)

Here are two sets of metric measurements. We assume that these are (metric) measurements from a test on **environmental awareness** (whatever that means). The measures are taken from two groups of people.

One after watching a political speech by Angela Merkel (Group A), the other after watching a political speech by Donald Trump (Group B). We are interested in whether our experimental manipulation (the type of video) has an effect on the eco-sensitivity measure.

```
x_A <- c(
  104, 105, 100, 91, 105, 118, 164, 168, 111, 107, 136, 149, 104, 114, 107, 95,
  83, 114, 171, 176, 117, 107, 108, 107, 119, 126, 105, 119, 107, 131
)
x_B <- c(
  133, 115, 84, 79, 127, 103, 109, 128, 127, 107, 94, 95, 90, 118, 124, 108,
  87, 111, 96, 89, 106, 121, 99, 86, 115, 136, 114
)
```

We are interested in the question of whether the mean eco-sensitivity is different across groups. Based on descriptive statistics, this might be the case:

```
c(
  mean_A = mean(x_A),
  mean_C = mean(x_B)
)

##   mean_A   mean_C
## 118.9333 107.4444
```

Notice that we have different numbers of measures in each group and that we do not have strong reasons to assume that these groups have the same variance. A Bayesian model for inferences about the likely difference in mean of eco-sensitivity between groups is shown in Figure 12.2. Notice that this model uses priors that are data-aware. This is **not** generally a good choice; especially not if you have prior knowledge to bring to bear on the situation. We do this here, to obtain priors that make fitting (with `greta` maximally painless), but stress that ideally, if prior knowledge exists, priors of the model should encode it.

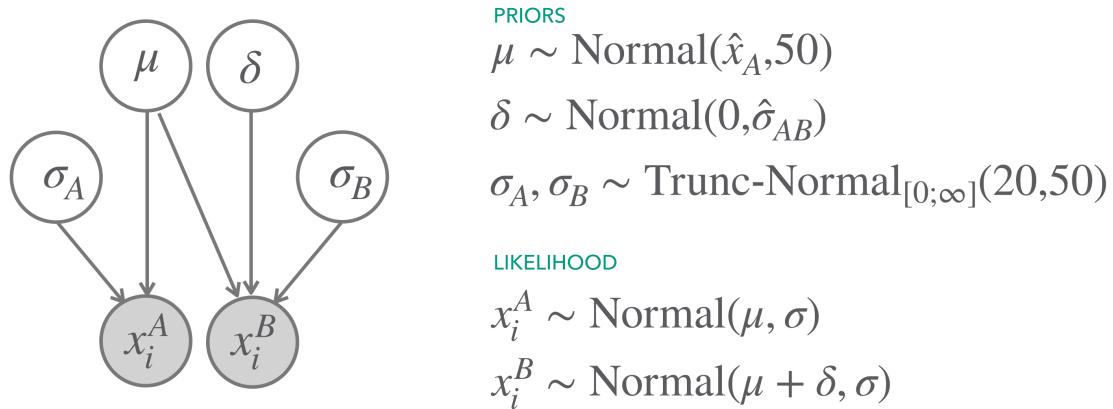


Figure 12.2.: Bayesian T-Test model for inferences about the difference in means in the eco-sensitivity data.

Here is the model from Figure 12.2 implemented in `greta`:

```
# data as greta array
y0 <- as_data(x_A)
y1 <- as_data(x_B)
# priors (regularizing (data-informed) for smooth fitting)
sd_delta <- sd(c(x_A,x_B))
mean_0    <- normal(mean(x_A), 10)
delta      <- normal(0, sd_delta)
sigma_0   <- normal(sd(x_A), 10, truncation = c(0, Inf))
sigma_1   <- normal(sd(x_B), 10, truncation = c(0, Inf))
mean_1    <- mean_0 + delta
# likelihood
distribution(y0) <- normal(mean_0, sigma_0)
distribution(y1) <- normal(mean_1, sigma_1)
# model
m <- model(delta)
```

Our research question is whether there is a difference in mean between groups. We therefore focus on the point-value $\delta = 0$. We consider a ROPE around this value with $\epsilon = 2$ (completely arbitrary choice, since the data is made-up anyway).

12.2. Testing as posterior estimation

12.2.1. Example: 24/7

The following repeats code and calculations from Chapter 9. We can calculate the 95% HDI as follows (via conjugacy of Beta and Binomial)

12. Bayesian hypothesis testing

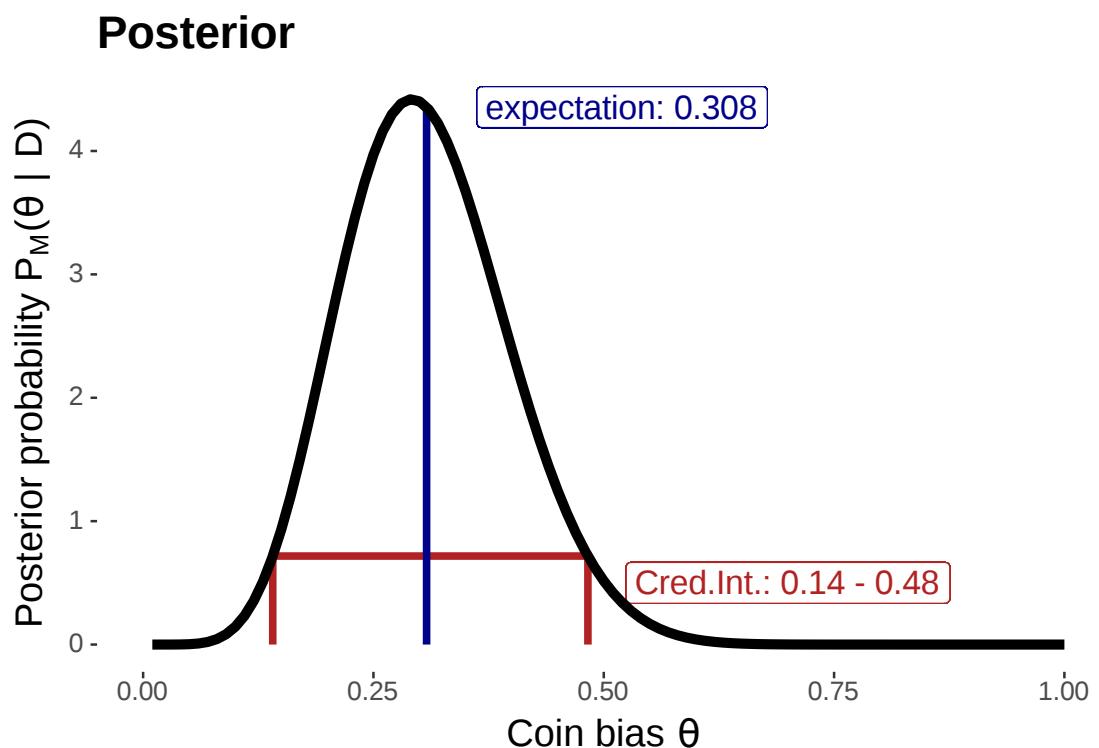
```

estimates_24_7 <- tibble(
  `lower_Bayes` = HDInterval::hdi(function(x) qbeta(x, 8,18))[1],
  `upper_Bayes` = HDInterval::hdi(function(x) qbeta(x, 8,18))[2],
) %>%
  pivot_longer(
    everything(),
    names_pattern = "(.*)_.*",
    names_to = c(".value", "approach")
  )
estimates_24_7

## # A tibble: 1 x 3
##   approach lower upper
##   <chr>     <dbl> <dbl>
## 1 Bayes     0.141  0.483

```

Here is a plot of the posterior.



Using Lindley's approach we notice that the point-value of interested $\theta = 0.5$ is excluded by the 95% HDI and so reject the idea that the coin is fair as sufficiently unlikely to act as if it was false. Using the

ROPE-approach of Kruschke, we notice that our ROPE of $\theta = 0.5 \pm 0.01$ is also fully outside of the 95% HDI. We therefore conclude that the idea that the coin is fair is sufficiently unlikely to act as if it was false.

12.2.2. Example: Eco-sensitivity

We use `greta` to draw samples from the posterior distribution.

```
# sampling
draws_t_test_2 <- greta::mcmc(m, n_samples = 4000)
# cast results (type 'mcmc.list') into tidy tibble
tidy_draws_tt2 = ggmcmc::ggs(draws_t_test_2)
```

We then check the 95% HDI of the posterior.

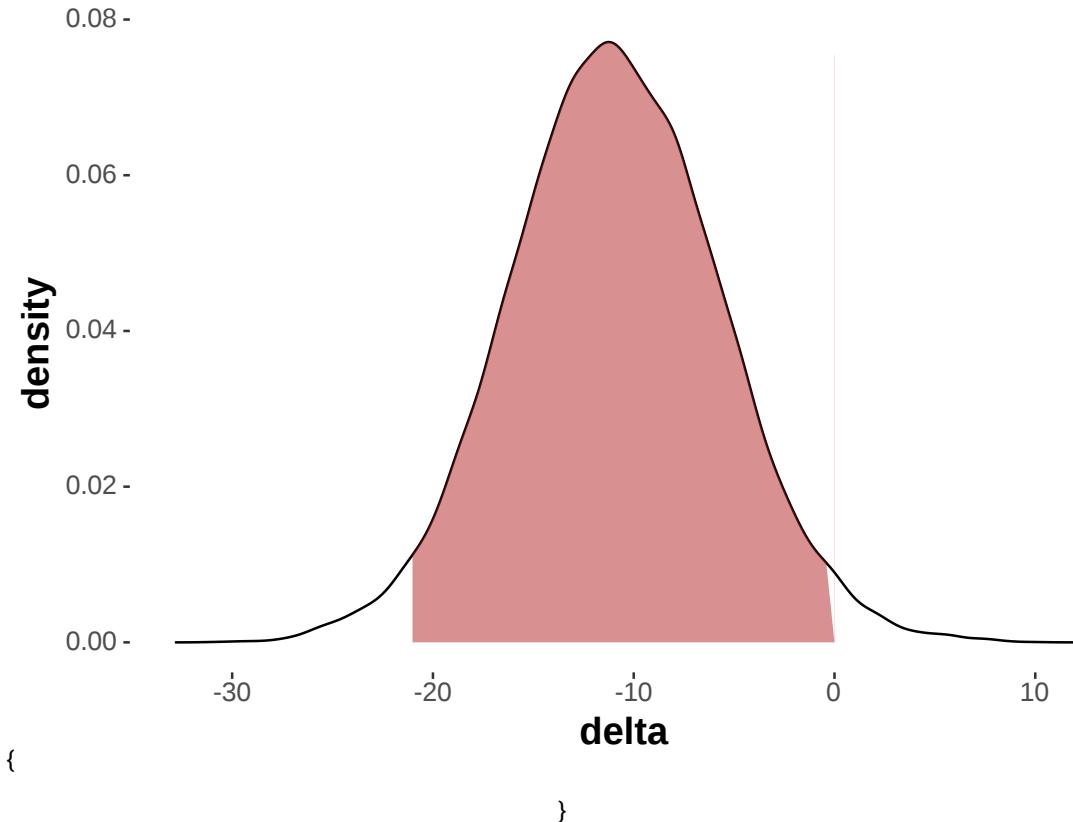
```
# get means and 95% HDI
Bayes_estimates_eco <- tidy_draws_tt2 %>%
  group_by(Parameter) %>%
  summarise(
    mean = mean(value),
    '|95%' = HDInterval::hdi(value)[1],
    '95|%' = HDInterval::hdi(value)[2]
  )
Bayes_estimates_eco
```

```
## # A tibble: 1 x 4
##   Parameter  mean `|95%` `95|%`
##   <fct>     <dbl>  <dbl>  <dbl>
## 1 delta      -10.9   -21.0  -0.384
```

Figure 12.2.2 also shows the posterior distribution and the 95% HDI (in red).

\begin{figure}

12. Bayesian hypothesis testing



\caption{Posterior density of δ parameter in Bayesian t -test model for (fictitious) eco-sensitivity data with the 95% HDI (in red).} \end{figure}

Using Lindley's approach, we'd conclude from the fact that the critical value of $\delta = 0$ is outside the 95% HDI that the idea of group-mean equality is sufficiently unlikely to be practically dismissed. The ROPE $\delta = 0 \pm 2$, however, is neither fully included, nor fully outside of the 95% HDI. Using Krushke's approach, we therefore withhold judgement.

12.3. The Savage-Dickey method

The Savage-Dickey method is a very convenient way of computing Bayes factors for nested models, especially when models only differ with respect to one parameter.

12.3.1. Nested (Bayesian) models

Suppose that there are n continuous parameters of interest $\theta = \langle \theta_1, \dots, \theta_n \rangle$. M_1 is a (Bayesian) model defined by $P(\theta | M_1)$ & $P(D | \theta, M_1)$. Then M_0 is properly nested under M_1 if:

- M_0 assigns fixed values to parameters $\theta_i = x_i, \dots, \theta_n = x_n$
- $P(D | \theta_1, \dots, \theta_{i-1}, M_0) = P(D | \theta_1, \dots, \theta_{i-1}, \theta_i = x_i, \dots, \theta_n = x_n, M_1)$

- $\lim_{\theta_i \rightarrow x_i, \dots, \theta_n \rightarrow x_n} P(\theta_1, \dots, \theta_{i-1} \mid \theta_i, \dots, \theta_n, M_1) = P(\theta_1, \dots, \theta_{i-1} \mid M_0)$

Notice that the last condition is satisfied in particular when M_1 's prior over $\theta_1, \dots, \theta_{i-1}$ is independent of the values for the remaining parameters.

12.3.2. Savage-Dickey theorem

Theorem 12.1 (Savage-Dickey Bayes factors for nested models). *Let M_0 be properly nested under M_1 s.t. M_0 fixes $\theta_i = x_i, \dots, \theta_n = x_n$. The Bayes factor BF_{01} in favor of M_0 over M_1 is then given by the ratio of posterior probability to prior probability of the parameters $\theta_i = x_i, \dots, \theta_n = x_n$ from the point of view of the nesting model M_1 :*

$$BF_{01} = \frac{P(\theta_i = x_i, \dots, \theta_n = x_n \mid D, M_1)}{P(\theta_i = x_i, \dots, \theta_n = x_n \mid M_1)}$$

Proof. Let's assume that M_0 has parameters $\theta = \langle \phi, \psi \rangle$ with $\phi = \phi_0$, and that M_1 has parameters $\theta = \langle \phi, \psi \rangle$ with ϕ free to vary. If M_0 is properly nested under M_1 , we know that $\lim_{\phi \rightarrow \phi_0} P(\psi \mid \phi, M_1) = P(\psi \mid M_0)$. We can then rewrite the marginal likelihood under M_0 as follows:

$$\begin{aligned} P(D \mid M_0) &= \int P(D \mid \psi, M_0) P(\psi \mid M_0) d\psi && [\text{marginalization}] \\ &= \int P(D \mid \psi, \phi = \phi_0, M_1) P(\psi \mid \phi = \phi_0, M_1) d\psi && [\text{assumption of nesting}] \\ &= P(D \mid \phi = \phi_0, M_1) && [\text{marginalization}] \\ &= \frac{P(\phi = \phi_0 \mid D, M_1) P(D \mid M_1)}{P(\phi = \phi_0 \mid M_1)} && [\text{Bayes rule}] \end{aligned}$$

The result follows if we divide by $P(D \mid M_1)$ on both sides of the equation. \square

12.3.3. Example: 24/7

Here is an example, based on the 24/7 data. For a nesting model with a flat prior ($\theta \sim^{M_1} \text{Beta}(1, 1)$), and a point hypothesis $\theta^* = 0.5$, we calculate:

```
# point-value of interest
theta_star <- 0.5
# posterior probability in nesting model
posterior_theta_star <- dbeta(theta_star, 8, 18)
# prior probability in nesting model
prior_theta_star <- dbeta(theta_star, 1, 1)
# Bayes factor (using Savage Dickey)
```

12. Bayesian hypothesis testing

```
BF_01 <- posterior_theta_star / prior_theta_star
BF_01
```

```
## [1] 0.5157351
```

This is very minor evidence in favor of the alternative model (Bayes factor $BF_{10} \approx 1.94$). We would not like to draw any (strong) categorical conclusions from this result, regarding the question of whether the coin might be fair. Figure 12.3 also shows the relation between prior and posterior at the point-value of interest.

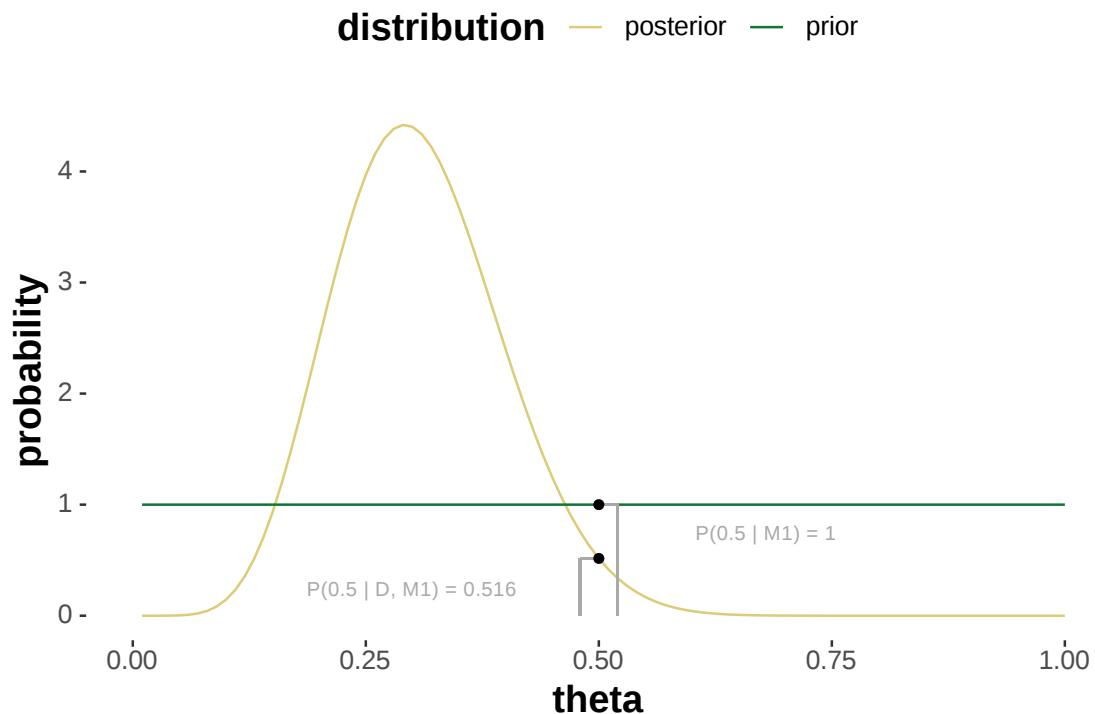


Figure 12.3.: Illustration of the Savage-Dickey method of Bayes factor computation for the 24/7 case.

12.3.4. Example: eco-sensitivity

To apply the Savage-Dickey method to the eco-sensitivity model, we have to obtain an estimate for the posterior density at the critical value $\delta = 0$ from the posterior samples. An approximate method for obtaining this value is implemented in the `polyspline` package (using poly-nomial splines to approximate the posterior curve).

```
# extract the samples for parameter delta
delta_samples <- tidy_draws_tt2 %>%
  filter(Parameter == "delta") %>%
  pull(value)
# estimating density of posterior at delta = 0 with Savage-Dickey
fit.posterior <- polspline::logspline(delta_samples)
posterior_delta_null <- polspline::dlogspline(0, fit.posterior)
prior_delta_null <- dnorm(0,0, sd_delta)
BF_delta_null = posterior_delta_null / prior_delta_null
BF_delta_null

## [1] 0.4421741
```

We conclude from this that there is only very mild (unnoteworthy) evidence in favor of the alternative hypothesis.

12.4. Bayes factors for ROPE-d hypotheses through encompassing models

The Savage-Dickey method can be generalized to cover also interval-valued hypotheses in general, and therefore also ROPE-d hypotheses in particular. The previous literature has focused on inequality-based intervals/hypotheses (like $\theta \geq 0.5$) (Klugkist, Kato, and Hoijtink 2005; Wetzels, Grasman, and Wagenmakers 2010; Oh 2014). Here, we show that this method also extends to arbitrary intervals. The advantage of this method is that we can use samples from the posterior distribution to approximate integrals which is more robust than having to estimate point-values of posterior density.

The main idea, following previous work (Klugkist, Kato, and Hoijtink 2005; Wetzels, Grasman, and Wagenmakers 2010; Oh 2014) is to use so-called **encompassing priors**. Let θ be a single parameter of interest (for simplicity), which can in principle take on any real value. We are interested in the interval-based hypotheses:

- $H_0: \theta \in [a; b]$, and
- $H_a: \theta \notin [a; b]$

An **encompassing model** M_e has a suitable likelihood function $P_{M_e}(D | \theta, \omega)$ (where ω is a vector of other parameters, beside the parameter θ of interest). It also defines a prior $P_{M_e}(\theta, \omega)$, for which crucially:

$$0 < P_{M_e}(\theta, \omega) < 1$$

This latter constraint makes sure that the parameter ranges of H_0 and H_a are not ruled out *a priori*.

Generalizing over the Savage-Dickey approach, we construct two models, one for each hypothesis, *both* of which are nested under the encompassing model:

12. Bayesian hypothesis testing

- M_0 has prior $P_{M_0}(\theta, \omega) = P_{M_e}(\theta, \omega \mid \theta \in [a; b])$
- M_a has prior $P_{M_0}(\theta, \omega) = P_{M_e}(\theta, \omega \mid \theta \notin [a; b])$

Both M_0 and M_a have the same likelihood function as M_e , which is why we drop the model index for readability in the following.

Figure 12.4 shows an example of the priors of an encompassing model for two nested models based on a ROPE-d hypothesis testing approach.

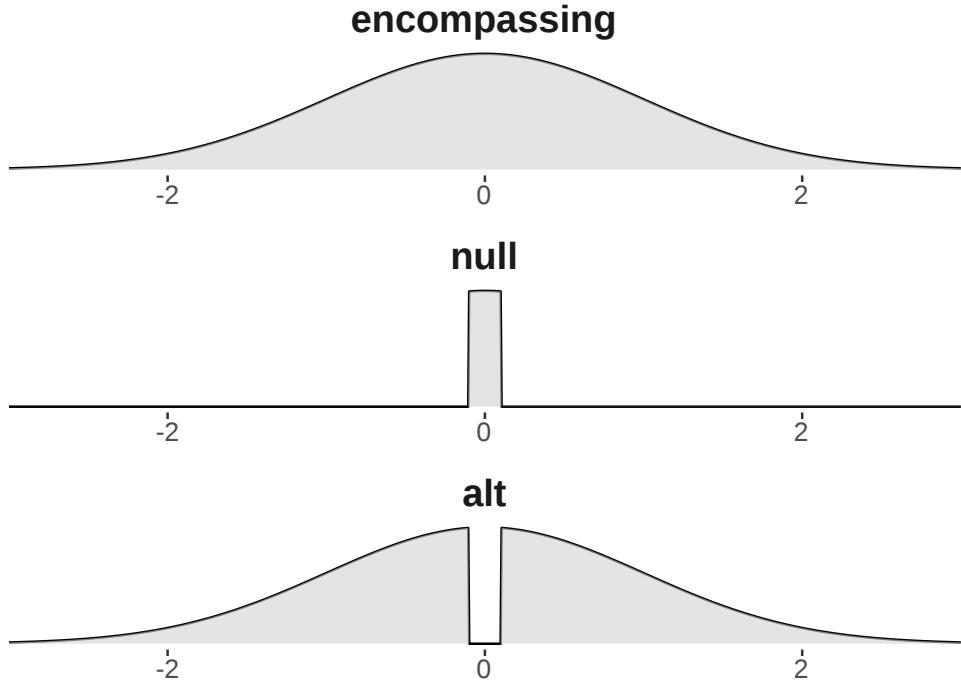


Figure 12.4.: Example of the prior of an encompassing model and the priors of two models nested under it.

Proof

Theorem 12.2. Fix a Bayesian model M (the encompassing model) with prior $P_M(\theta, \omega)$ and likelihood function $P_M(D \mid \theta, \omega)$, where θ is the parameter of interest and ω is a vector of other (nuisance) parameters. Assume that priors over θ are independent of the nuisance parameters ω . For an interval-valued hypothesis $H_0: \theta = \theta^* \pm \epsilon$, the Bayes Factor in favor of this hypothesis over its negation $H_a: \theta \neq \theta^* \pm \epsilon$ can be expressed as:

$$\begin{aligned} BF_{01} &= \frac{\text{posterior-odds of } H_0}{\text{prior-odds of } H_0} \\ &= \frac{P_M(\theta = \theta^* \pm \epsilon \mid D)}{P_M(\theta \neq \theta^* \pm \epsilon \mid D)} \frac{P_M(\theta \neq \theta^* \pm \epsilon)}{P_M(\theta = \theta^* \pm \epsilon)} \end{aligned}$$

Proof. TBD □

12.4.1. Example: 24/7

Using the ROPE-d method to compute the interval-valued hypothesis $\theta = 0.5 \pm \epsilon$ is:

```
# set the scene
theta_null <- 0.5
epsilon <- 0.01          # epsilon margin for ROPE
upper <- theta_null + epsilon    # upper bound of ROPE
lower <- theta_null - epsilon    # lower bound of ROPE
# calculate prior odds of the ROPE-d hypothesis
prior_of_hypothesis <- pbeta(upper, 1, 1) - pbeta(lower, 1, 1)
prior_odds <- prior_of_hypothesis / (1 - prior_of_hypothesis)
# calculate posterior odds of the ROPE-d hypothesis
posterior_of_hypothesis <- qbeta(upper, 8, 18) - qbeta(lower, 8, 18)
posterior_odds <- posterior_of_hypothesis / (1 - posterior_of_hypothesis)
# calculate Bayes Factor
bf_ROPEd_hypothesis <- posterior_odds / prior_odds
bf_ROPEd_hypothesis

## [1] 0.2243274
```

This is mild evidence in favor of the alternative hypothesis (Bayes factor $BF_{10} \approx 4.46$).

12.4.2. Example: eco-sensitivity

```
# estimating BF for ROPE-d hypothesis with encompassing priors
delta_null <- 0
epsilon <- 0.25          # epsilon margin for ROPE
upper <- delta_null + epsilon    # upper bound of ROPE
lower <- delta_null - epsilon    # lower bound of ROPE
# calculate prior odds of the ROPE-d hypothesis
prior_of_hypothesis <- pnorm(upper, 0, sd_delta) - pnorm(lower, 0, sd_delta)
prior_odds <- prior_of_hypothesis / (1 - prior_of_hypothesis)
# calculate posterior odds of the ROPE-d hypothesis
posterior_of_hypothesis <- mean( lower <= delta_samples & delta_samples <= upper )
posterior_odds <- posterior_of_hypothesis / (1 - posterior_of_hypothesis)
# calculate Bayes Factor
bf_ROPEd_hypothesis <- posterior_odds / prior_odds
bf_ROPEd_hypothesis
```

12. Bayesian hypothesis testing

```
## [1] 0.478977
```

This is only minor evidence in favor of the alternative hypothesis (Bayes factor $\text{BF}_{10} \approx 2.09$).

Part IV.

Applied (generalized) linear modeling

13. Simple linear regression

This chapter introduces the basics of (simple) linear regression modeling with one explanatory variable. It covers ordinary least-squares regression, a frequentist maximum-likelihood approach, as well as a Bayesian approach. It addresses how hypotheses about the values of a regression model's parameters (so-called coefficients) can be addressed in a frequentist and a Bayesian approach.

13.1. Data set: murder data

As a running example we use data on murder rates in cities of different population size, also containing further socio-economic information.

```
murder_data <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-data-analysis/1.0.0/data/murder.csv')) %>%  
  rename(murder_rate = annual_murder_rate_per_million_inhabitants,  
         low_income = percentage_low_income,  
         unemployment = percentage_unemployment) %>%  
  select(murder_rate, low_income, unemployment, population)
```

We take a look at the data:

```
murder_data  
  
## # A tibble: 20 x 4  
##   murder_rate low_income  unemployment population  
##       <dbl>      <dbl>        <dbl>      <dbl>  
## 1       11.2      16.5        6.2    587000  
## 2       13.4      20.5        6.4    643000  
## 3       40.7      26.3        9.3    635000  
## 4       5.3       16.5        5.3    692000  
## 5       24.8      19.2        7.3   1248000  
## 6       12.7      16.5        5.9    643000  
## 7       20.9      20.2        6.4   1964000  
## 8       35.7      21.3        7.6   1531000  
## 9       8.7       17.2        4.9    713000  
## 10      9.6       14.3        6.4    749000  
## 11     14.5      18.1        6     7895000
```

13. Simple linear regression

```
## 12      26.9      23.1      7.4    762000
## 13      15.7      19.1      5.8    2793000
## 14      36.2      24.7      8.6    741000
## 15      18.1      18.6      6.5    625000
## 16      28.9      24.9      8.3    854000
## 17      14.9      17.9      6.7    716000
## 18      25.8      22.4      8.6    921000
## 19      21.7      20.2      8.4    595000
## 20      25.7      16.9      6.7    3353000
```

Each row in this data set shows data from a city. The information in the columns is:

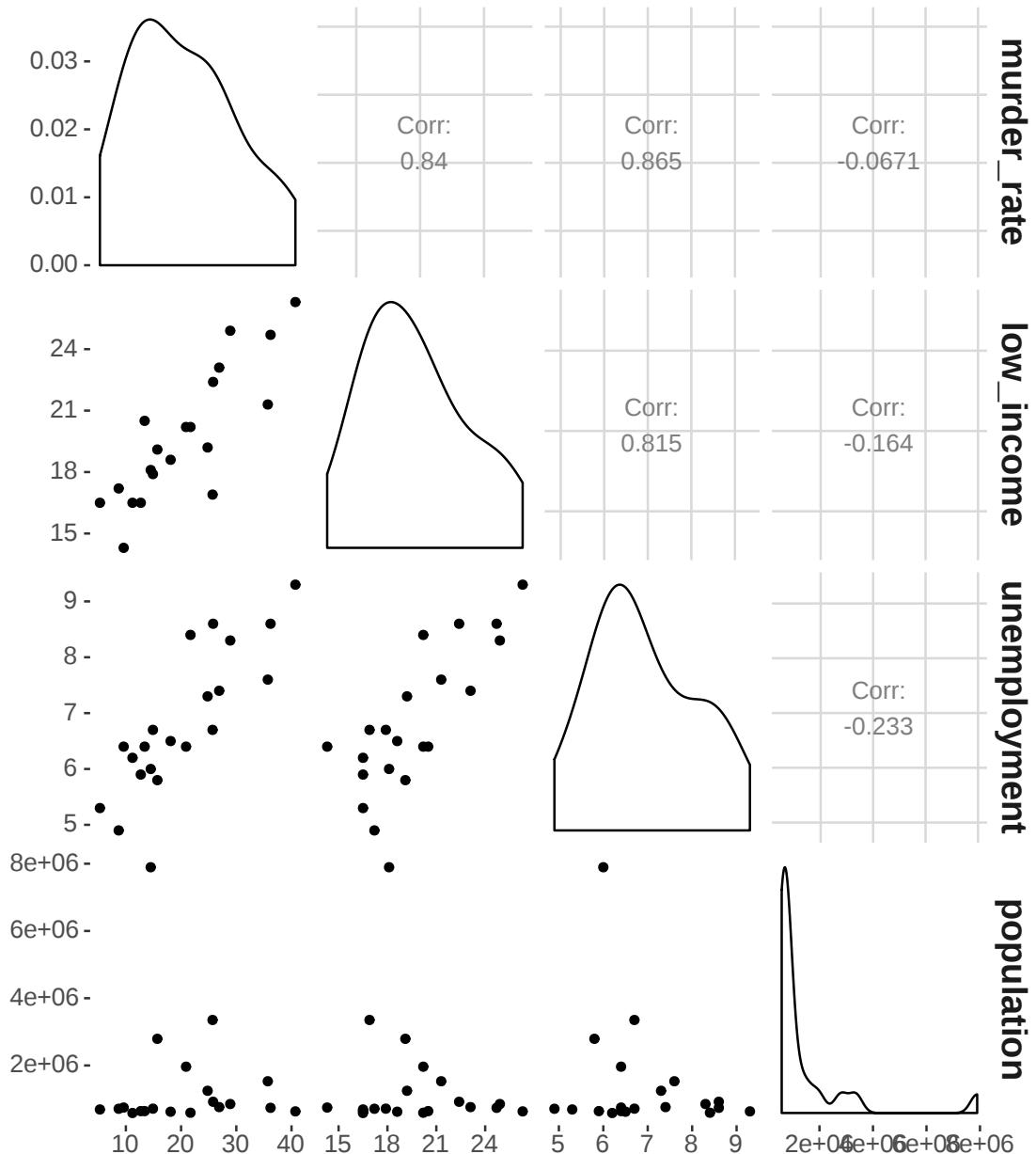
- `murder_rate`: annual murder rate per million inhabitants
- `low_income`: percentage of inhabitants with a low income (however that is defined)
- `unemployment`: percentage of unemployed inhabitants
- `population`: number of inhabitants of a city

Here's a nice way of plotting each variable against each other:

```
GGally::ggpairs(murder_data, title = "Murder rate data")
```

Murder rate data

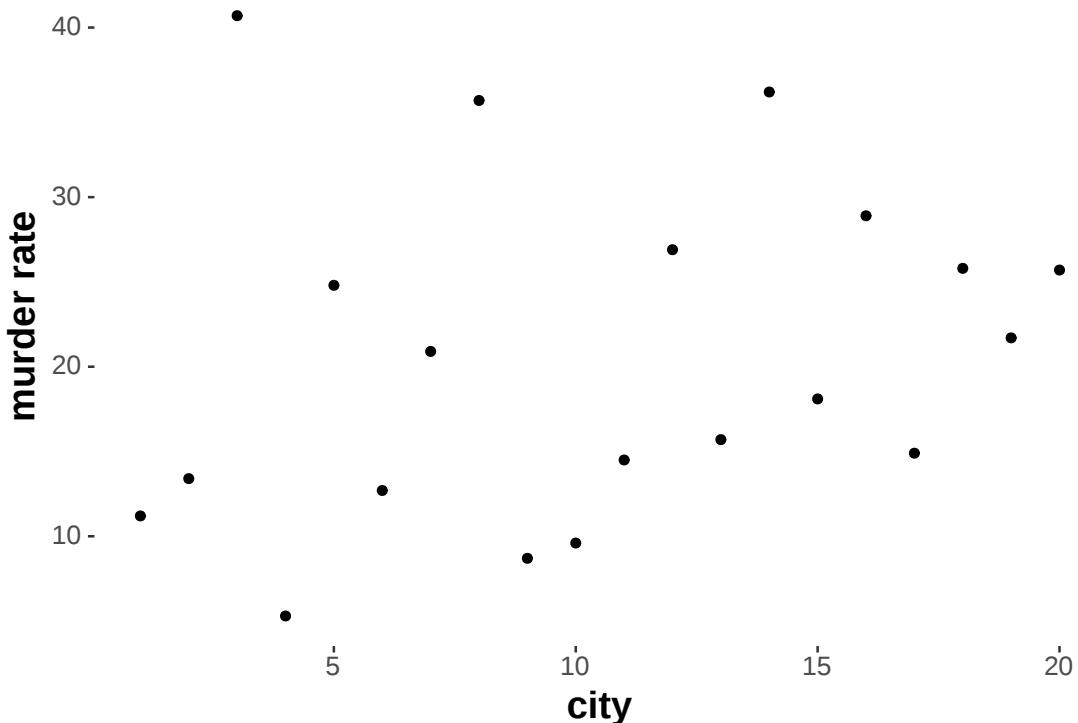
`murder_rate` `low_income` `unemployment` `population`



The diagonal of this graph shows the density curve of the data in each column. Scatter plots below the diagonal show pairs of values from two columns plotted against each other. The information above the diagonal gives the correlation score of each pair of variables.

13.2. What is a (simple) linear regression?

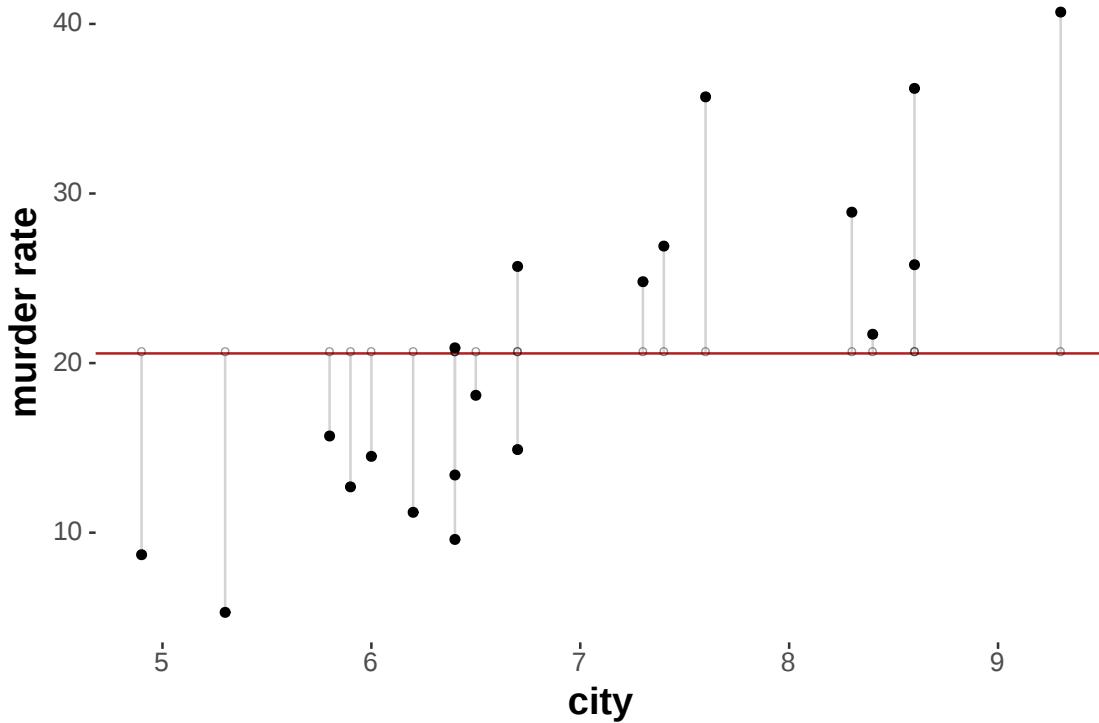
We are interested in explaining or predicting the murder rates in a city. Suppose we have nothing to explain it with, i.e., we only have a vector of murder rates. Let's plot the murder rate for every city (just numbered consecutively):



13.2.1. Prediction without any further information

Suppose we knew all observed murder rates. If we then wanted to predict the murder rate of a random city , but had no further information about that city, our best guess would be the mean of the observed murder rates, because this is what minimizes (on average) the distance to the observed murder rates.

The plot below visualizes the prediction we make by this naive approach. The black dots show the data points, the red line shows the prediction we make (the mean murder rate), the small hollow dots show the specific predictions for each observed value x_i and the gray lines show the distance between our prediction and the actual data observation.



The mean distance could be captured in terms of the **total sum of squares** like this, where y is the n -placed vector observed murder rates and \bar{y} is its mean:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$$

In the case at hand, that is:

```
y <- murder_data %>% pull(murder_rate)
n <- length(y)
tss_simple <- sum((y - mean(y))^2)
tss_simple

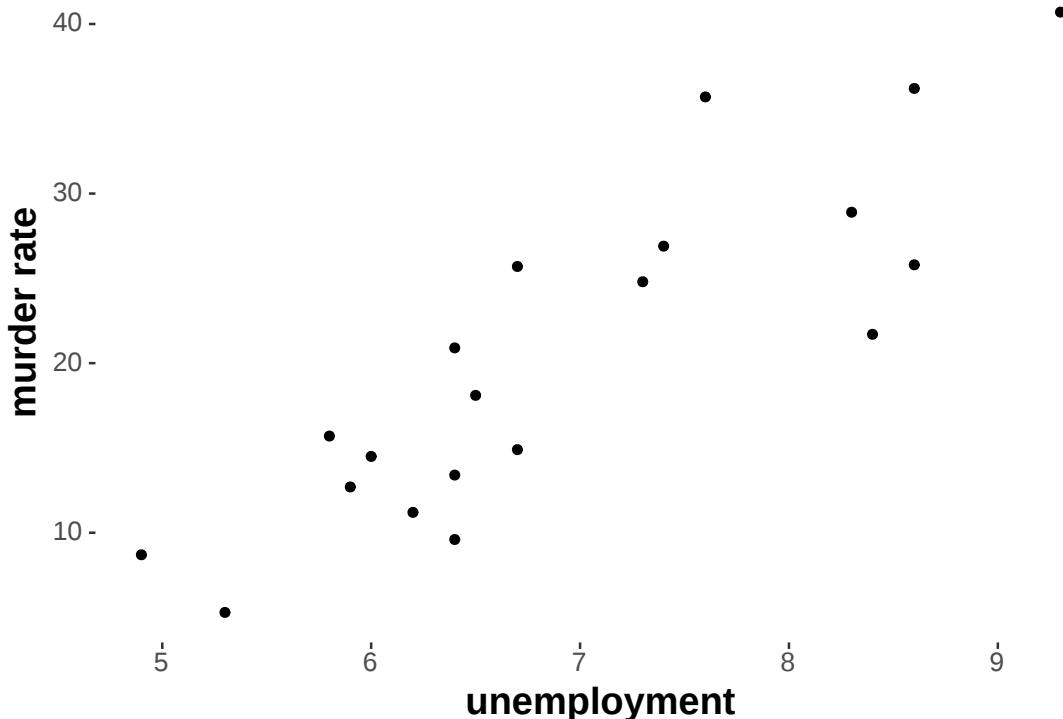
## [1] 1855.202
```

13.2.2. Prediction with knowledge of unemployment rate

We might not be very content with this prediction error. Suppose we could use some piece of information about the random city whose murder rate we are trying to predict. E.g., we might happen to know the value of the variable `unemployment`. How could that help us make a better prediction?

13. Simple linear regression

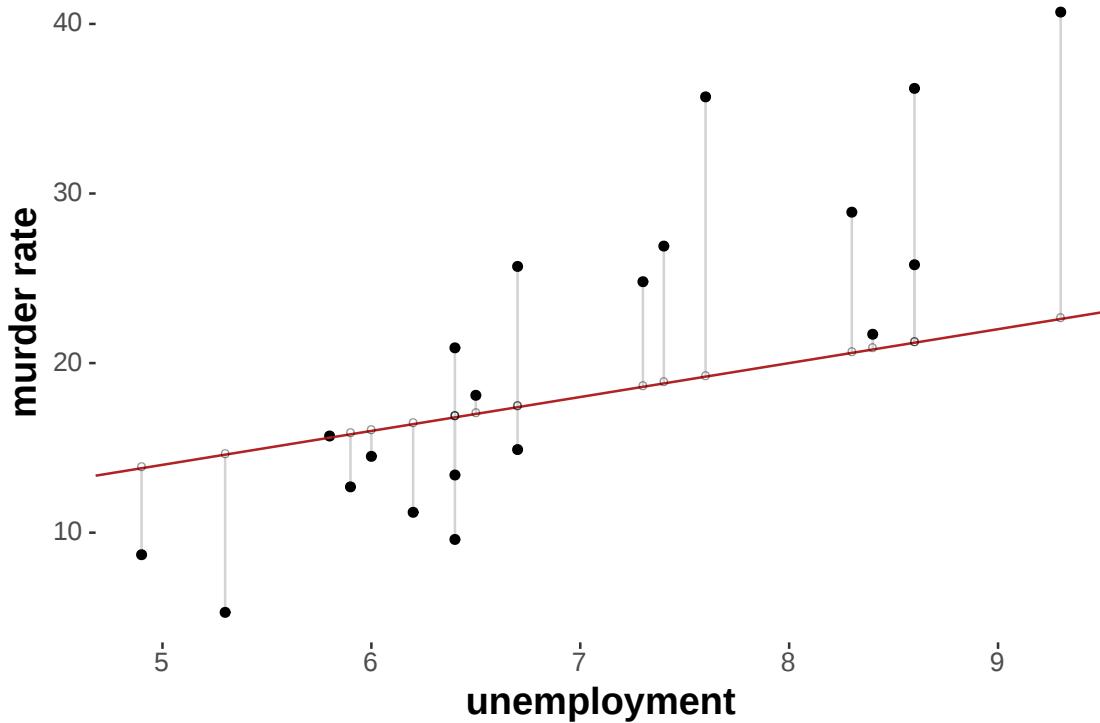
There does seem to be some useful information in the unemployment rate, which may lead to better predictions of the murder rate. We see this in a scatter plot:



Let us assume, for the sake of current illustration, that we expect a very particular functional relationship between the variables `murder_rate` and `unemployment`. For some reason or other, we hypothesize that even with 0% unemployment, the murder rate would be positive, namely at 4 murders per million inhabitants. We further hypothesize that with each increase of 1% in the unemployment percentage, the murder rate per million increases by 2. The functional relationship between dependent variable y (= murder rate) and predictor variable x (= unemployment) would then be expressible as a linear function (the hat on variable y indicates that these are not data observations but predictions):

$$\hat{y}_i = 2x_i + 4$$

Here is a graphical representation of this functional relationship. Again, the black dots show the data points, the red line the linear function $f(x) = 2x + 4$, the small hollow dots show the specific predictions for each observed value x_i and the gray lines show the distance between our prediction and the actual data observation. (Notice that there are data points for which the unemployment rate is the same, but we observed different murder rates.)



We can again quantify our prediction error in terms of a sum of squares like we did before. For the case of a prediction vector \hat{y} , the quantity in question is called **residual sum of squares**.

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y})^2$$

Here is how we can calculate RSS in R:

```
y <- murder_data %>% pull(murder_rate)
x <- murder_data %>% pull(unemployment)
predicted_y <- 2 * x + 4
n <- length(y)
rss_guesswork <- sum((y - predicted_y)^2)
rss_guesswork

## [1] 1327.74
```

Compared to the previous prediction, which was based on the mean \bar{y} only, this linear function reduces the prediction error (measured here geometrically in terms of a sum of squares).

13.2.3. Simple linear regression: general problem formulation

Suppose we have k predictor variables x_1, \dots, x_k and dependent variable y . We consider the simple linear relation (where the hat on top of vector y symbolizes that this is a vector of predictions):

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

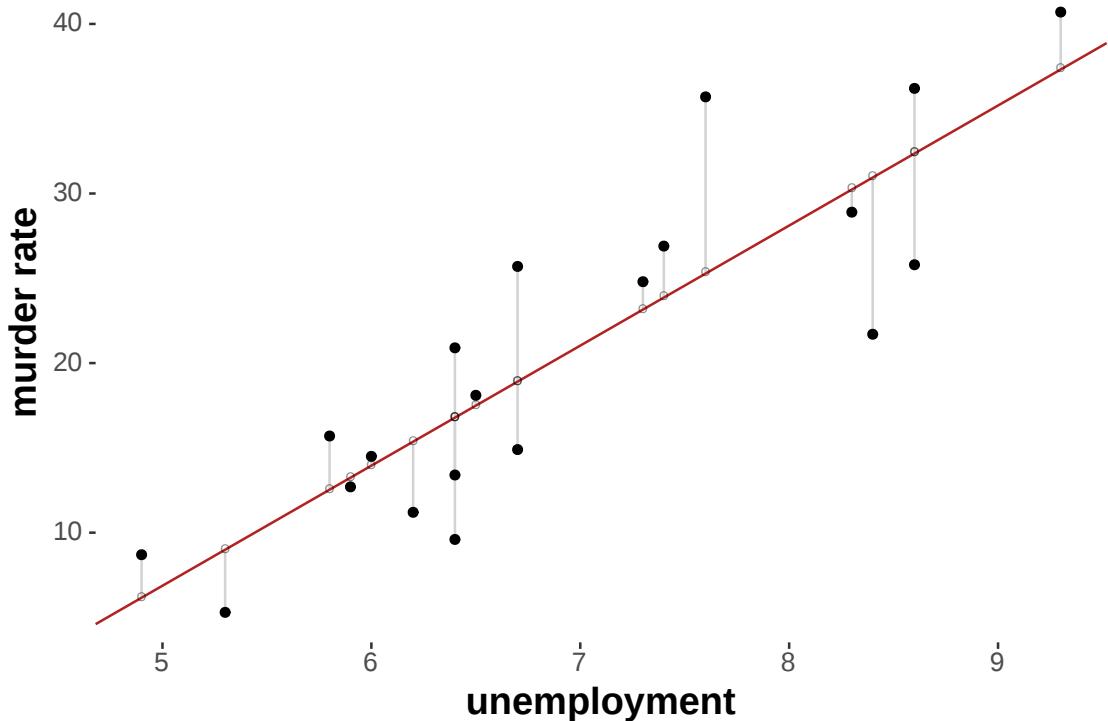
The parameters $\beta_0, \beta_1, \dots, \beta_k$ of this equation are called **regression coefficients**. In particular, β_0 is called the **regression intercept** and β_1, \dots, β_k are **regression slope coefficients**. Based on the predictions of a parameter vector $\langle \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k \rangle$, we consider the residual sum of squares as a measure of prediction error:

$$\text{RSS}_{\langle \beta_0, \beta_1, \dots, \beta_k \rangle} = \sum_{i=1}^k (y_i - \hat{y}_i)^2$$

We would like to find the best parameter values (denoted traditionally by a hat on the parameter's variable: $\hat{\beta}_i$) in the sense of minimizing the residual sum of squares:

$$\langle \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k \rangle = \arg \min_{\langle \beta_0, \beta_1, \dots, \beta_k \rangle} \text{RSS}_{\langle \beta_0, \beta_1, \dots, \beta_k \rangle}$$

For instance, the example started above, where we regressed `murder_rate` against `unemployment` has two regression coefficients: an intercept term and a slope for `unemployment`. The optimal solution for these (see next section) delivers the regression line in the graph below:



The total sum of squares for the best fitting parameters is:

```
## [1] 467.6023
```

The next section will explain how we find best-fitting parameter values in the sense above.

13.3. Ordinary least-squares regression

This section looks at different ways of finding values for regression coefficients that minimize the residual sum of squares.

13.3.1. Finding optimal parameters with `optim`

We can use the `optim` function to find the best-fitting parameter values for a simple linear regression.

Here is an example based on the murder data.

```
# data to be explained / predicted
y <- murder_data %>% pull(murder_rate)
# data to use for prediction / explanation
```

13. Simple linear regression

```
x <- murder_data %>% pull(unemployment)
# function to calculate residual sum of squares
get_rss = function(y, x, beta_0, beta_1) {
  yPred = beta_0 + x * beta_1
  sum((y-yPred)^2)
}
# finding best-fitting values for TSS
fit_rss = optim(par = c(0, 1),
  fn = function(par) {
    get_rss(y, x, par[1], par[2])
  }
)
# output the results
message(
  "Best fitting parameter values:",
  "\n\tIntercept: ", fit_rss$par[1] %>% signif(5),
  "\n\tSlope: ", fit_rss$par[2] %>% signif(5),
  "\nRSS for best fit: ", fit_rss$value %>% signif(5)
)

## Best fitting parameter values:
## Intercept: -28.528
## Slope: 7.0795
## RSS for best fit: 467.6
```

13.3.2. Fitting OLS regression lines with lm

R also has a built-in function `lm` which fits (simple) linear regression models via RSS minimization. Here is how you call this function for the running example:

```
# fit an OLS regression
fit_lm <- lm(
  # the formula argument specifies dependent and independent variables
  formula = murder_rate ~ unemployment,
  # we also need to say where the data (columns) should come from
  data = murder_data
)
# output the fitted object
fit_lm

##
## Call:
```

```
## lm(formula = murder_rate ~ unemployment, data = murder_data)
##
## Coefficients:
##   (Intercept)  unemployment
##             -28.53          7.08
```

The output of the fitted object shows the best-fitting values (compare them to what we obtained by hand).

It also shows the function call by which this fit was obtained. There is more information in the object `fit_lm` and we will return to this later when we consider hypothesis testing on regression coefficients.

But it might be interesting to take a quick preview already:

```
summary(fit_lm)

##
## Call:
## lm(formula = murder_rate ~ unemployment, data = murder_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2415 -3.7728  0.5795  3.2207 10.4221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.5267    6.8137  -4.187 0.000554 ***
## unemployment  7.0796    0.9687   7.309 8.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.097 on 18 degrees of freedom
## Multiple R-squared:  0.748, Adjusted R-squared:  0.7339
## F-statistic: 53.41 on 1 and 18 DF,  p-value: 8.663e-07
```

13.3.3. Finding optimal parameter values with math

It is also possible to determine the OLS-fits by a mathematical derivation.

Theorem 13.1 (OLS solution). *For a simple linear regression model with just one predictor, the solution for:*

$$\arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^k (y_i - (\beta_0 + \beta_1 x_i))^2$$

is given by:

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof. TBD □

Let's use these formulas to calculate regression coefficients for the running example as well:

```
tibble(
  beta_1 = cov(x,y) / var(x),
  beta_0 = mean(y) - beta_1 * mean(x)
)

## # A tibble: 1 x 2
##   beta_1 beta_0
##     <dbl>  <dbl>
## 1    7.08 -28.5
```

A similar result exists also for regression with more than one predictor variable. This explains why the computation of best-fitting regression coefficients with a built-in function like `lm` is lightning fast (as compared to using `optim` or a Bayesian approach which relies on MCMC sampling).

13.4. A maximum-likelihood approach

In order to be able to extend regression modeling to predictor variables other than metric variables (so-called generalized linear regression models), the geometric approach needs to be abandoned in favor of a likelihood-based approach.

There are two equivalent formulation of a (simple) linear regression model, using a likelihood-based approach. The first is more explicit, showing clearly that the model assumes that for each observation y_i , the model assumes an error term ϵ_i , which is an iid sample from a Normal distribution. (Notice that the likelihood-based model assumes an additional parameter σ , the standard deviation of the error terms.)

likelihood-based regression [explicit version]

$$\begin{aligned} y_{\text{pred}} &= \beta_0 + \beta_1 x \\ y_i &= y_{\text{pred}} + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \sigma) \end{aligned}$$

The second, equivalent version of this writes this more compactly, suppressing the explicit mentioning of iid error terms:

likelihood-based regression [compact version]

$$\begin{aligned}y_{\text{pred}} &= \beta_0 + \beta_1 x \\y &\sim \text{Normal}(\mu = y_{\text{pred}}, \sigma)\end{aligned}$$

We can use `optim` to find maximum likelihood estimates:

```
# data to be explained / predicted
y <- murder_data %>% pull(murder_rate)
# data to use for prediction / explanation
x <- murder_data %>% pull(unemployment)
# function to calculate negative log-likelihood
get_nll = function(y, x, beta_0, beta_1, sd) {
  if (sd <= 0) {return( Inf )}
  yPred = beta_0 + x * beta_1
  nll = -dnorm(y, mean=yPred, sd=sd, log = T)
  sum(nll)
}
# finding MLE
fit_lh = optim(par = c(0, 1, 1),
  fn = function(par) {
    get_nll(y, x, par[1], par[2], par[3])
  }
)
# output the results
message(
  "Best fitting parameter values:",
  "\n\tIntercept: ", fit_lh$par[1] %>% signif(5),
  "\n\tSlope: ", fit_lh$par[2] %>% signif(5),
  "\nNegative log-likelihood for best fit: ", fit_lh$value %>% signif(5)
)

## Best fitting parameter values:
## Intercept: -28.517
## Slope: 7.0783
## Negative log-likelihood for best fit: 59.898
```

13. Simple linear regression

It is no coincidence that these fitted values are (modulo number imprecision) the same as for the geometric OLS approach.

Theorem 13.2 (MLE solution). *For a simple linear regression model with just one predictor, the solution for:*

$$\arg \max_{\langle \beta_0, \beta_1, \sigma \rangle} \prod_{i=1}^k \text{Normal}(\mu = \beta_0 + \beta_1 x_i, \sigma)$$

is the same as for the OLS approach:

$$\hat{\beta}_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof. TBD □

The equivalence also extends to cases with more than one explanatory variable.

R also has a built-in way of approaching simple linear regression with a maximum-likelihood approach, namely by using the function `glm` (generalized linear model). Notice that the output looks slightly different from that of `lm`.

```
fit_glm <- glm(murder_rate ~ unemployment, data = murder_data)
fit_glm

##
## Call: glm(formula = murder_rate ~ unemployment, data = murder_data)
##
## Coefficients:
## (Intercept)  unemployment
##           -28.53          7.08
##
## Degrees of Freedom: 19 Total (i.e. Null); 18 Residual
## Null Deviance:      1855
## Residual Deviance: 467.6      AIC: 125.8
```

We might also risk a peek at the summary of `fit_glm`:

```
summary(fit_glm)
```

```

## 
## Call:
## glm(formula = murder_rate ~ unemployment, data = murder_data)
## 
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -9.2415  -3.7728   0.5795   3.2207  10.4221
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -28.5267    6.8137  -4.187 0.000554 ***
## unemployment  7.0796    0.9687   7.309 8.66e-07 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 25.9779)
## 
## Null deviance: 1855.2 on 19 degrees of freedom
## Residual deviance: 467.6 on 18 degrees of freedom
## AIC: 125.8
## 
## Number of Fisher Scoring iterations: 2

```

13.5. A Bayesian approach

A Bayesian model for (simple) linear regression looks very much like the previous likelihood-based model, just that it also adds prior information. We have already seen a Bayesian linear regression model in Chapter 8.5.6. It is repeated here in Figure 13.1.

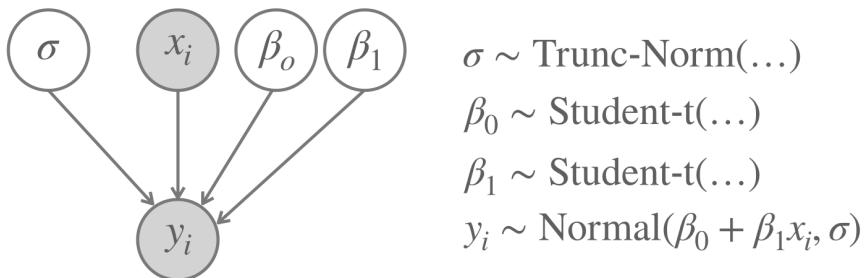


Figure 13.1.: Bayesian Simple Linear Regression Model (repeated from before).

13.5.1. Implementation in greta

Here is an implementation of a Bayesian regression model for the running example murder data using greta:

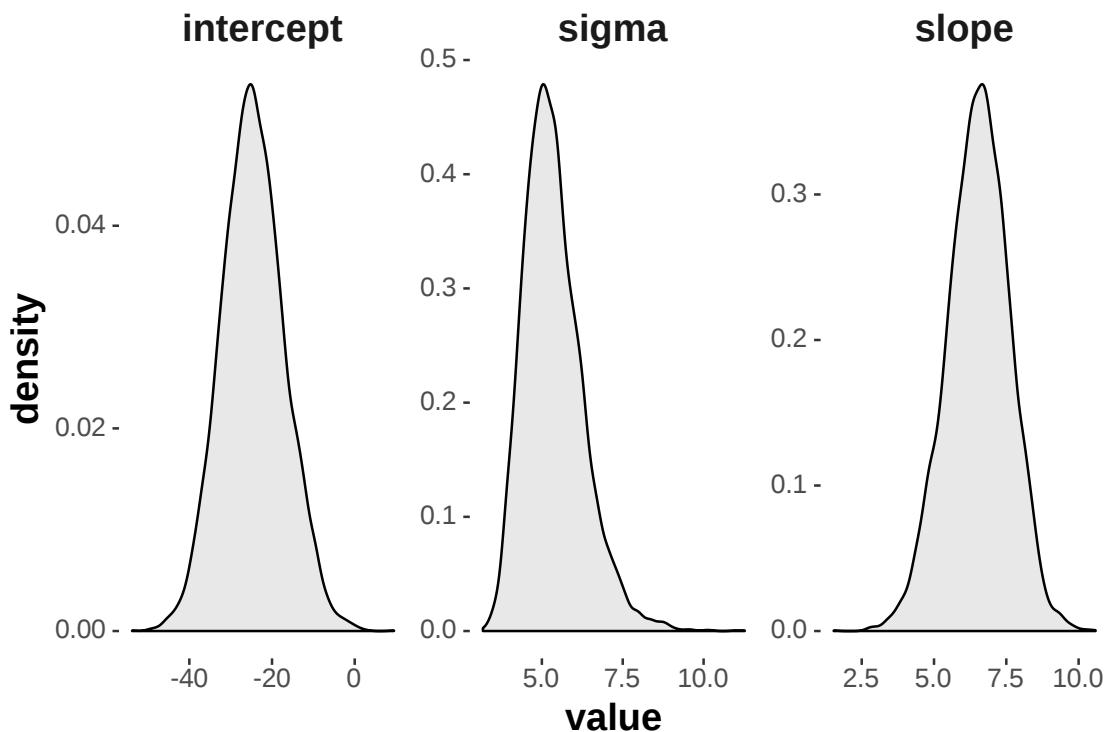
```
# data to be explained / predicted
y <- murder_data %>% pull(murder_rate)
# data to use for prediction / explanation
x <- murder_data %>% pull(unemployment)
y_greta      <- as_data(y)
x_greta      <- as_data(x)
# latent variables and priors
intercept <- student(df= 1, mu = 0, sigma = 10)
slope       <- student(df= 1, mu = 0, sigma = 10)
sigma       <- normal(0, 5, truncation = c(0, Inf))
# derived latent variable (linear model)
y_pred <- intercept + slope * x_greta
# likelihood
distribution(y) <- normal(y_pred, sigma)
# finalize model, register which parameters to monitor
murder_model <- model(intercept, slope, sigma)
```

We can draw samples from the posterior distribution as usual:

```
# draw samples
draws_murder_data <- greta::mcmc(
  murder_model,
  n_samples = 2000,
  chains = 4,
  warmup = 1000
)
# cast results (type 'mcmc.list') into tidy tibble
tidy_draws_murder_data <- ggmc::ggs(draws_murder_data)
```

Here is a plot of the posterior:

```
# plot posterior
tidy_draws_murder_data %>%
  ggplot(aes(x = value)) +
  geom_density(fill = "lightgray", alpha = 0.5) +
  facet_wrap(~ Parameter, scales = "free")
```



13.5.2. Using the `brms` package

Instead of hand-coding each Bayesian regression model, we can also use the `brms` package (Bürkner 2017). The main function of this package is `brm` (short for Bayesian regression model). It behaves very similarly to the `glm` function we saw above.¹ Here is an example for the current case study:

```
fit_brms_murder <- brm(
  # specify what to explain in terms of what
  # using the formula syntax
  formula = murder_rate ~ unemployment,
  # which data to use
  data = murder_data
)
```

The function `brm` returns a model-fit object, similar to `glm`. We can inspect it to get more information, using the `summary` function:

¹Actually, `brm` behaves similar to the more general `lmer` function from the `lme4` package, which is even more general than `glm`. Both `lmer` and `brm` also cover so-called hierarchical regression models.

```

summary(fit_brms_murder)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: murder_rate ~ unemployment
## Data: murder_data (Number of observations: 20)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     -28.48      7.32   -42.05  -13.79 1.00    3014    2362
## unemployment     7.07      1.04     4.97    9.04 1.00    2978    2451
##
## Family Specific Parameters:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma        5.42      0.96     3.88    7.63 1.00    2664    2196
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

```

This output tells us about which model we fitted, states some properties of the MCMC sampling routine used to obtain samples from the posterior distribution, and it also gives us information about the posteriors under the heading “Population-Level Effects” we get summary statistics of the set of posterior samples for each regression coefficient, including the mean (“Estimate”), and the 95% interquartile range (“l-95%” is the lower bound and “u-95%” is the upper bound).²

13.6. Testing coefficients

13.6.1. Bayesian approach

```

# get means and 95% HDI
Bayes_estimates <- tidy_draws_murder_data %>%
  group_by(Parameter) %>%
  summarise(
    '|95%' = HDInterval::hdi(value)[1],
    mean = mean(value),

```

²Notice that the 95% interquartile range is not necessarily the same as the 95% HDI, but for large sample sizes the two will coincide.

```
'95|%' = HDInterval::hdi(value)[2]
)
Bayes_estimates
```

```
## # A tibble: 3 x 4
##   Parameter `|95%`  mean `|95%` 
##   <fct>     <dbl>  <dbl>  <dbl>
## 1 intercept -39.1  -24.6  -9.51
## 2 sigma      3.78   5.36   7.20
## 3 slope      4.34   6.53   8.53
```

13.6.2. Frequentist approach

Figure 13.2 shows a frequentist model for testing the hypotheses that the regression coefficients are equal to zero.

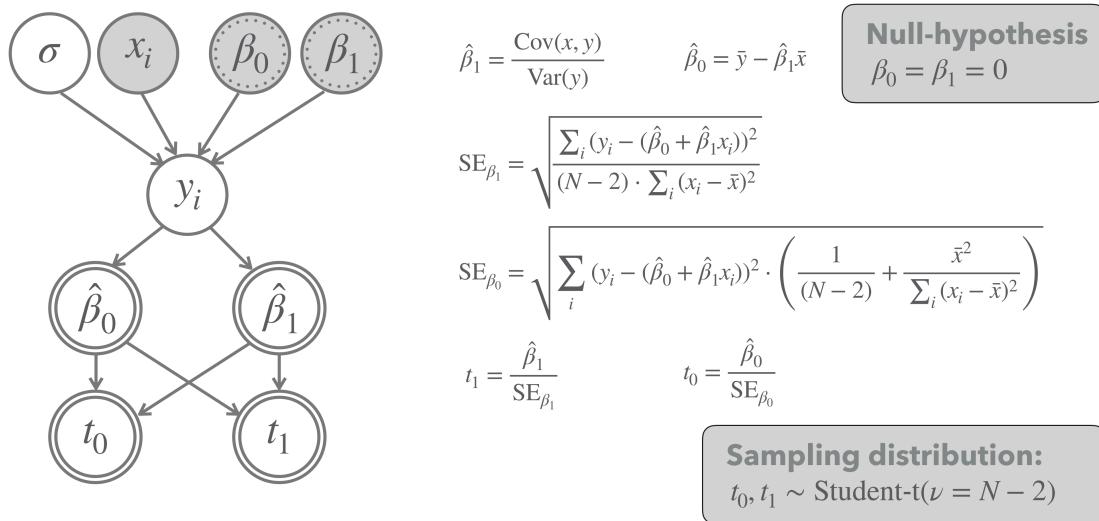


Figure 13.2.: Frequentist model for testing whether regression coefficients are plausibly equal to zero.

We use this model to compute the test statics for the observed data:

```
# observed data
y_obs <- murder_data %>% pull(murder_rate)
x_obs <- murder_data %>% pull(unemployment)
n_obs <- length(y_obs)
# best-fitting coefficients
beta_1_hat <- cov(x_obs, y_obs) / var(x_obs)
beta_0_hat <- mean(y_obs) - beta_1_hat * mean(x_obs)
```

13. Simple linear regression

```
# calculating t-scores
MSE <- sum((y_obs - beta_0_hat - beta_1_hat * x_obs)^2) / (n_obs-2)
S_xx <- sum((x_obs - mean(x_obs))^2)
SE_beta_1_hat <- sqrt(MSE / S_xx)
SE_beta_0_hat <- sqrt(MSE * (1/n_obs + mean(x_obs)^2 / S_xx))
t_slope = (beta_1_hat) / SE_beta_1_hat
t_intercept = beta_0_hat / SE_beta_0_hat
tibble(t_slope, t_intercept)

## # A tibble: 1 x 2
##   t_slope t_intercept
##       <dbl>      <dbl>
## 1     7.31     -4.19
```

Calculate p -values (two sided!) for both of these values:

```
p_value_intercept = pt(t_intercept, df = n_obs -2) + 1-pt(-t_intercept, df = n_obs -2)
p_value_slope     = pt(-t_slope, df = n_obs -2) + 1-pt(t_slope, df = n_obs -2)
tibble(p_value_intercept, p_value_slope)

## # A tibble: 1 x 2
##   p_value_intercept p_value_slope
##       <dbl>        <dbl>
## 1     0.000554    0.000000866
```

We compare the manual calculation to the that of the built-in functions `lm` and `glm`:

```
summary(lm(murder_rate ~ unemployment, data = murder_data))

##
## Call:
## lm(formula = murder_rate ~ unemployment, data = murder_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.2415 -3.7728  0.5795  3.2207 10.4221 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -28.5267    6.8137  -4.187 0.000554 ***
## unemployment  7.0796    0.9687   7.309 8.66e-07 ***
```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.097 on 18 degrees of freedom
## Multiple R-squared: 0.748, Adjusted R-squared: 0.7339
## F-statistic: 53.41 on 1 and 18 DF, p-value: 8.663e-07

summary(glm(murder_rate ~ unemployment, data = murder_data))

##
## Call:
## glm(formula = murder_rate ~ unemployment, data = murder_data)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -9.2415   -3.7728    0.5795   3.2207   10.4221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.5267    6.8137  -4.187 0.000554 ***
## unemployment  7.0796    0.9687   7.309 8.66e-07 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 25.9779)
##
## Null deviance: 1855.2 on 19 degrees of freedom
## Residual deviance: 467.6 on 18 degrees of freedom
## AIC: 125.8
##
## Number of Fisher Scoring iterations: 2

```


14. Beyond simple linear regression

This chapter showcases how linear regression models can be applied flexibly to a variety of data analysis problems.

- categorical predictors
 - two-groups (eco-sensitivity data)
 - more groups (mental chronometry)
 - interactions (Winter data)
- metric and categorical predictors combined
 - avocado data
- logistic regression
 - KoF data

14.1. Two categorical predictors

Let's revisit the (fictitious) eco-sensitivity data from Chapter 12:

```
x_A <- c(  
  104, 105, 100, 91, 105, 118, 164, 168, 111, 107, 136, 149, 104, 114, 107, 95,  
  83, 114, 171, 176, 117, 107, 108, 107, 119, 126, 105, 119, 107, 131  
)  
x_B <- c(  
  133, 115, 84, 79, 127, 103, 109, 128, 127, 107, 94, 95, 90, 118, 124, 108,  
  87, 111, 96, 89, 106, 121, 99, 86, 115, 136, 114  
)
```

Remember that we are interested in the question whether there is a difference in means of group A and group B. Previously we used a *t*-test for this two-group comparison (frequentist or Bayesian). But we can also cast this as a regression problem. Here is how.

Let's first squeeze the data into a tibble:

14. Beyond simple linear regression

```
eco_sensitivity_data <- tibble(
  group = c(rep("A", length(x_A)), rep("B", length(x_B))),
  measurement = c(x_A, x_B)
)
eco_sensitivity_data

## # A tibble: 57 x 2
##   group measurement
##   <chr>     <dbl>
## 1 A          104
## 2 A          105
## 3 A          100
## 4 A          91
## 5 A          105
## 6 A          118
## 7 A          164
## 8 A          168
## 9 A          111
## 10 A         107
## # ... with 47 more rows
```

Notice that this tibble contains the data in a tidy format, i.e., each row contains a tuple of associated measurements. We want to explain or predict the variable `measurement` in terms of the variable `group`. We can then run a regression model with formula `measurement ~ group`. Here's such a model using the Bayesian approach:

```
fit_brms_eco_sensitivity <- brm(
  # specify what to explain in terms of what
  # using the formula syntax
  formula = measurement ~ group,
  # which data to use
  data = eco_sensitivity_data
)
```

Let's inspect the summary information for the posterior samples:

```
# just showing information for the so-called 'fixed effects'
summary(fit_brms_eco_sensitivity)$fixed[,c("Estimate", "l-95% CI", "u-95% CI")]

##           Estimate l-95% CI    u-95% CI
## Intercept 118.68533 111.42867 125.8595713
## groupB    -11.37756 -22.47929  -0.6354608
```

Compare this with the summary of the posterior estimates we obtained from the Bayesian *t*-test model for this data which we implemented in `greta` and ran in Chapter 12.

```
Bayes_estimates_eco
```

```
## # A tibble: 1 x 4
##   Parameter `|95%` mean `|95%` 
##   <fct>     <dbl> <dbl> <dbl>
## 1 delta      -21.0 -10.9 -0.384
```

The mean of the δ parameter looks suspiciously similar to the ominous `groupB` parameter shown in the output of the `bmrs` model fit. The 95% HDIs of the estimated posterior for the δ parameter also look suspiciously like the values of the 95% interquartile range for the `groupB` parameter. This is no coincidence! In fact, the regression model that `bmrs` calculates here is essentially the same as the *t*-test model we implemented in `greta` by hand except for slight different in the choice of the priors and inessential differences in mathematical formulation of the group mean comparison.

The model computed implicitly in the call to `brm` above is a linear regression model of the following form:

$$\hat{y}_i = \beta_0 + \beta_1 x_i \quad y_i \sim \text{Normal}(\mu = \hat{y}_i, \sigma)$$

The only important point is that the vector x , which corresponds to the group information in the column `group` (which contains entries of strings "A" and "B"), is implicitly treated as a vector of zeros and ones. Implicitly, `brm` has chosen the string "A" as the **reference category** which is encoded as zero. The string "B" is the other category, encoded as number 1. As a consequence, the linear model's intercept parameter β_0 can be interpreted as the predicted mean of the reference category: if for some i we have $x_i = 0$, then the predictor \hat{y}_i will just be $\hat{y}_i = \beta_0$; whence that the intercept β_0 will be fitted to the mean of the reference category. If for some i we have $x_i = 1$ instead, the predicted value will be computed as $\hat{y}_i = \beta_0 + \beta_1$, so that the slope term β_1 will effectively play the role of the difference δ between the mean of the groups. Modulo choice of priors and variable naming, the `brm` model encodes a Bayesian *t*-test model exactly like we previously did using `greta`. The upshot is, that we can conceive of a *t*-test as a special case of a linear regression model!

Of course, nothing in this correspondence depends on a Bayesian analysis.

```
fit_glm_eco <- glm(
  # specify what to explain in terms of what
  # using the formula syntax
  formula = measurement ~ group,
  # which data to use
  data = eco_sensitivity_data
)
summary(fit_glm_eco)
```

14. Beyond simple linear regression

```
##  
## Call:  
## glm(formula = measurement ~ group, data = eco_sensitivity_data)  
##  
## Deviance Residuals:  
##      Min       1Q   Median       3Q      Max  
## -35.933  -13.933   -4.444   10.556   57.067  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  118.933     3.783   31.44 <2e-16 ***  
## groupB      -11.489     5.497   -2.09   0.0412 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for gaussian family taken to be 429.3552)  
##  
## Null deviance: 25490 on 56 degrees of freedom  
## Residual deviance: 23615 on 55 degrees of freedom  
## AIC: 511.27  
##  
## Number of Fisher Scoring iterations: 2
```

The p -value associated with the test of whether the slope coefficient (the difference between means) is plausibly zero corresponds to the result we get from a t -test (when assuming equal variance in groups; an assumption that the linear modeling approach makes per default).

```
t.test(x_A, x_B, paired = F, var.equal = T)
```

```
##  
## Two Sample t-test  
##  
## data: x_A and x_B  
## t = 2.0901, df = 55, p-value = 0.04124  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##  0.4732249 22.5045529  
## sample estimates:  
## mean of x mean of y  
## 118.9333 107.4444
```

14.2. More than two categorical predictors

A *t*-test is only applicable to at most two groups. The classical approach to generalizing frequentist testing to the comparison of more than two groups is to use ANOVA, as briefly discussed in Chapter @ (ch-03-05-hypothesis-testing-ANOVA). But we can also use a linear regression approach for this, as demonstrated here based on the mental chronometry data.

We load the data as usual, but also immediately mutate the column `block` (which captures the experimental manipulation we want to use to explain the dependent variable `RT` (= reaction times)) so that the “goNoGo” condition will come first in alphabetic order. This has the effect that, later in regression modeling, this condition will be treated as the reference level to compare other groups against. This makes sense, because our main question of interest is whether these inequalities are supported by the data:

$$\text{RT in 'reaction' < RT in 'goNoGo' < RT in 'discrimination'}$$

So we are interested in the δ s, so to speak, between ‘reaction’ and ‘goNoGo’ and between ‘discrimination’ and ‘goNoGo’.

```
mc_url <- url('https://raw.githubusercontent.com/michael-franke/intro-data-analysis/master/data/mental_chronometry.csv')
mc_data_cleaned <- read_csv(mc_url) %>%
  # renaming to make 'goNoGo' the reference level
  # (dirty hack, but simpler than to messing with contrast coding)
  mutate(
    block = case_when(
      block == "reaction" ~ "B_reaction",
      block == "goNoGo" ~ "A_goNoGo",
      block == "discrimination" ~ "C_discrimination"
    )
  )
```

To fit this model with `brm` we then just need a simple function call with the formula `RT ~ block` that precisely describes what we are interested: to explain reaction times as a function of the experimental condition:

```
fit_brms_mc <- brm(
  # model 'RT' as a function of 'block'
  formula = RT ~ block,
  data = mc_data_cleaned
)
```

To inspect the posterior fits of this model, we can extract the relevant summary statistics as before:

14. Beyond simple linear regression

```
summary(fit_brms_mc)$fixed[,c("Estimate", "l-95% CI", "u-95% CI")]
##                                     Estimate  l-95% CI  u-95% CI
## Intercept                  427.26357 419.6334 434.92959
## blockB_reaction      -127.38043 -136.8158 -117.98286
## blockC_discrimination 60.38725   50.8000  70.04778
```

Notice that there is an intercept term, as before. This corresponds to the mean reaction time of the reference level (here: 'goNoGo'). There are two slope coefficients, one for the difference between the 'goNoGo' and the 'reaction' condition ('blockB_reaction') and another for the difference between the 'goNoGo' and the 'discrimination' condition ('blockC_discrimination').

As we may have expected, the 95% interquartile range for both slope coefficients (which, given the amount of data we have, is almost surely almost identical to the 95% HDI) does not include 0 by a very wide margin. We could therefore conclude, based on a Bayesian approach to hypothesis testing in terms of posterior estimation, that the reaction times of conditions are credibly different.

The function call for a frequentist analysis with `glm` is almost identical:

```
fit_glm_mc <- glm(
  # model 'RT' as a function of 'block'
  formula = RT ~ block,
  data = mc_data_cleaned
)
```

The summary of the model fit reveals *p*-values for both slope coefficients, which indicate (unsurprisingly) that there is very strong evidence against the null hypothesis of no difference between the means of the two pairs of conditions we compare with this model:

```
summary(fit_glm_mc)

##
## Call:
## glm(formula = RT ~ block, data = mc_data_cleaned)
##
## Deviance Residuals:
##       Min        1Q     Median        3Q       Max
## -219.64    -52.85    -15.43     41.15    396.36
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 427.226    4.059 105.25  <2e-16 ***
## blockB_reaction            -127.381    4.977 -25.60  <2e-16 ***
```

```

## blockC_discrimination 60.414      4.989    12.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8007.877)
##
## Null deviance: 36589210  on 2403  degrees of freedom
## Residual deviance: 19226912  on 2401  degrees of freedom
## AIC: 28435
##
## Number of Fisher Scoring iterations: 2

```

14.3. Interaction terms in factorial designs

The following content is a distilled version of a short tutorial on Bayesian regression modeling for factorial designs (Franke and Roettger 2019), which can be downloaded here. We consider data on voice pitch in a 2×2 factorial design, with factors `gender` and `context`. This is laboratory data measuring the voice pitch of male and female speakers (factor `gender`) in two different kinds of linguistic contexts, namely a polite and an informal situation (factor `context`).

We load the data, inspect and plot it.

```

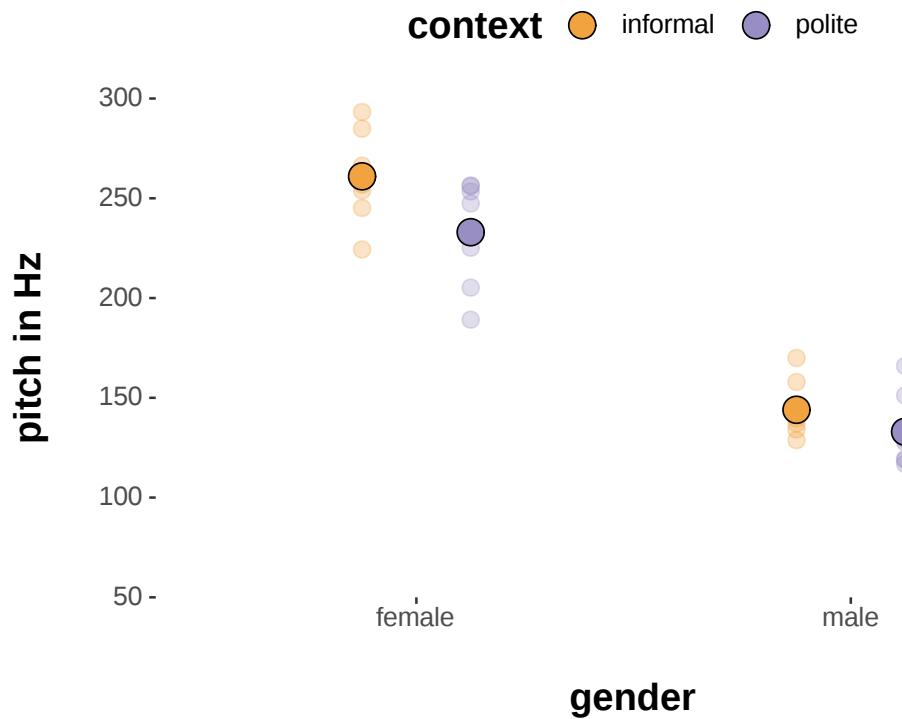
politeness_url <- url('https://raw.githubusercontent.com/michael-franke/intro-data-analysis/main/data/politeness.csv')
politeness_data <- read_csv(politeness_url))

```

```

## # A tibble: 83 x 5
##   subject gender sentence context  pitch
##   <chr>    <chr>  <chr>    <chr>   <dbl>
## 1 F1       F      S1       pol     213.
## 2 F1       F      S1       inf     204.
## 3 F1       F      S2       pol     285.
## 4 F1       F      S2       inf     260.
## 5 F1       F      S3       pol     204.
## 6 F1       F      S3       inf     287.
## 7 F1       F      S4       pol     251.
## 8 F1       F      S4       inf     277.
## 9 F1       F      S5       pol     232.
## 10 F1      F      S5       inf     252.
## # ... with 73 more rows

```



In a 2×2 factorial design like this, there are essentially four pairs of factor levels (so-called **design cells**): female speakers in informal contexts, female speakers in polite contexts, male speakers in informal contexts and male speakers in polite contexts. Different schemes exists by means of which different comparisons of means of design cells (or single factors) can be probed. A simple coding scheme for differences in our 2×2 design is shown in Figure 14.1. We consider the cell “female+informal” as the reference level and so model its mean as intercept β_0 . We then have a slope term β_{pol} which encodes the difference between female pitch in informal and female pitch in polite contexts. Similarly for β_{male} . Finally, we also include a so-called **interaction term**, denoted as $\beta_{\text{pol}\&\text{male}}$ in Figure 14.1. The interaction term quantifies how much a change away from the reference level in both variables differs from the sum of unilateral changes.

We can fit a regression model with this coding scheme using the formula `pitch ~ gender * context`. Importantly the star `*` between explanatory variables `gender` and `context` indicates that we also want to include the interaction term.

```
fit_brms_politeness <- brm(
  # model 'pitch' as a function of 'gender' and 'context',
  # also including the interaction between `gender` and `context` 
  formula = pitch ~ gender * context,
  data = politeness_data
)
```

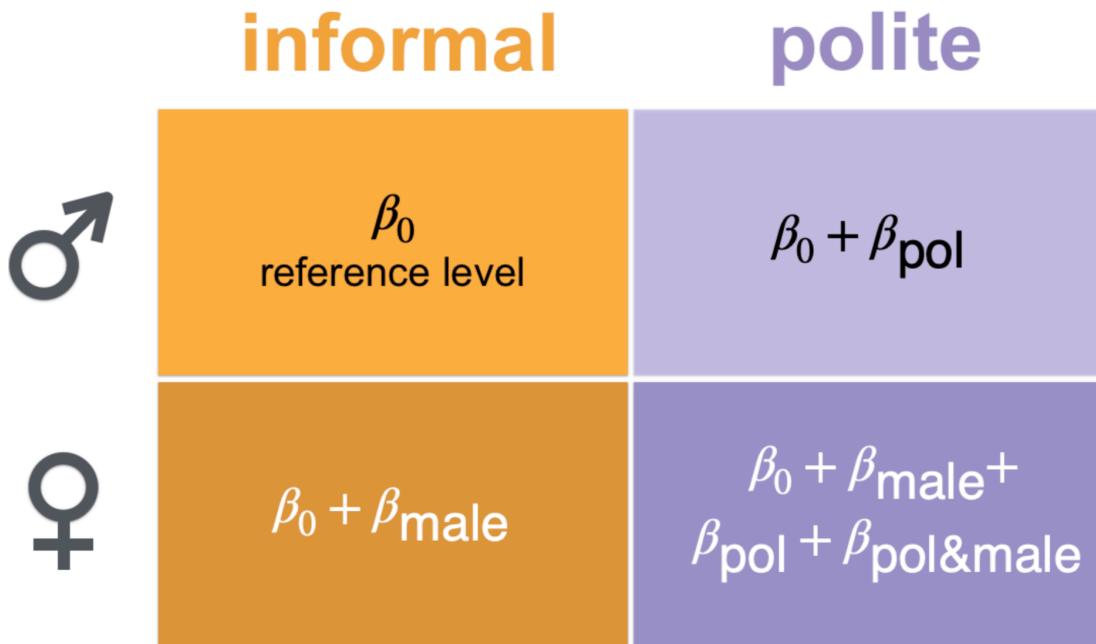


Figure 14.1.: Regression coefficients for a factorial design (using so-called 'treatment coding').

The summary statistics below lists the model parameters indicated in Figure 14.1.

```
summary(fit_brms_politeness)$fixed[,c("Estimate", "1-95% CI", "u-95% CI")]

##                                     Estimate   1-95% CI   u-95% CI
## Intercept                   261.02993 245.65872 276.800178
## genderM                    -116.53009 -138.41661 -93.764509
## contextpol                  -27.69013  -49.44973  -4.936747
## genderM:contextpol       16.16356  -15.79220  46.823184
```

We could conclude from this that, given model and data, it is plausible to think that male speakers had lower voices than female speakers in informal contexts: this shows in the exclusion of 0 in the 95% interquantile range for parameter `genderM`. We may also conclude that, given model and data, it is plausible to think that female speakers used lower voices in polite contexts than in formal ones (`parameter contextpol`). The posterior of the interaction term `genderM:contextpol` gives no indication to think that 0, or any value near it, is not plausible. This can be interpreted as saying that there is no indication, given model and data, to believe that male speakers' voice pitch changes differently from informal to polite contexts than female speakers' voice pitch does.

->

A. Further useful material

A.1. Material on *Introduction to Probability*:

- “Introduction to Probability” by J.K. Blitzstein and J. Hwang (Blitzstein and Hwang 2014)
- “Probability Theory: The Logic of Science” by E.T. Jaynes (Jaynes 2003)

A.2. Material on *Bayesian Data Analysis*:

- “Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan” by J. Kruschke (Kruschke 2015)
- “Baysian Data Analysis” by A. Gelman et al. (Gelman et al. 2013)
- “Statistical Rethinking: A Bayesian Course with Examples in R and Stan” by R. McElreath (McElreath 2015)
 - webbook based on McElreath’s book: Statistical Rethinking with brms, ggplot2, and the tidyverse by Solomon Kurz

A.3. Material on *frequentist statistics*:

- “Statistics for LinguistsL: An introduction using R”, by B. Winter (Winter 2019-)

A.4. Material on *R, tidyverse, etc.*:

- official R manual: An Introduction to R
- “R for Data Science: Import, Tidy, Transform, Visualize, and Model Data” by H. Wickham and G. Grolemund (Wickham and Grolemund 2016)
- RStudio’s Cheat Sheets
- “Data Visualization” by K. Healy (Healy 2018)
- webbook Learning Statistics with R by Danielle Navarro
- webbook with focus on visualization: Data Science for Psychologists by Hansjörg Neth

A. *Further useful material*

A.5. Further information for RStudio

- *Keyboard shortcuts* for Windows and Mac in RStudio: “Tools -> Keyboard Shortcuts Help” or also on the RStudio support site

A.6. Resources on WebPPL

- official website
- documentation
- short introduction tutorial
- Bayesian Data Analysis using Probabilistic Programs: Statistics as pottery by webbook on BDA with WebPPL by MH Tessler

B. Common probability distributions

This chapter summarizes common probability distributions, which occur at central places in this book.

B.1. Selected continuous distributions of random variables

B.1.1. Normal distribution

One of the most important distribution families is the *gaussian* or *normal family* because it fits many natural phenomena. Furthermore the sampling distributions of many estimators depend on the normal distribution. On the one hand because they are derived from normally distributed random variables or on the other hand because they can be asymptotically approximated by a normal distribution for large samples (*Central limit theorem*).

Distributions of the normal family are symmetric with range $(-\infty, +\infty)$ and have two parameters μ and σ that are referred to, respectively, as the *mean* and the *standard deviation* of the normal random variable. These parameters are examples of *location* and *scale* parameters. The normal distribution is located at μ and its width is scaled by choice of σ . The distribution is symmetric with most observations lying around the central peak μ and more extreme values are further away depending on σ .

$$X \sim \text{Normal}(\mu, \sigma^2)$$

Fig.~B.1 shows the probability density function of three normal distributed random variables with different parameters. Fig.~B.2 shows the corresponding cumulative function of the three normal distributions.

```
rv_normal <- tibble(
  x = seq(from = -15, to = 15, by = .01),
  y1 = dnorm(x),
  y2 = dnorm(x, mean = 2, sd = 2),
  y3 = dnorm(x, mean = -2, sd = 3)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(0,1)",
```

B. Common probability distributions

```

        parameter == "y2" ~ "(2,2)",
        parameter == "y3" ~ "(-2,3)")
    )

ggplot(rv_normal, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Normal", y = "Density")

```

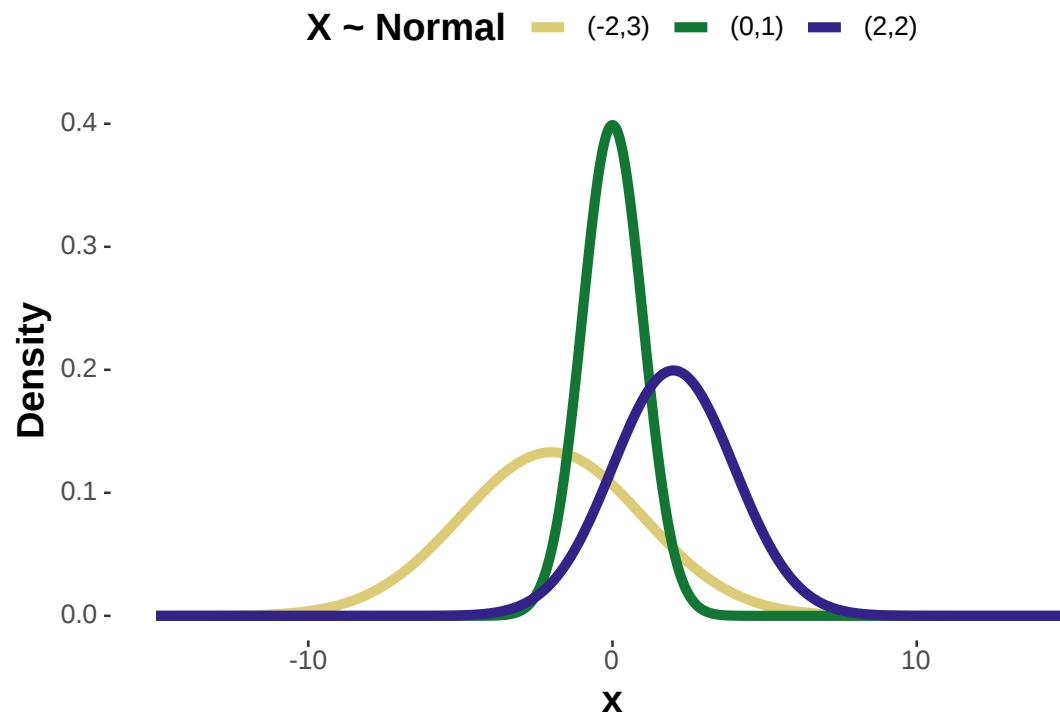


Figure B.1.: Examples of probability density function of normal distributions.

```

rv_normal %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Normal", y = "y")

```

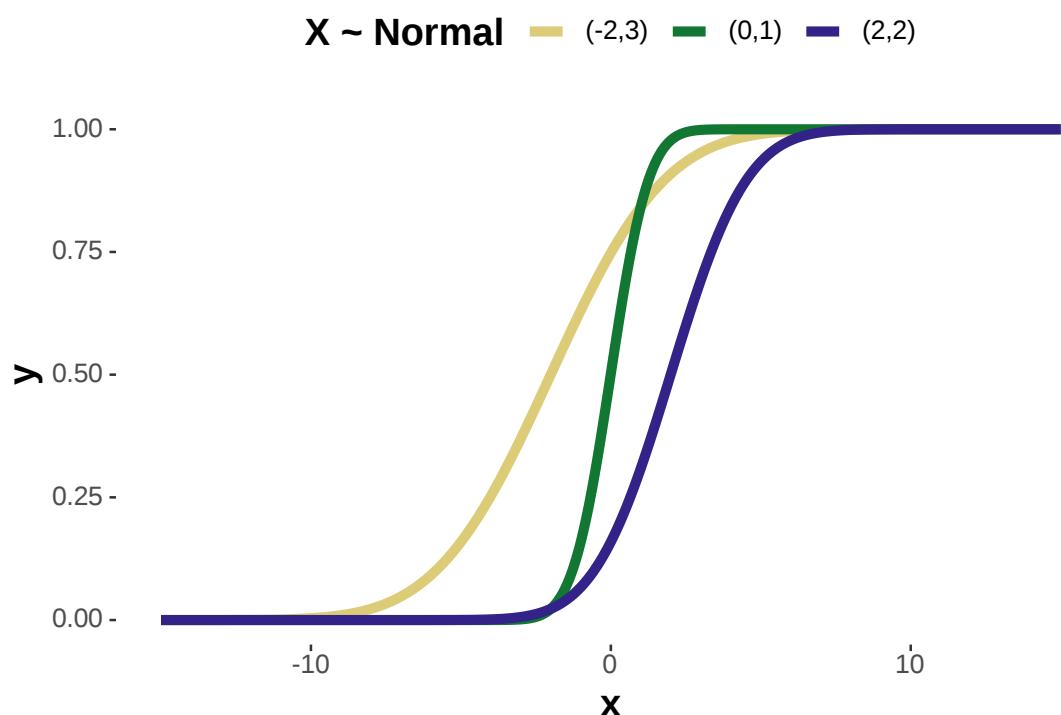


Figure B.2.: Examples of the cumulative distribution function of normal distributions corresponding to the previous probability density functions.

B. Common probability distributions

A special case of normal distributed random variables is the *standard normal* distributed variable with $\mu = 0$ and $\sigma = 1$: $Y \sim Normal(0, 1)$. Each normal distribution can be converted into a standard normal distribution by *z-standardization* (see equation below). The advantage of standardization is that values from different scales can be compared, because they become *scale independent* by z-transformation.

Probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Cumulative distribution function

$$F(x) = \int_{-\infty}^x f(t)dt$$

Expected value $E(X) = \mu$

Variance $Var(X) = \sigma^2$

Z-transformation $Z = \frac{X-\mu}{\sigma}$

Deviation and *Coverage The normal distribution is often associated with the *68-95-99.7 rule*. The values refer to the probability of a random data point landing within *one*, *two* or *three* standard deviations of the mean (Fig.~B.3 depicts these three intervals). For example, about 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean μ .

- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6827$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$

```
# plot normal distribution with intervals
ggplot(NULL, aes(x = c(-10, 10))) +
  # plot area under the curve
  stat_function(fun = dnorm, args = list(mean = 0, sd = 2),
                geom = "area",
                fill = project_colors[1],
                xlim = c(-6, 6)) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 2),
                geom = "area",
                fill = project_colors[2],
                xlim = c(-4, 4)) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 2),
                geom = "area",
                fill = project_colors[3],
```

```

xlim = c(-2, 2)) +
# plot the curve
stat_function(fun = dnorm, args = list(mean = 0, sd = 2),
              geom = "line",
              xlim = c(-10, 10),
              size = 2) +
# scale x-axis
xlim(-10, 10) +
# label x-axis
xlab("X") +
# label ticks of x-axis
scale_x_continuous(breaks = c(-6,-4,-2,0,2,4,6),
                   labels = c(expression(-3~sigma),expression(-2~sigma),
                             expression(-sigma),"0",expression(sigma),
                             expression(2~sigma),expression(3~sigma)))

```

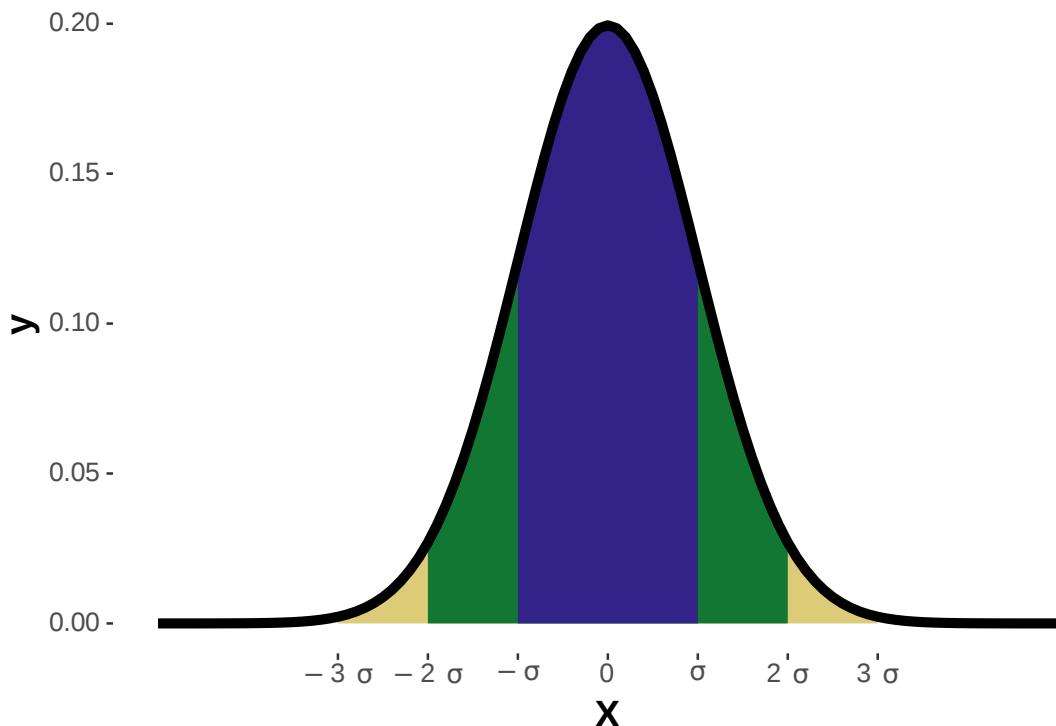


Figure B.3.: Coverage of normal distribution

Linear transformations

1. If $X \sim Normal(\mu, \sigma^2)$ is linear transformed by $Y = a*X + b$, then the new random variable is again

B. Common probability distributions

- normal distributed with $Y \sim Normal(a\mu + b, a^2\sigma^2)$.
2. Are $X \sim Normal(\mu_x, \sigma^2)$ and $Y \sim Normal(\mu_y, \sigma^2)$ normal distributed and independent, then their sum is again normal distributed with $X + Y \sim Normal(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

B.1.1.1. Hands On

```
knitr::include_app("https://istats.shinyapps.io/NormalDist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.1.2. Chi-squared distribution

The χ^2 -distribution is widely used in hypothesis testing in inferential statistics, because many test statistics are approximately distributed as χ^2 -distribution.

The χ^2 -distribution is directly related to the standard normal distribution: The sum of n independent and standard normal distributed random variables X_1, X_2, \dots, X_n is distributed according to a χ^2 distribution with n degrees of freedom:

$$Y = X_1^2 + X_2^2 + \dots + X_n^2.$$

The χ^2 distribution is a skew probability distribution with range $[0, +\infty)$ and only one parameter: n , the degrees of freedom: (If $n = 1$, then $(0, +\infty)$).

$$X \sim \chi^2(n).$$

Fig.~B.4 shows the probability density function of three chi-squared distributed random variables with different values for the parameter. Notice, that with increasing degrees of freedom the chi-squared distribution approximates the normal distribution. For $n \geq 30$ the chi-squared distribution can be approximated by a normal distribution. Fig.~B.5 shows the corresponding cumulative function of the three chi-squared density distributions.

```
rv_chisq <- tibble(
  x = seq(from = 0, to = 20, by = .01),
  y1 = dchisq(x, df = 2),
  y2 = dchisq(x, df = 4),
  y3 = dchisq(x, df = 9)
) %>%
  pivot_longer(cols = starts_with("y"),
```

```

  names_to  = "parameter",
  values_to = "y") %>%
mutate(
  parameter = case_when(parameter == "y1" ~ "(2)",
                         parameter == "y2" ~ "(4)",
                         parameter == "y3" ~ "(9)")
)

# dist plot
ggplot(rv_chisq, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Chi-Squared", y = "Density")

```

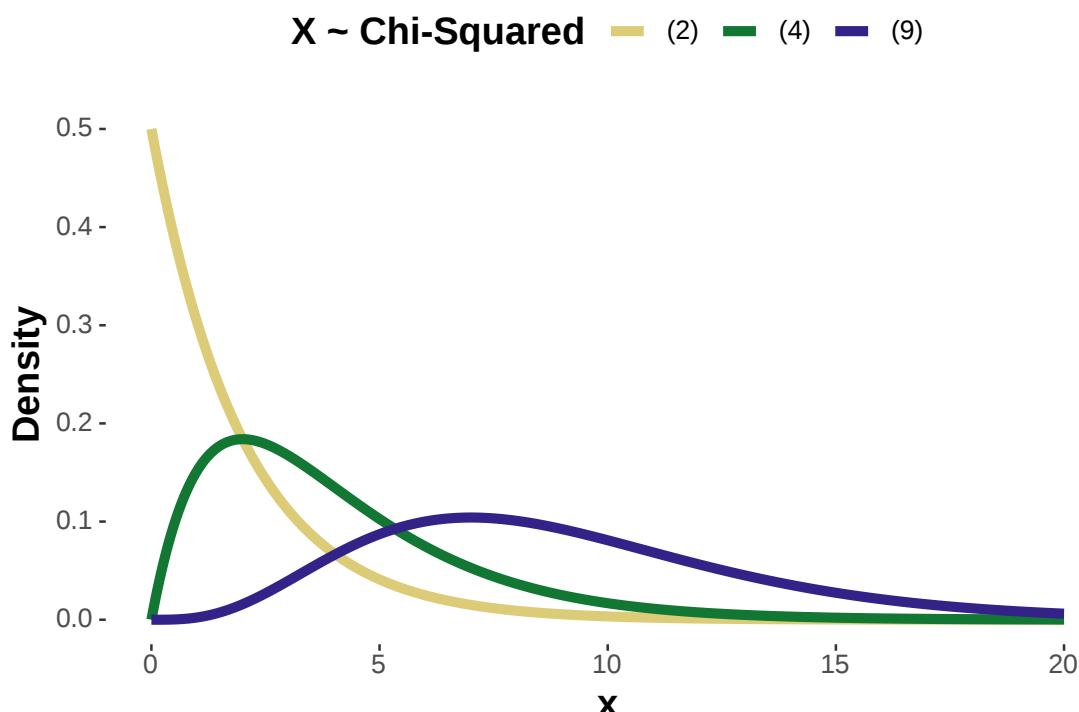


Figure B.4.: Examples of probability density function of chi-squared distributions.

```

rv_chisq %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%

```

B. Common probability distributions

```
ungroup() %>%
ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Chi-Squared", y = "y")
```

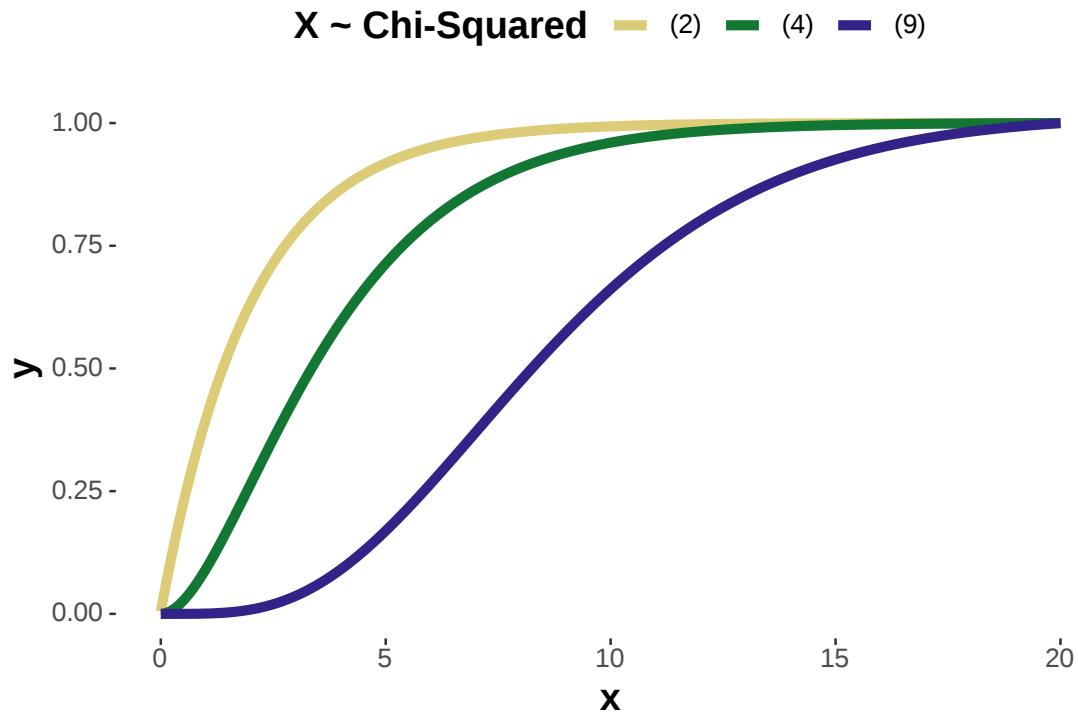


Figure B.5.: Examples of the cumulative distribution function of chi-squared distributions corresponding to the previous probability density functions.

Probability density function

$$f(x) = \begin{cases} \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Where $\Gamma(\frac{n}{2})$ denotes the Gamma function.

Cumulative distribution function

$$F(x) = \frac{\gamma(\frac{n}{2}, \frac{x}{2})}{\Gamma(\frac{n}{2})}$$

with $\gamma(s, t)$ being the lower incomplete gamma function:

$$\gamma(s, t) = \int_0^t t^{s-1} e^{-t} dt.$$

Expected value $E(X) = n$

Variance $Var(X) = 2n$

Transformations The sum of two χ^2 -distributed random variables $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$ is again a chi^2 -distributed random variable $X + Y = \chi^2(m + n)$.

B.1.2.1. Hands On

```
knitr:::include_app("https://istats.shinyapps.io/ChisqDist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.1.3. F distribution

The F distribution, named after R.A. Fisher, is used in particular in regression and variance analysis. It is defined by the ratio of two chi^2 -distributed random variables $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$, each divided by its degree of freedom:

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}}.$$

The F distribution is a continuous skew probability distribution with range $(0, +\infty)$ and two parameters m and n , corresponding to the degrees of freedom of the two chi^2 -distributed random variables:

$$X \sim F(m, n).$$

Fig.~B.6 shows the probability density function of three F distributed random variables with different parameter values. For a small number of degrees of freedom the density distribution is skewed to the left side. When the number increases, the density distribution gets more and more symmetric. Fig.~B.7 shows the corresponding cumulative function of the three F density distributions.

B. Common probability distributions

```

rv_F <- tibble(
  x = seq(from = 0, to = 7, by = .01),
  y1 = df(x, df1 = 2, df2 = 4),
  y2 = df(x, df1 = 4, df2 = 6),
  y3 = df(x, df1 = 12, df2 = 12)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(2,4)",
                           parameter == "y2" ~ "(4,6)",
                           parameter == "y3" ~ "(12,12)")
  )

# dist plot
ggplot(rv_F, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ F", y = "Density")

rv_F %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ F", y = "y")

```

Probability density function

$$F(x) = m^{\frac{m}{2}} n^{\frac{n}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(mx + n)^{\frac{m+n}{2}}} \text{ for } x > 0.$$

Where $\Gamma(x)$ denotes the gamma function.

Cumulative distribution function

$$F(x) = I\left(\frac{m \cdot x}{m \cdot x + n}, \frac{m}{2}, \frac{n}{2}\right),$$

with $I(z, a, b)$ being the regularized incomplete beta function:

$$I(z, a, b) = \frac{1}{B(a, b)} \cdot \int_0^z t^{a-1} (1-t)^{b-1} dt.$$

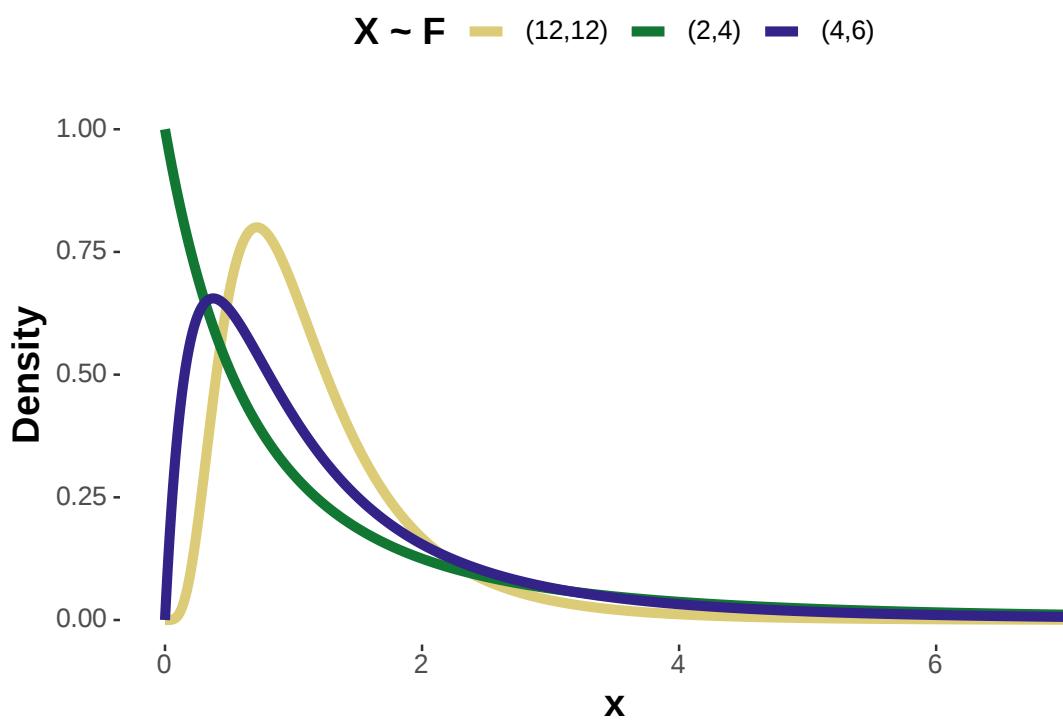


Figure B.6.: Examples of probability density function of F distributions.

B. Common probability distributions

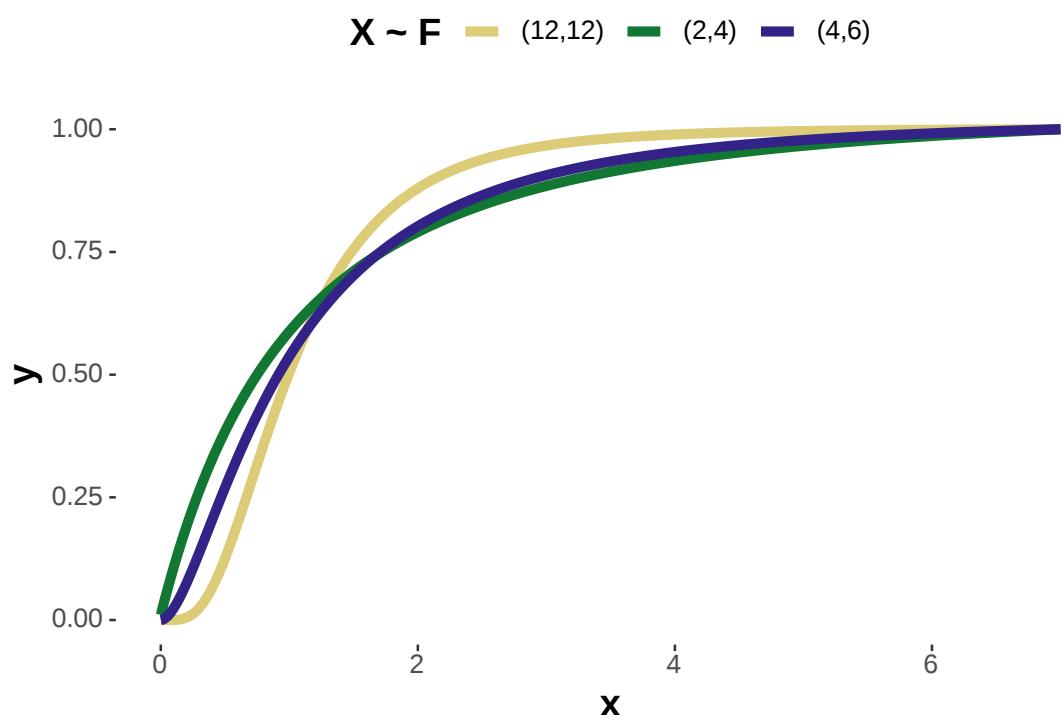


Figure B.7.: Examples of the cumulative distribution function of F distributions corresponding to the previous probability density functions.

$$\text{Expected value } E(X) = \frac{n}{n-2} \text{ (for } n \geq 3\text{)}$$

$$\text{Variance } Var(X) = \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2} \text{ (for } n \geq 5\text{)}$$

B.1.3.1. Hands On

```
knitr::include_app("https://istats.shinyapps.io/FDist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.1.4. Student's *t*-distribution

The *t* or Student's *t* distribution was discovered by William S. Gosset in 1908 (Vallverdú 2016), who published his work under the pseudonym "Student". He worked at the Guinness factory and had to deal with the problem of small sample sizes, where using a normal distribution as an approximation can be too crude. To overcome this problem Gosset conceived of the *t* distribution. Accordingly, this distribution is used in particular when the sample size is small and the variance unknown, which is often the case in reality. Its shape resembles the normal bell shape and has a peak at zero, but the *t* distribution is a bit lower and wider (bigger tails) than the normal distribution.

The *t* distribution consists of a standard-normally distributed random variable $X \sim \text{Normal}(0, 1)$ and a χ^2 -distributed random variable $Y \sim \chi^2(n)$ (X and Y are independent):

$$T = \frac{X}{\sqrt{Y/n}}.$$

The *t* distribution has range $(-\infty, +\infty)$ and one parameter ν , the degrees of freedom. The degrees of freedom can be calculated by the sample size n minus one:

$$t \sim \text{Student-}t(\nu = n - 1).$$

Fig.~B.8 shows the probability density function of three *t*-distributed random variables with different parameters, and Fig.~B.9 shows the corresponding cumulative function of the three *t* density distributions. Notice that for small degrees of freedom ν , the *t*-distribution has bigger tails. This is because the *t* distribution was specially designed to provide more conservative test results when analyzing small samples. When the degrees of freedom increases, the *t*-distribution approaches a normal distribution. For $\nu \geq 30$ this approximation is quite good.

```
rv_student <- tibble(
  x = seq(from = -6, to = 6, by = .01),
  y1 = dt(x, df = 1),
```

B. Common probability distributions

```
y2 = dt(x, df = 2),
y3 = dt(x, df = 10)
) %>%
pivot_longer(cols = starts_with("y"),
             names_to = "parameter",
             values_to = "y") %>%
mutate(
  parameter = case_when(parameter == "y1" ~ "(1)",
                         parameter == "y2" ~ "(2)",
                         parameter == "y3" ~ "(10)")
)

# dist plot
ggplot(rv_student, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ t", y = "Density")
```

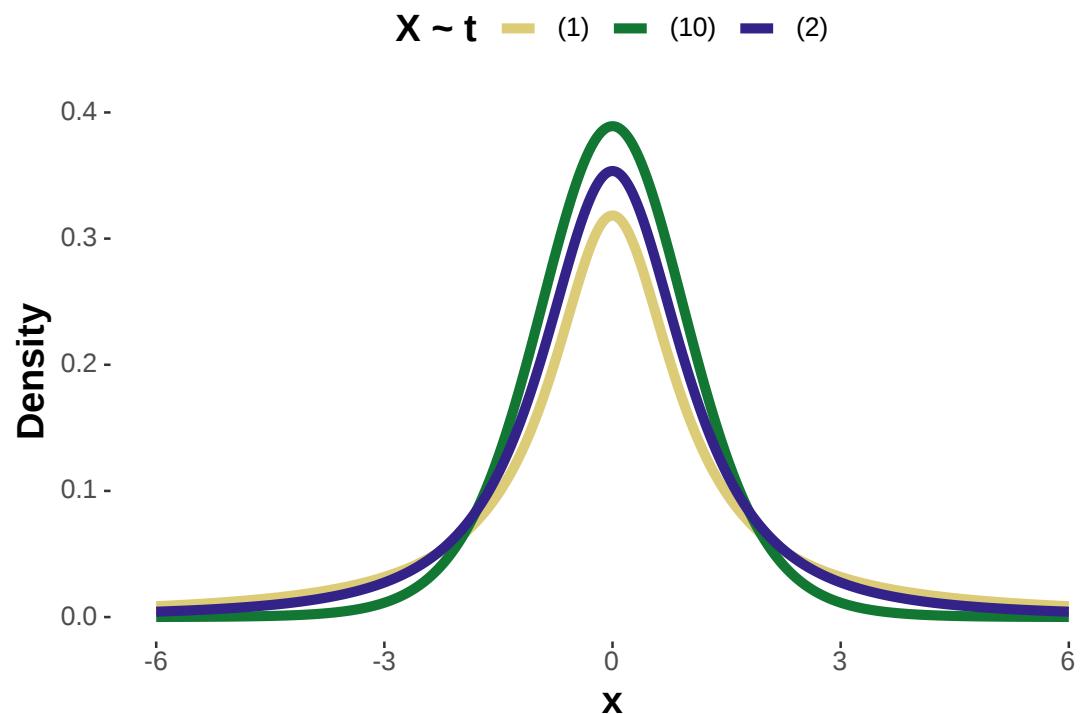


Figure B.8.: Examples of probability density functions of t distribution.

```
rv_student %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ t", y = "y")
```

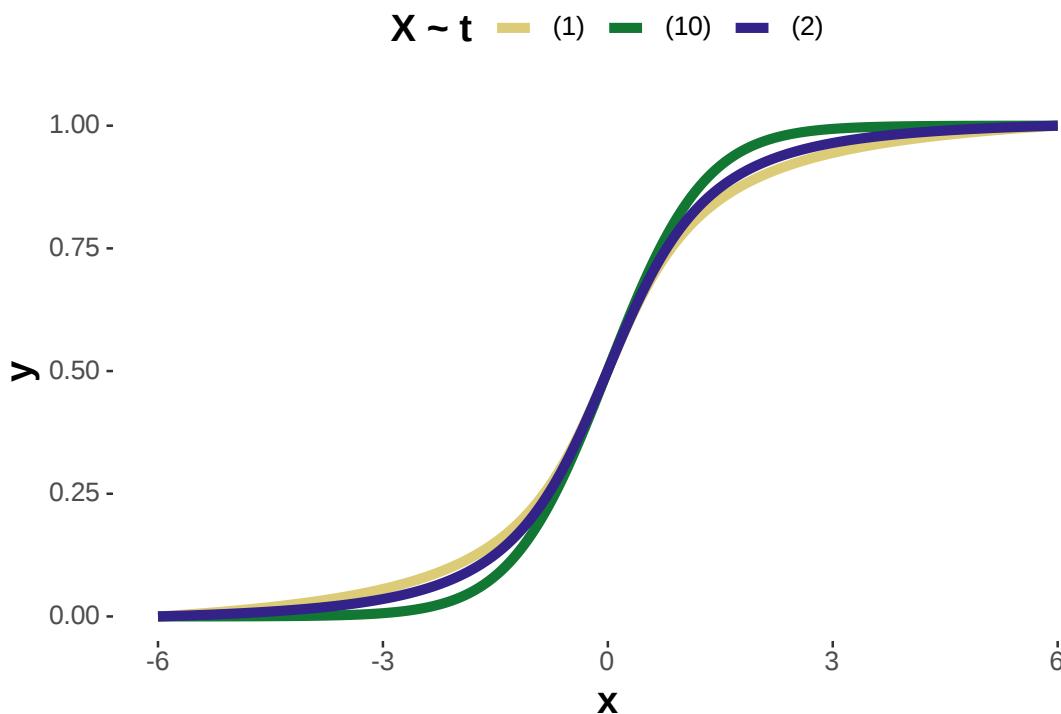


Figure B.9.: Examples of the cumulative distribution function of t distributions corresponding to the previous probability density functions.

Probability density function

$$f(x, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \cdot \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

with $\Gamma(x)$ denoting the gamma function.

B. Common probability distributions

Cumulative distribution function

$$F(x, \nu) = I\left(\frac{x + \sqrt{x^2 + \nu}}{2\sqrt{x^2 + \nu}}, \frac{\nu}{2}, \frac{\nu}{2}\right),$$

where $I(z, a, b)$ denotes the regularized incomplete beta function:

$$I(z, a, b) = \frac{1}{B(a, b)} \cdot \int_0^z t^{a-1} (1-t)^{b-1} dt.$$

Expected value $E(X) = 0$

Variance $Var(X) = \frac{n}{n-2}$ (for $n \geq 30$)

B.1.4.1. Hands On

```
knitr::include_app("https://istats.shinyapps.io/tdist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.1.5. Beta distribution

The beta distribution creates a continuous distribution of numbers between 0 and 1, therefore this distribution is useful if the uncertain quantity is bounded by 0 and 1 (or 100%), is continuous, and has a single mode. In Bayesian Data Analysis the beta distribution has a special standing as prior distribution for a bernoulli or binomial (see discrete distributions) likelihood. The reason for this is that a combination of a beta prior and a bernoulli (or binomial) likelihood results in a posterior distribution with the same form as the beta distribution. Such priors are referred to as *conjugate priors*.

A beta distribution has two parameters a and b :

$$X \sim Beta(a, b).$$

The two parameters can be interpreted as the number of observations made, such that: $n = a + b$. If a and b get bigger, the beta distribution gets narrower. If only a gets bigger the distribution moves rightward and if only b gets bigger the distribution moves leftward. Thus, the parameters define the shape of the distribution, therefore they are also called *shape parameters*. A Beta(1,1) is equivalent to a uniform distribution. Fig.~B.10 shows the probability density function of four beta distributed random variables with different parameter values. Fig.~B.11 shows the corresponding cumulative functions.

```
rv_beta <- tibble(
  x = seq(from = 0, to = 1, by = .01),
  y1 = dbeta(x, shape1 = 1, shape2 = 1),
  y2 = dbeta(x, shape1 = 4, shape2 = 4),
```

```

y3 = dbeta(x, shape1 = 4, shape2 = 2),
y4 = dbeta(x, shape1 = 2, shape2 = 4)
) %>%
pivot_longer(cols = starts_with("y"),
             names_to = "parameter",
             values_to = "y") %>%
mutate(
  parameter = case_when(parameter == "y1" ~ "(1,1)",
                         parameter == "y2" ~ "(4,4)",
                         parameter == "y3" ~ "(4,2)",
                         parameter == "y4" ~ "(2,4)")
)
# dist plot
ggplot(rv_beta, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Beta", y = "Density")

```

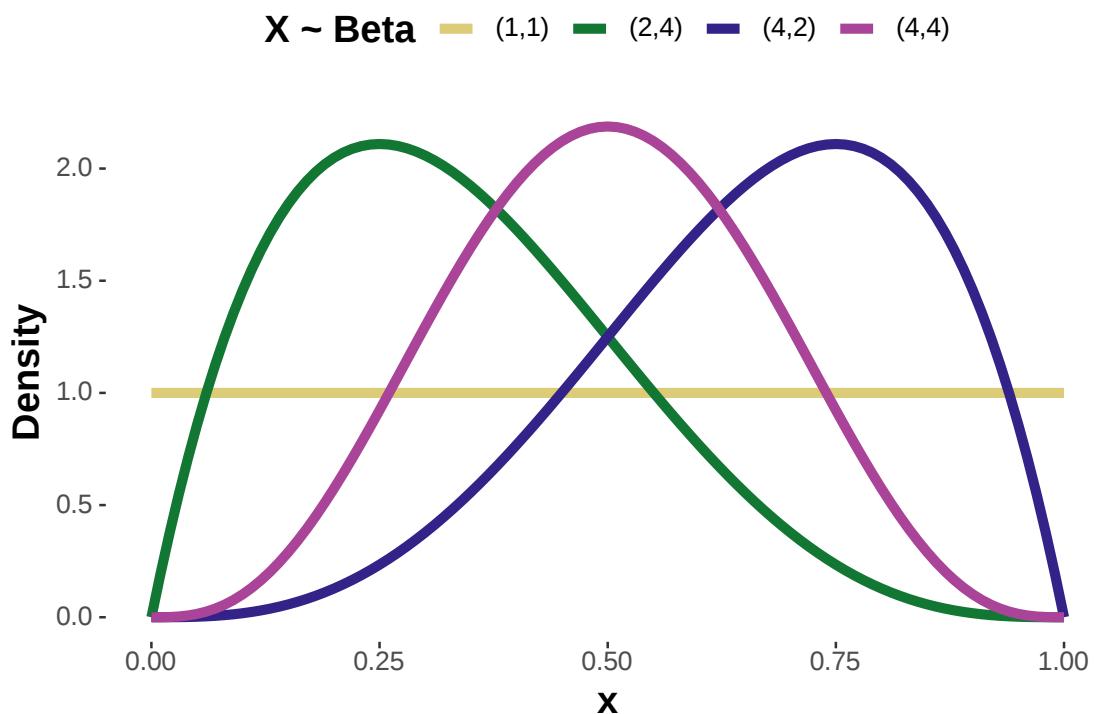


Figure B.10.: Examples of probability density function of beta distributions.

B. Common probability distributions

```
rv_beta %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Beta", y = "y")
```

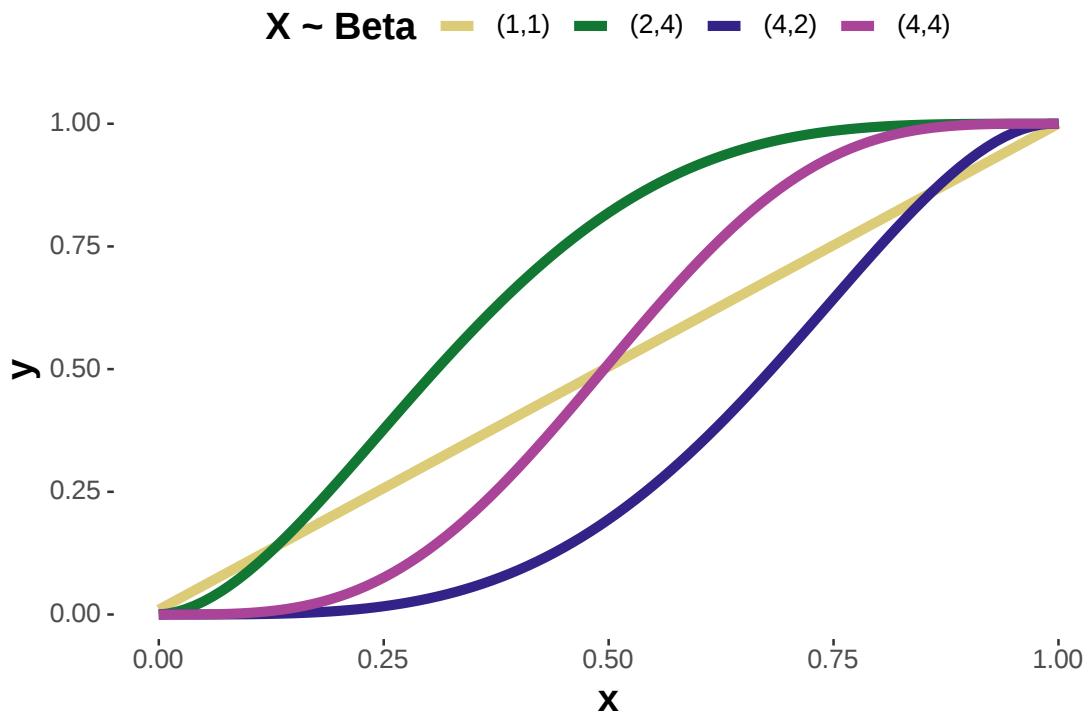


Figure B.11.: Examples of the cumulative distribution function of beta distributions corresponding to the previous probability density functions.

Probability density function

$$f(x) = \frac{\theta^{(a-1)}(1-\theta)^{(b-1)}}{B(a,b)},$$

where $B(a, b)$ is the beta function:

$$B(a, b) = \int_0^1 \theta^{(a-1)}(1-\theta)^{(b-1)} d\theta.$$

Cumulative distribution function

$$F(x) = \frac{B(x; a, b)}{B(a, b)},$$

where $B(x; a, b)$ is the *incomplete beta function*:

$$B(x; a, b) = \int_0^x t^{(a-1)}(1-t)^{(b-1)} dt,$$

and $B(a, b)$ the (complete) *beta function*

$$B(a, b) = \int_0^1 \theta^{(a-1)}(1-\theta)^{(b-1)} d\theta.$$

Expected value Mean: $E(X) = \frac{a}{a+b}$ Mode: $\omega = \frac{(a-1)}{a+b-2}$

Variance Variance: $Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$ Concentration: $\kappa = a + b$ (related to variance such that, the bigger a and b are, the narrower the distribution)

Reparameterization of the beta distribution Sometimes it is helpful (and more intuitive) to write the beta distribution in terms of its mode ω and concentration κ instead of a and b :

$$\text{Beta}(a, b) = \text{Beta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1), \text{ for } \kappa > 2.$$

B.1.6. Uniform distribution

The (continuous) uniform distribution takes values within a specified range a and b that have constant probability. Sometimes the distribution is also called rectangular distribution, due to its shape of a rectangle. The uniform distribution is in particular common for random number generation. In Bayesian Data Analysis it is often used as prior distribution to express *ignorance*. This can be thought in the following way: When different events are possible but no (reliable) information exists about their probability of occurrence, the most conservative (and also intuitive) choice would be to assign probability such that all events are equally likely to occur. The uniform distribution model this intuition, it generates a completely random number in some interval $[a, b]$.

B. Common probability distributions

The distribution is specified by two parameters: the end points a (minimum) and b (maximum):

$$X \sim Unif(a, b).$$

When $a = 0$ and $b = 1$ the distribution is referred to as *standard* uniform distribution. Fig.~B.12 shows the probability density function of two uniform distributed random variables with different parameter values.

Fig.~B.13 shows the corresponding cumulative functions.

```
rv_unif <- tibble(
  x = seq(from = -.1, to = 4.2, by = .01),
  y1 = dunif(x),
  y2 = dunif(x, min = 2, max = 4)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = ifelse(parameter == "y1",
                        "(0,1)", "(2,4)")
  )

# dist plot
ggplot(rv_unif, aes(x, y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Uniform", y = "Density")

rv_unif %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_line(size = 2) +
  labs(color = "X ~ Uniform", y = "y")
```

Probability density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

Cumulative distribution function

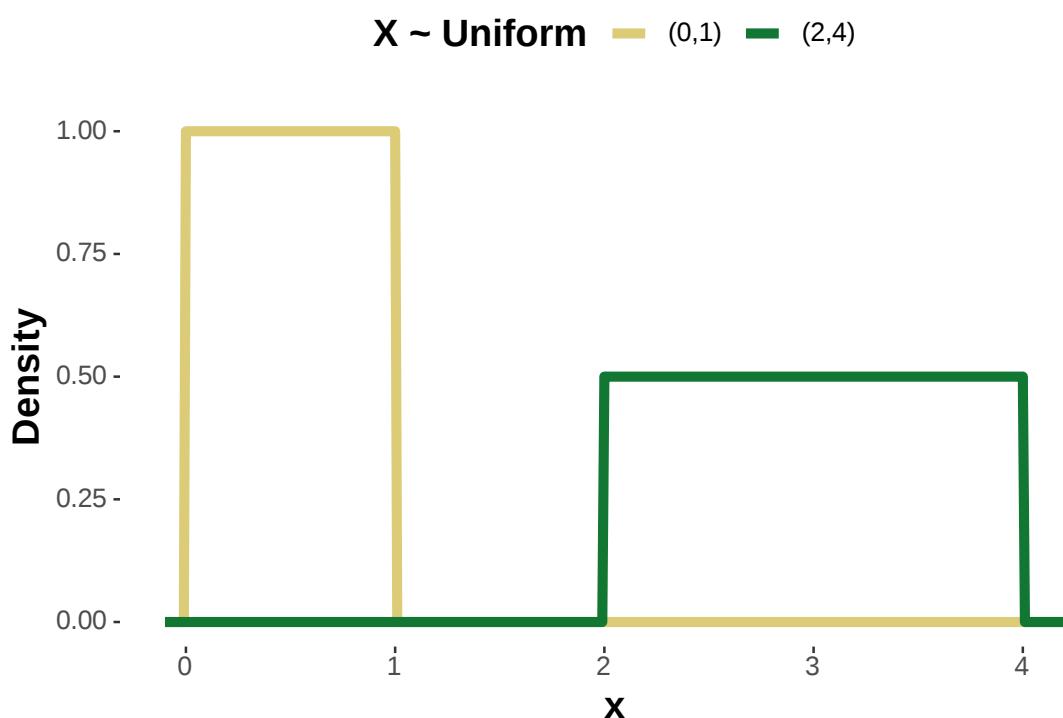


Figure B.12.: Examples of probability density function of uniform distributions.

B. Common probability distributions

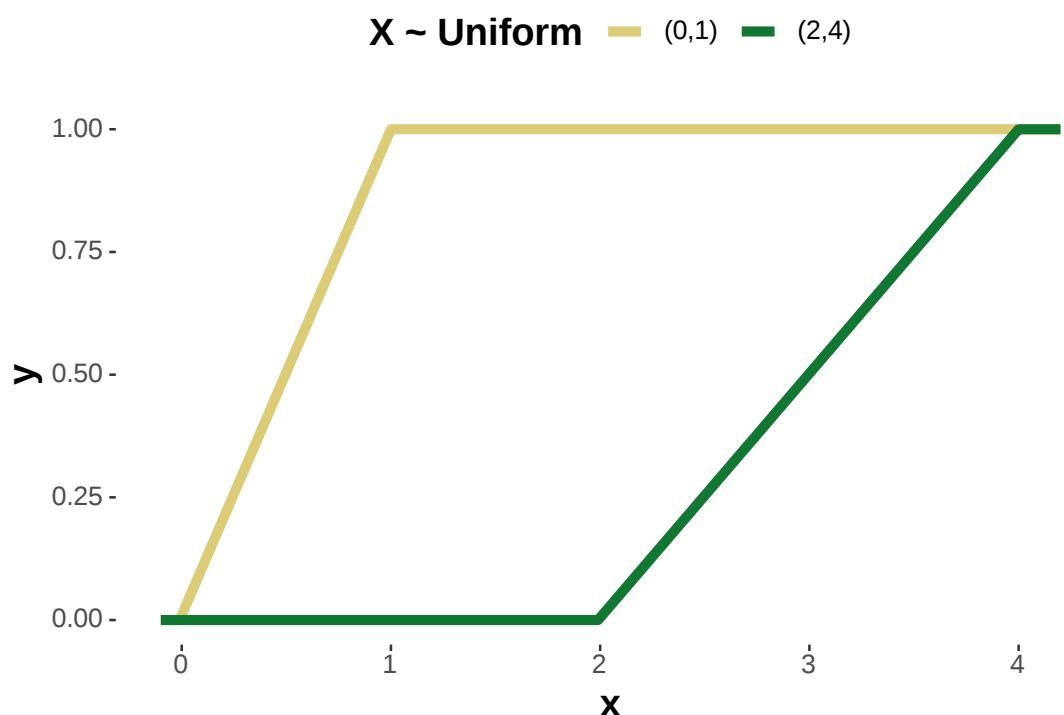


Figure B.13.: Examples of the cumulative distribution function of uniform distributions corresponding to the previous probability density functions.

$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x < b, \\ 1 & \text{for } x \geq b. \end{cases}$$

Expected value $E(X) = \frac{a+b}{2}$

Variance $Var(X) = \frac{(b-a)^2}{12}$

B.1.7. Dirichlet distribution

The Dirichlet distribution is a multivariate generalisation of the Beta distribution: While Beta distribution is a distribution over binomials, the Dirichlet is a distribution over Multinomials.

It can be used in any situation where an entity has to necessarily fall into one of $n + 1$ mutually exclusive subclasses, and the goal is to study the proportion of entities belonging to the different subclasses.

The Dirichlet distributions are commonly used as *prior distributions* in Bayesian statistics, as this family is a *conjugate prior* for the categorical distribution and the multinomial distribution.

The Dirichlet distribution $\mathcal{D}ir(\alpha)$ is a family of continuous multivariate probability distributions, parameterized by a vector α of positive reals. Thus, it is a distribution with k positive parameters α^k with respect to a k -dimensional space.

$$X \sim \mathcal{D}ir(\alpha)$$

The probability density function (see formula below) of the Dirichlet distribution for k random variables is a $k - 1$ dimensional probability *simplex* that exists on a k dimensional space. How does the parameter α influence the Dirichlet distribution?

- Higher values of α_i lead to greater “weight” of X_i and greater amount of the total “mass” assigned to it. (see plot 1)
- If $\alpha_1 = \dots = \alpha_k = 1$, then the points are uniformly distributed. (see plot 2)
- If all α_i are equal, the distribution is symmetric. (see plot 3 for asymmetric plot)
- Values of $\alpha_i < 1$ can be thought as anti-weight that pushes away x_i toward extremes. (see plot 4)

```
C <- matrix(c(10, 1, 2, 0.15, 10, 1, 10, 0.15, 10, 1, 5, 0.15), 4, 3)
```

```
f1 <- function(v) ddirichlet(v, C[1,])
f2 <- function(v) ddirichlet(v, C[2,])
f3 <- function(v) ddirichlet(v, C[3,])
f4 <- function(v) ddirichlet(v, C[4,])
```

B. Common probability distributions

```

mesh1 <- simplex_mesh(.0025) %>% as_tibble()
mesh2 <- simplex_mesh(.0025) %>% as_tibble()
mesh3 <- simplex_mesh(.0025) %>% as_tibble()
mesh4 <- simplex_mesh(.0025) %>% as_tibble()

mesh1$f1 <- mesh1 %>% apply(1, function(v) f1(bary2simp(v)))
mesh2$f2 <- mesh2 %>% apply(1, function(v) f2(bary2simp(v)))
mesh3$f3 <- mesh3 %>% apply(1, function(v) f3(bary2simp(v)))
mesh4$f4 <- mesh4 %>% apply(1, function(v) f4(bary2simp(v)))

points <- map_df(seq(nrow(C)), function(m){
  rdirichlet(250, C[m,]) %>%
    simp2bary() %>%
    as_tibble() %>%
    transmute(
      mesh = paste0("f", m),
      x = V1,
      y = V2
    )
}) %>%
  mutate(
    mesh = recode(mesh, f1 = "(10, 10, 10)",
                  f2 = "(1, 1, 1)",
                  f3 = "(2, 10, 5)",
                  f4 = "(.15, .15, .15)")
  )

meshes <- left_join(mesh1, mesh2, by = c("x", "y")) %>%
  left_join(mesh3, by = c("x", "y")) %>%
  left_join(mesh4, by = c("x", "y")) %>%
  pivot_longer(starts_with("f"), names_to = "mesh") %>%
  mutate(
    mesh = recode(mesh, f1 = "(10, 10, 10)",
                  f2 = "(1, 1, 1)",
                  f3 = "(2, 10, 5)",
                  f4 = "(.15, .15, .15)")
  )

ggplot(meshes, aes(x, y)) +
  geom_raster(aes(fill = value), show.legend = FALSE) +
  coord_equal(xlim = c(0,1), ylim = c(0, .85)) +
  geom_point(data = points, color = "orange", size = .3) +
  facet_wrap(~mesh)

```

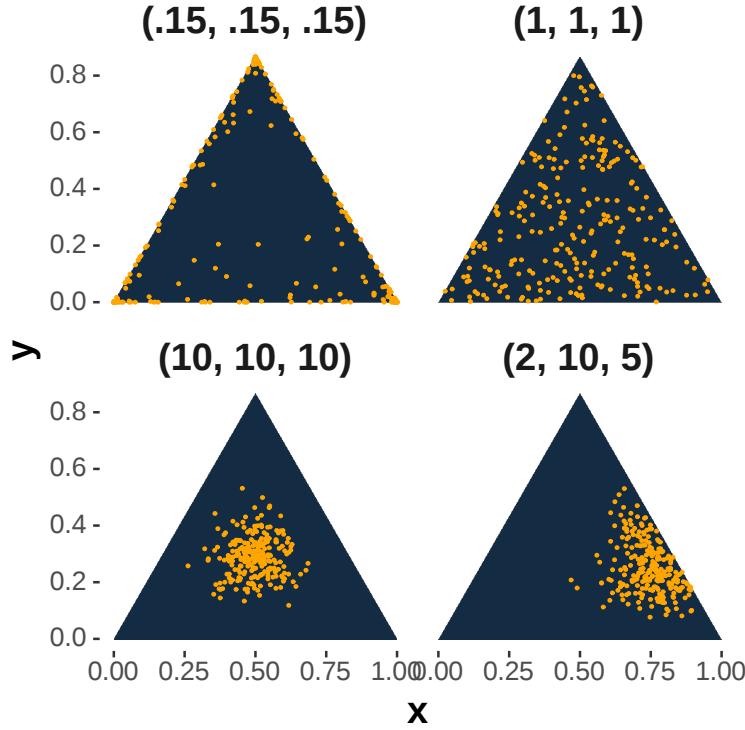


Figure B.14.: Examples of probability density function of dirichlet distributions.

Probability density function

$$f(x) = \frac{\Gamma(\sum_{i=1}^{n+1} \alpha_i)}{\prod_{i=1}^{n+1} \Gamma(\alpha_i)} \prod_{i=1}^{n+1} p_i^{\alpha_i - 1},$$

with *Gamma* denoting the gamma function and

$$p_i = \frac{X_i}{\sum_{j=1}^{n+1} X_j}, 1 \leq i \leq n,$$

where X_1, X_2, \dots, X_{n+1} are independent Gamma random variables, with $X_i \sim G(\alpha_i, 1)$.

Expected value $E(p_i) = \frac{\alpha_i}{t}$, with $t = \sum_{i=1}^{n+1} \alpha_i$

Variance $Var(p_i) = \frac{\alpha_i(t-\alpha_i)}{t^2(t+1)}$, with $t = \sum_{i=1}^{n+1} \alpha_i$

B.2. Selected discrete distributions of random variables

B.2.1. Binomial distribution

The binomial distribution is a useful model for binary decisions where the outcome is a choice between two alternatives (e.g. Yes/No, Left/Right, Present/Absent, Head/Tail, ...). The two outcomes are coded as 0 (failure) and 1 (success). Consequently, let the probability of occurrence of the outcome "success" be p , then the probability of occurrence of "failure" is $1 - p$. Consider a coin-flip experiment, with the outcomes "head" and "tail". If we flip a coin repeatedly, e.g. 30 times, the successive trials are independent of each other and the probability p is constant, then the resulting binomial distribution is a discrete random variable with outcomes $\{0, 1, 2, \dots, 30\}$.

The binomial distribution has two parameters "size" and "prob", often denoted as n and p , respectively.

The "size" refers to the number of trials and "prob" to the probability of success:

$$X \sim \text{Binomial}(n, p).$$

Fig.~B.15 shows the probability mass function of three binomial distributed random variables with different parameter values. As stated above, p refers to the probability of success. The higher this probability the more often we will observe the outcome coded with "1". Therefore the distribution tends toward the right side and vice-versa. The distribution gets more symmetrical if the parameter p approximates 0.5. Fig.~B.16 shows the corresponding cumulative functions.

```
# how many trials
trials = 30

rv_binom <- tibble(
  x = seq(0, trials),
  y1 = dbinom(x, size = trials, p = 0.2),
  y2 = dbinom(x, size = trials, p = 0.5),
  y3 = dbinom(x, size = trials, p = 0.8)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(n,0.2)",
                           parameter == "y2" ~ "(n,0.5)",
                           parameter == "y3" ~ "(n,0.8)")
  )

# dist plot
ggplot(rv_binom, aes(x, y, fill = parameter)) +
  geom_col(position = "identity", alpha = 0.8) +
  labs(fill = "X ~ Binomial", y = "Probability")
```

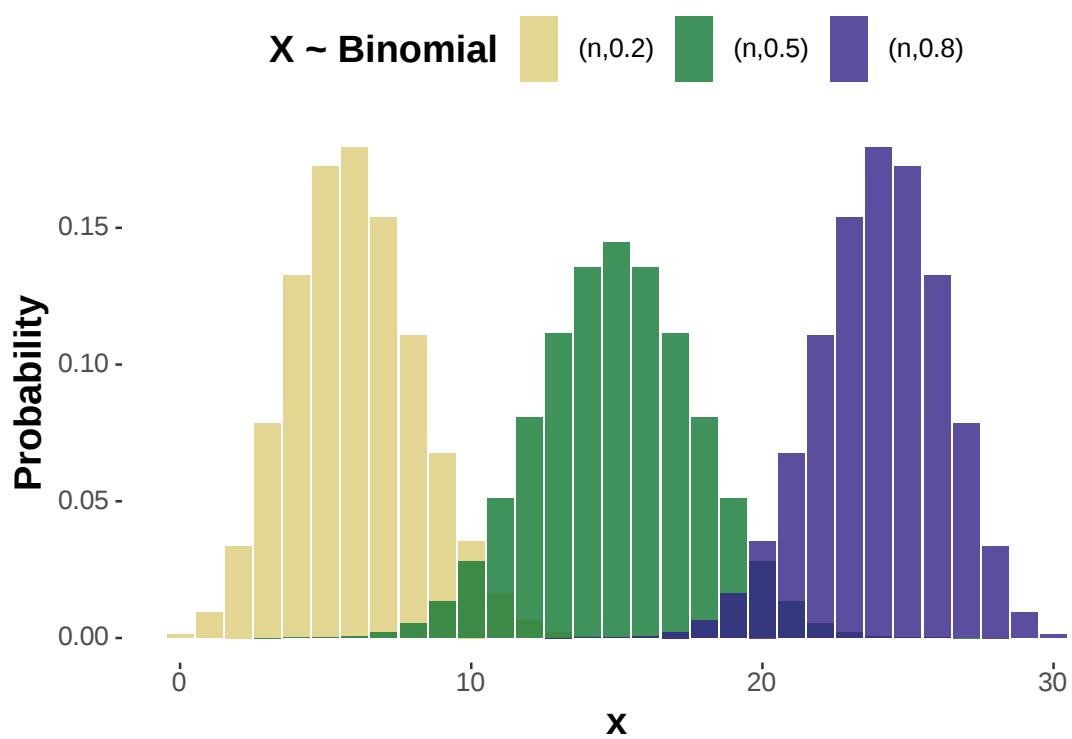


Figure B.15.: Examples of probability mass function of Binomial distributions.

B. Common probability distributions

```
rv_binom %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_step(size = 2) +
  labs(color = "X ~ Binomial", y = "y")
```

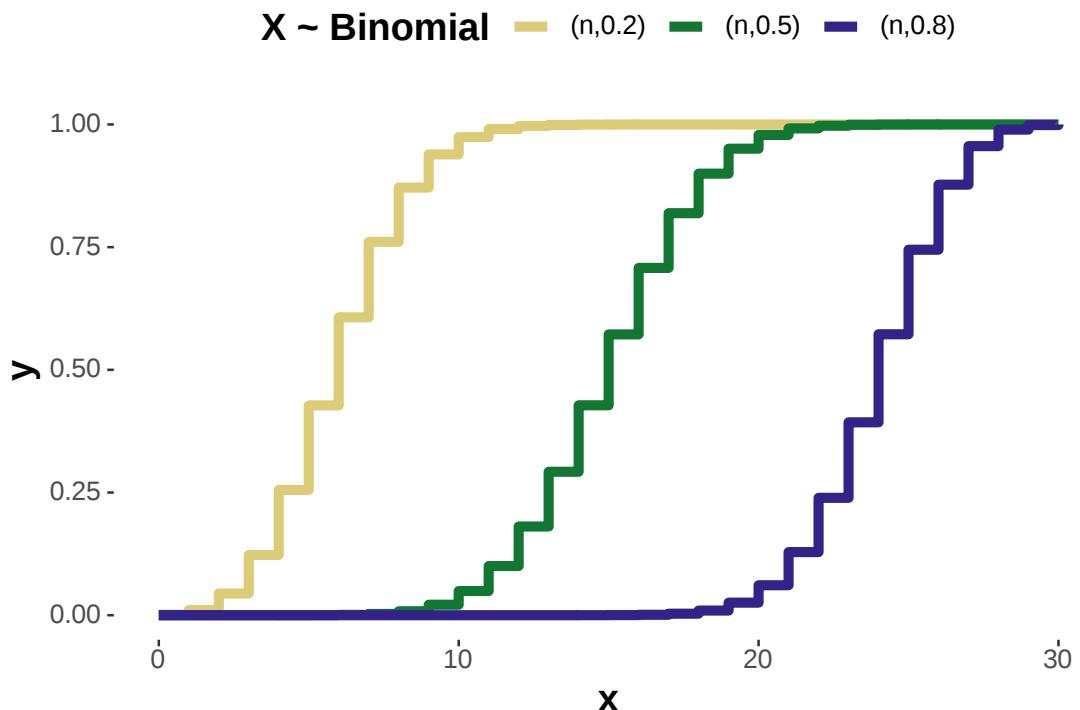


Figure B.16.: Examples of the cumulative distribution function of Binomial distributions corresponding to the previous probability mass functions.

Probability mass function

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $\binom{n}{x}$ is the binomial coefficient.

Cumulative function

$$F(x) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k}$$

Expected value $E(X) = n \cdot p$

Variance $Var(X) = n \cdot p \cdot (1 - p)$

B.2.1.1. Hands On

```
knitr::include_app("https://istats.shinyapps.io/BinomialDist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.2.2. Multinomial distribution

The Multinomial distribution is a generalization of the Binomial distribution to the case of n repeated trials: While the Binomial distribution can have two outcomes, the Multinomial distribution can have multiple outcomes. Consider an experiment where each trial can result in any of k possible outcomes with a probability p_i , where $(i = 1, 2, \dots, k)$, with $\sum_{i=1}^k p_i = 1$. For n repeated trials, let k_i denote the number of times $X = x_i$ was observed, where $i = 1, 2, \dots, m$. It follows that $\sum_{i=1}^m k_i = n$.

Probability mass function

The probability of observing a vector of outcomes $\mathbf{k} = [k_1, \dots, k_m]^T$ is

$$f(\mathbf{k}|\mathbf{p}) = \binom{n}{k_1 \cdot k_2 \cdot \dots \cdot k_m} \prod_{i=1}^m p_i^{k_i},$$

where $\binom{n}{k_1 \cdot k_2 \cdot \dots \cdot k_m}$ is the *multinomial coefficient*: $\binom{n}{k_1 \cdot k_2 \cdot \dots \cdot k_m} = \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_m!}$. It is a generalization of the *binomial coefficient* $\binom{n}{k}$.

Expected Values: $E(X) = n \cdot p_i$

Variance: $Var(X) = n \cdot p_i \cdot (1 - p_i)$

B.2.3. Bernoulli distribution

The Bernoulli distribution is a special case of the binomial distribution with $size = 1$, therefore the outcome of a bernoulli random variable is either 0 or 1. Apart from that the same information holds as for the binomial distribution. As the “size” parameter is now negligible, the bernoulli distribution has only one parameter, the probability of success p :

$$X \sim Bern(p).$$

Fig.~B.17 shows the probability mass function of three bernoulli distributed random variables with different parameters. Fig.~B.18 shows the corresponding cumulative distributions.

```
rv_bern <- tibble(
  x = seq(from = 0, to = 1),
  y1 = dbern(x, prob = 0.2),
  y2 = dbern(x, prob = 0.5),
  y3 = dbern(x, prob = 0.8)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(0.2)",
                           parameter == "y2" ~ "(0.5)",
                           parameter == "y3" ~ "(0.8)")
  )

# dist plot
ggplot(rv_bern, aes(x, y, fill = parameter)) +
  geom_col(position = "dodge", color = "white") +
  labs(fill = "X ~ Bernoulli", y = "Probability") +
  scale_x_continuous(breaks = c(0.0, 1.0), labels = c("0", "1"), limits = c(-0.5, 1.5))

rv_bern %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y),
    cum_y2 = cumsum(y)/sum(y)
  ) %>%
  add_column(
    x2 = c(1, 1, 1, 1.5, 1.5, 1.5)
  ) %>%
  ungroup() %>%
```

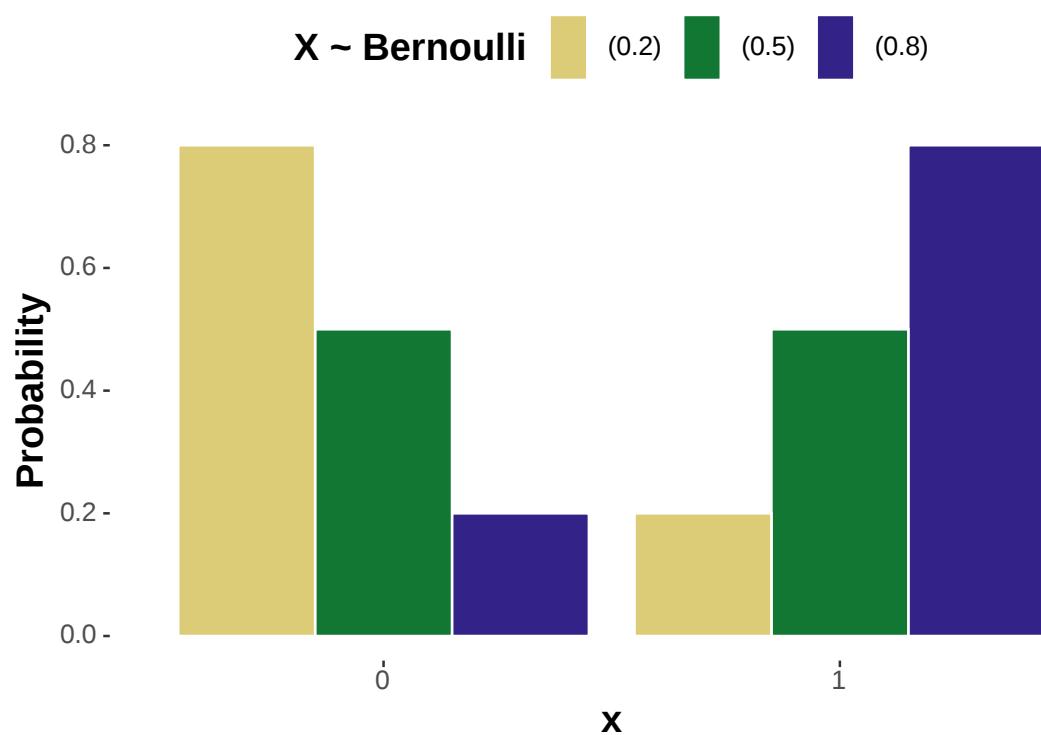


Figure B.17.: Examples of probability mass function of Bernoulli distributions.

B. Common probability distributions

```
ggplot(aes(x, cum_y, color = parameter)) +
  geom_segment(aes(xend = x2, yend = cum_y2), size = 1.5, linetype = "dashed") +
  geom_segment(aes(x = -0.5, y = 0, xend = 0.0, yend = 0), size = 1.5, linetype = "dashed") +
  geom_point(aes(x, cum_y), size = 4) +
  labs(color = "X ~ Bernoulli", y = "y") +
  scale_x_continuous(breaks = c(0.0, 1.0), labels = c("0", "1"), limits = c(-0.5, 1.5))
```

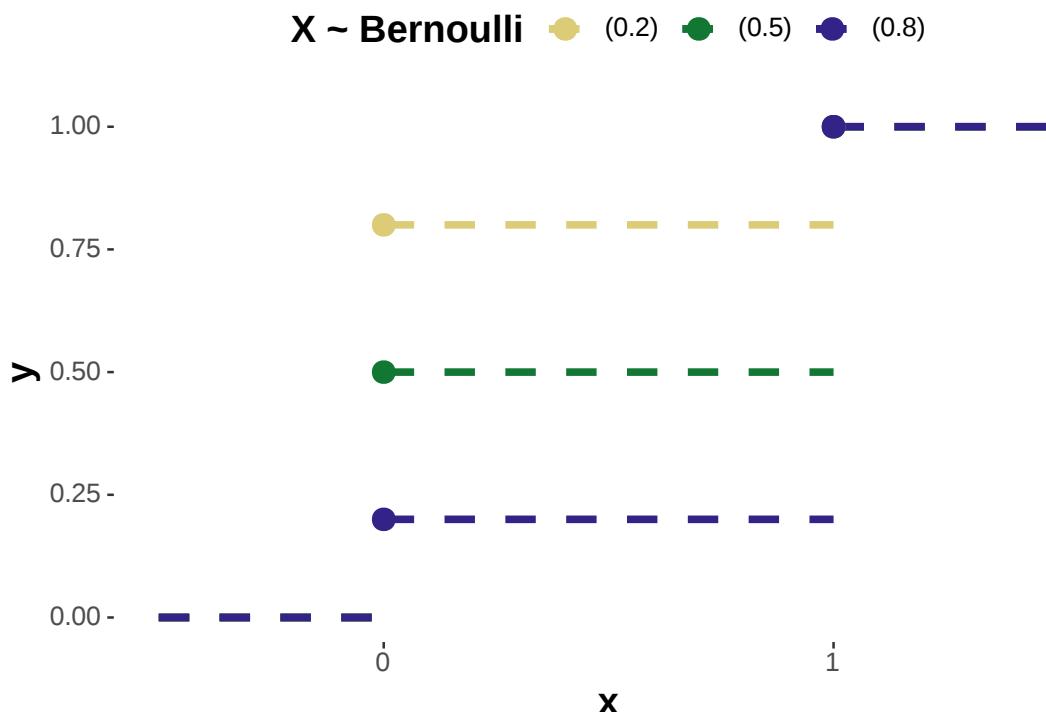


Figure B.18.: Examples of the cumulative distribution function of Bernoulli distributions corresponding to the previous probability mass functions.

Probability mass function

$$f(x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

Cumulative function

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1-p & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

Expected value $E(X) = p$

Variance $Var(X) = p \cdot (1 - p)$

B.2.4. Categorical distribution

The Categorical distribution, sometimes also referred to as *Multinoulli* distribution, is a generalization of the Bernoulli distribution for categorical random variables: While a Bernoulli distribution is a distribution over two alternatives, the Categorical is a distribution over multiple alternatives. For a single trial (e.g. a single die roll) the Categorical distribution is equal to the Multinomial distribution.

The Categorical distribution is parametrized by the probabilities assigned to each event. Let p_i the probability assigned to outcome i . The set of p_i 's are the parameters, and are constrained by $\sum_{i=1}^k p_i = 1$.

$$X \sim Cat(p)$$

```
rv_cat <- tibble(
  x = seq(from = 0, to = 20),
  y1 = dcat(x, prob = c(0.2,0.3,0.4,0.1)),
  y2 = dcat(x, prob = c(0.1,0.1,0.5,0.3)),
  y3 = dcat(x, prob = c(0.4,0.3,0.1,0.2))
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(0.2,0.3,0.4,0.1)",
                           parameter == "y2" ~ "(0.1,0.1,0.5,0.3)",
                           parameter == "y3" ~ "(0.4,0.3,0.1,0.2)")
  )

# dist plot
ggplot(rv_cat, aes(x, y, fill = parameter)) +
  geom_col(position = "dodge", color = "white") +
  labs(fill = "X ~ Categorical", y = "Probability") +
  scale_x_continuous(limits = c(0,5), breaks = c(1,2,3,4))
```

B. Common probability distributions

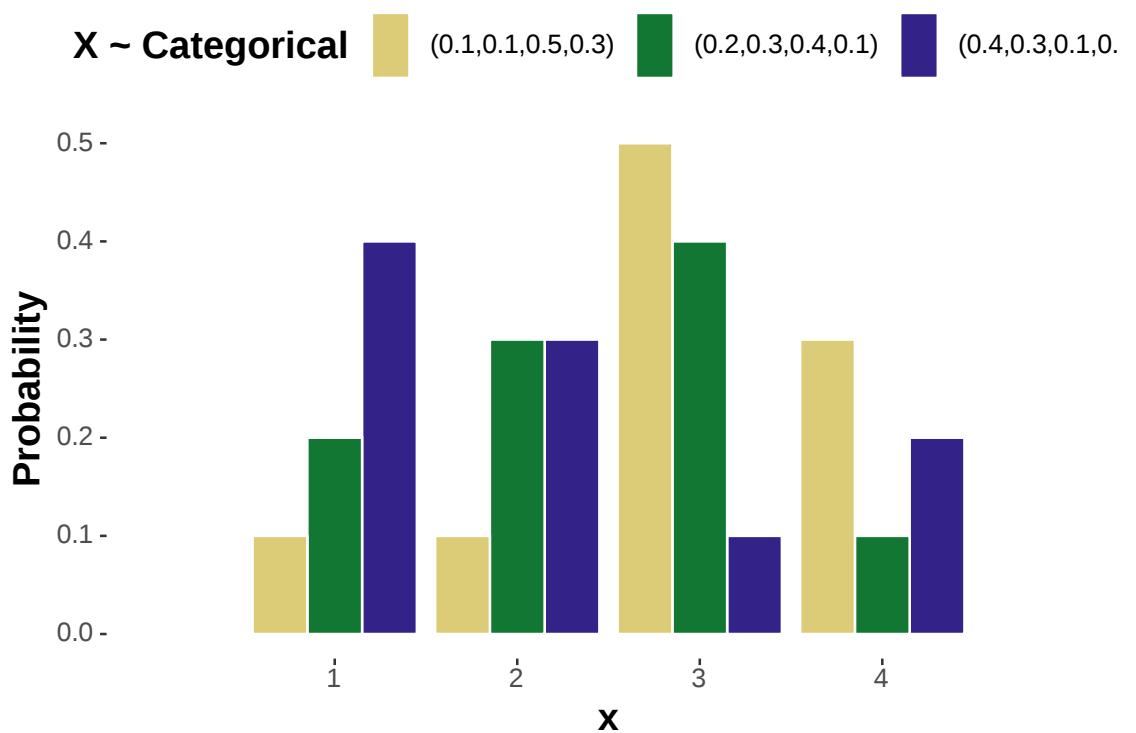


Figure B.19.: Examples of probability mass function of Categorical distributions.

```
rv_cat %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_step(size = 2) +
  labs(color = "X ~ Categorical", y = "y") +
  scale_x_continuous(limits = c(0,5), breaks = c(1,2,3,4))
```

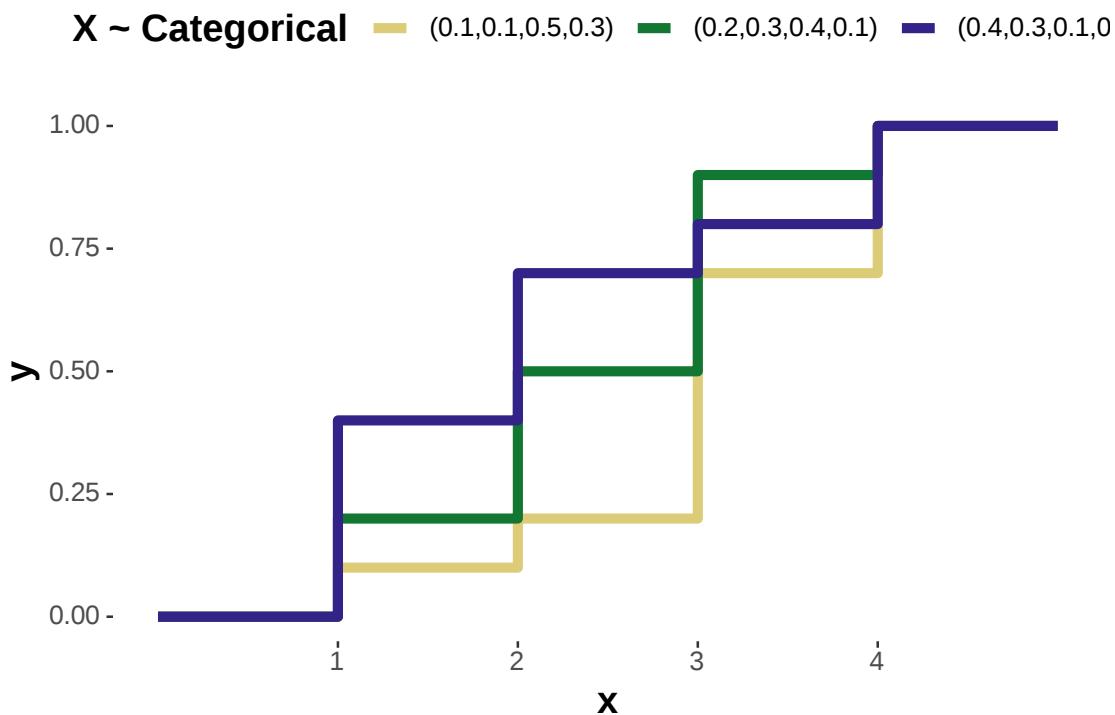


Figure B.20.: Examples of the cumulative distribution function of Categorical distributions corresponding to the previous probability mass functions.

Probability mass function

$$f(x|p) = \prod_{i=1}^k p_i^{\{x=i\}},$$

B. Common probability distributions

where $\{x = i\}$ evaluates to 1 if $x = i$, otherwise 0 and $p = p_1, \dots, p_k$ where p_i is the probability of seeing event i .

Expected Value: $E(x) = p$

Variance: $Var(x) = p \cdot (1 - p)$

B.2.5. Beta-Binomial distribution

The beta-binomial distribution, as the name already indicates, is a mixture of a binomial and beta distribution. Remember, a binomial distribution is useful to model a binary choice with outcomes "0" and "1". The binomial distribution has two parameters p , the probability of success ("1"), and n , the number of trials. Furthermore we assume that the successive trials are independent and p is constant. In a beta-binomial distribution p is not anymore assumed to be constant (or fixed) but changes from trial to trial. Thus, a further assumption about the distribution of p is made and here the beta distribution comes into play: the probability p is assumed to be randomly drawn from a beta distribution with parameters a and b . Therefore, the beta-binomial distribution has three parameters n, a and b :

$$X \sim BetaBinom(n, a, b).$$

For large values of a and b the distribution approaches a binomial distribution. When $a = 1$ and $b = 1$ the distribution equals a discrete uniform distribution from 0 to n . When $n = 1$, the distribution equals a bernoulli distribution.

Fig.~B.21 shows the probability mass function of three beta-binomial distributed random variables with different parameter values. Fig.~B.22 shows the corresponding cumulative distributions.

```
# how many trials
trials = 30

rv_betabinom <- tibble(
  x = seq(from = 0, to = trials),
  y1 = dbbinom(x, size = trials, alpha = 4, beta = 4),
  y2 = dbbinom(x, size = trials, alpha = 2, beta = 4),
  y3 = dbbinom(x, size = trials, alpha = 1, beta = 1)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(n,4,4)",
                           parameter == "y2" ~ "(n,2,4)",
                           parameter == "y3" ~ "(n,1,1)")
  )
```

```
# dist plot
ggplot(rv_betabinom, aes(x, y, fill = parameter)) +
  geom_col(position = "identity", alpha = 0.7) +
  labs(fill = "X ~ Beta-Binomial", y = "Probability")
```

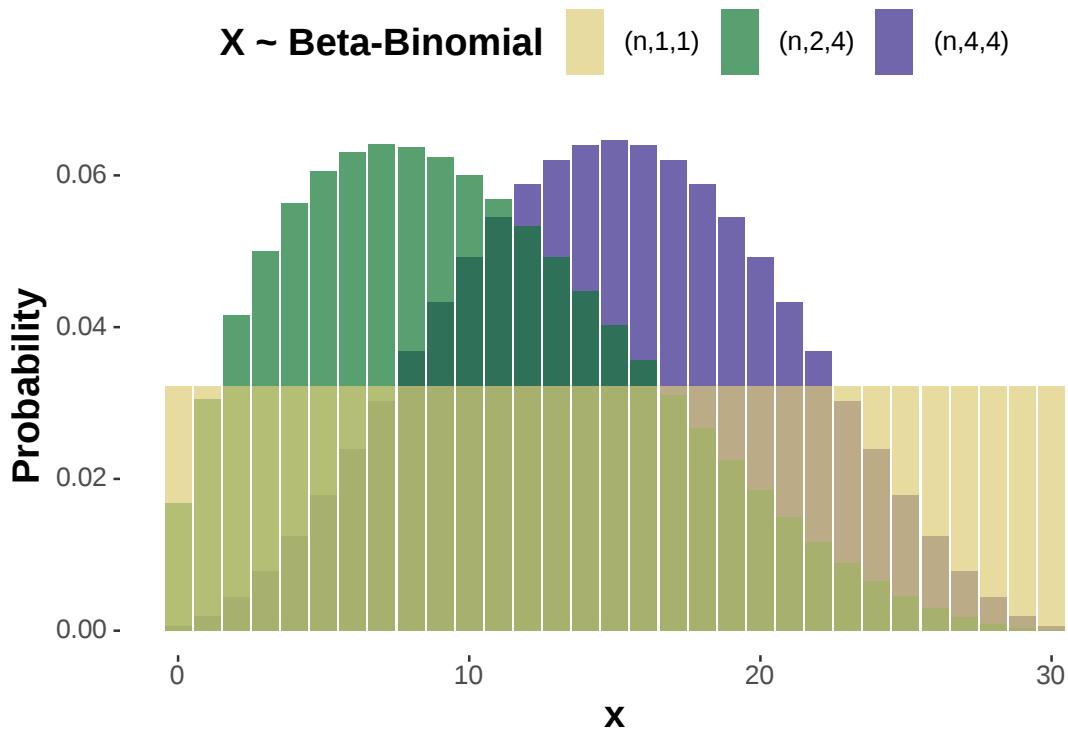


Figure B.21.: Examples of probability mass function of Beta-Binomial distributions.

```
rv_betabinom %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_step(size = 2) +
  labs(color = "X ~ Beta-Binomial", y = "y")
```

Probability mass function

B. Common probability distributions

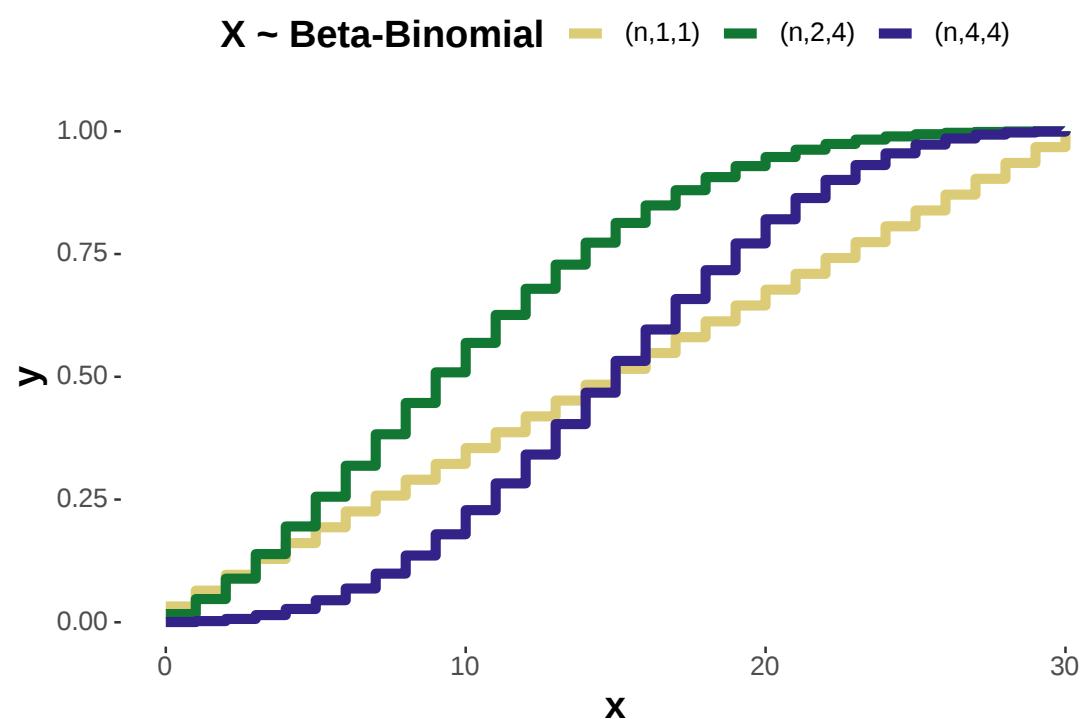


Figure B.22.: Examples of the cumulative distribution function of Beta-Binomial distributions corresponding to the previous probability mass functions.

$$f(x) = \binom{n}{x} \frac{B(a+x, b+n-x)}{B(a, b)},$$

where $\binom{n}{x}$ is the binomial coefficient and $B(x)$ the beta function (see beta distribution).

Cumulative function

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \binom{n}{x} \frac{B(a+x, b+n-x)}{B(a, b)} {}_3F_2(n, a, b) & \text{if } 0 \leq x < n, \\ 1 & \text{if } x \geq n. \end{cases}$$

Where ${}_3F_2(n, a, b)$ is the generalized hypergeometric function.

Expected value $E(X) = n \frac{a}{a+b}$

Variance $Var(X) = n \frac{ab}{(a+b)^2} \frac{a+b+n}{a+b+1}$

B.2.6. Poisson distribution

A poisson distributed random variable represents the number of successes occurring in a given *time interval*. It gives the probability of a given number of events happening in a fixed interval of time. The poisson distribution is a limiting case of the binomial distribution when the number of trials becomes very large and the probability of success is small. For example the number of car accidents in Osnabrueck in the next month, the number of typing errors on a page, the number of interruptions generated by a CPU during T seconds, etc. Events described by a poisson distribution must fullfill the following conditions: they occur in non-overlapping intervals, they can not occur simultaneously and each event occurs at a constant rate.

The poisson distribution has one parameter, the rate λ , sometimes also referred to as *intensity*:

$$X \sim Po(\lambda).$$

The parameter λ can be thought of as the expected number of events in the time interval. Consequently, changing the rate parameter changes the probability of seeing different numbers of events in one interval.

See Fig.~B.23 for the probability mass function of three poisson distributed random variables with different parameter values. Notice, that the higher λ the more symmetrical gets the distribution. In fact, the poisson distribution can be approximated by a normal distribution for a rate paramter ≥ 10 . Fig.~B.24 shows the corresponding cumulative distributions.

B. Common probability distributions

```

rv_pois <- tibble(
  x = seq(from = 0, to = 30, by = 1),
  y1 = dpois(x, lambda = 2),
  y2 = dpois(x, lambda = 8),
  y3 = dpois(x, lambda = 15)
) %>%
  pivot_longer(cols = starts_with("y"),
               names_to = "parameter",
               values_to = "y") %>%
  mutate(
    parameter = case_when(parameter == "y1" ~ "(2)",
                           parameter == "y2" ~ "(8)",
                           parameter == "y3" ~ "(15)")
  )

# dist plot
ggplot(rv_pois, aes(x, y, fill = parameter)) +
  geom_col(alpha = 0.7, position = "identity") +
  labs(fill = "X ~ Poisson", y = "Density")

# cumdist plot
rv_pois %>%
  group_by(parameter) %>%
  mutate(
    cum_y = cumsum(y)/sum(y)
  ) %>%
  ungroup() %>%
  ggplot(aes(x, cum_y, color = parameter)) +
  geom_step(size = 2) +
  labs(color = "X ~ Poisson", y = "y")

```

Probability mass function

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Cumulative function

$$F(x) = \sum_{k=0}^x \frac{\lambda^k}{k!} e^{-\lambda}$$

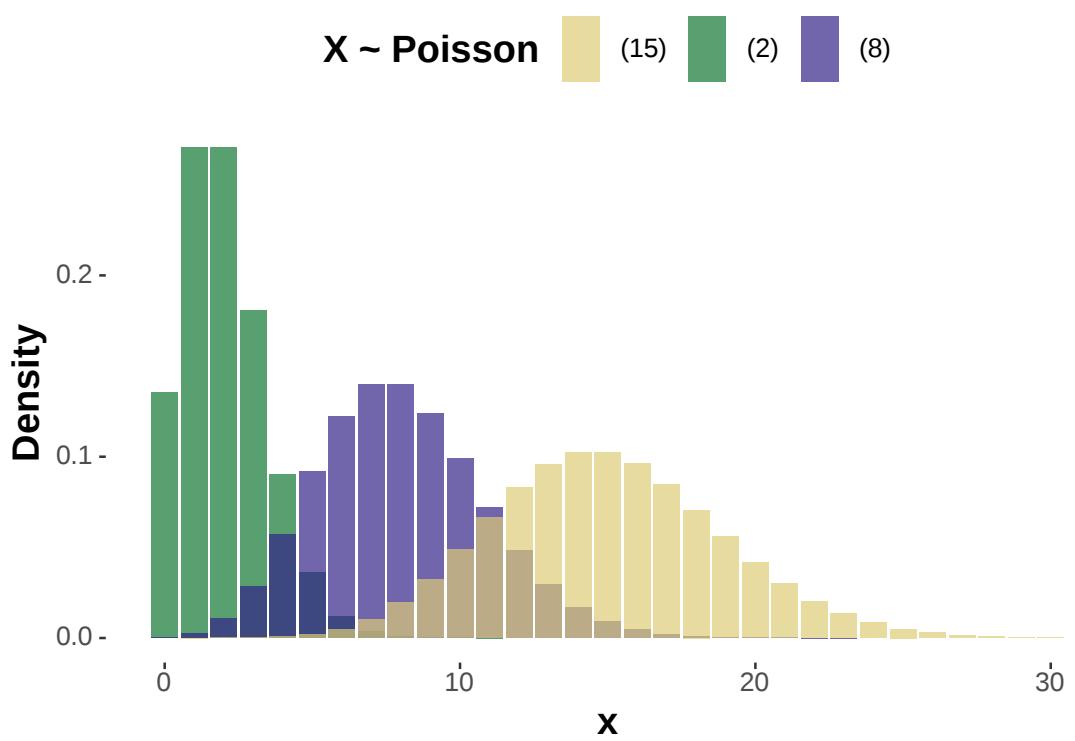


Figure B.23.: Examples of probability mass function of Poisson distributions.

B. Common probability distributions

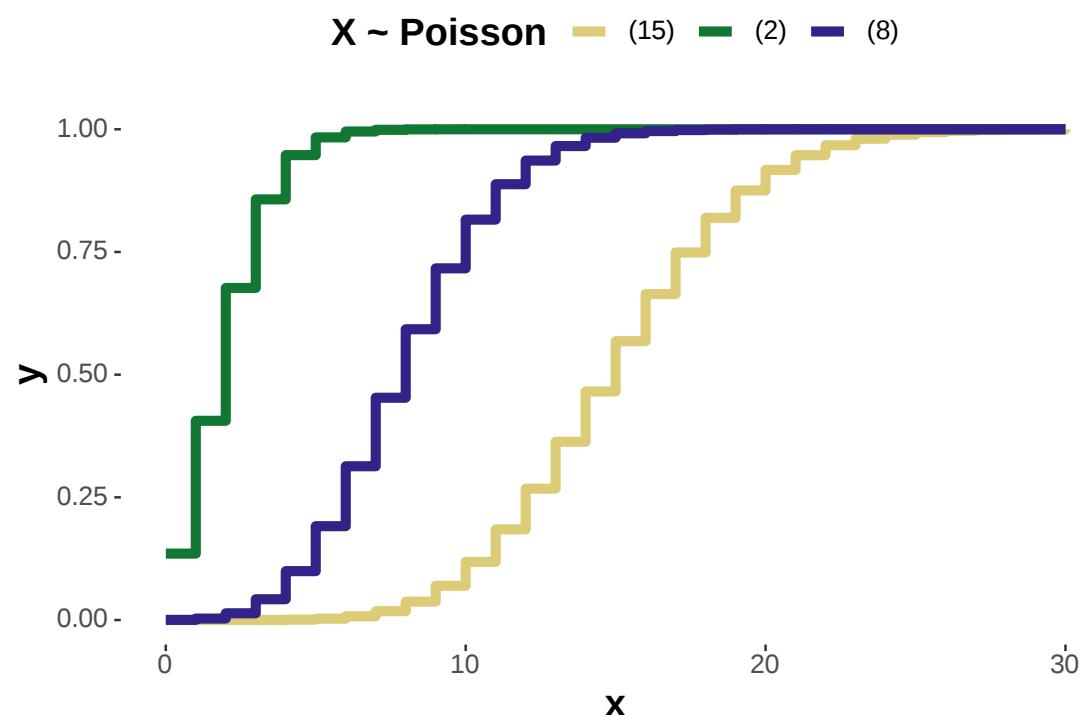


Figure B.24.: Examples of the cumulative distribution function of Poisson distributions corresponding to the previous probability mass functions.

Expected value $E(X) = \lambda$

Variance $Var(X) = \lambda$

B.2.6.1. Hands On

```
knitr::include_app("https://istats.shinyapps.io/PoissonDist/", height = "800px")
```

App taken from <http://www.artofstat.com/webapps.html> (Klingenberg, n.d.)

B.3. Understanding distributions as random variables

```
rv_normal_transform = tibble(
  x_axis = seq(from = -5, to = 5, by = .01),
  rv_norm_x = dnorm(x = x_axis, mean = 1, sd = 2),
  rv_norm_y = dnorm(x = x_axis, mean = 1.75, sd = 2.75),
  rv_linear_transform = 3*rv_norm_x+0.25,
  rv_summ_transform = rv_norm_x + rv_norm_y
) %>%
  pivot_longer(cols = starts_with("rv"),
               names_to = "random_variables",
               values_to = "y") %>%
  mutate(
    random_variables = case_when(random_variables == "rv_norm_x" ~ "X ~ N(1,2)",
                                 random_variables == "rv_norm_y" ~ "Y ~ N(1.75,2.75)",
                                 random_variables == "rv_linear_transform" ~ "3*X+0.25~ N(3*1+0.25,3^2*2.75)",
                                 random_variables == "rv_summ_transform" ~ "X+Y ~ N(1+1.75,2^2+2.75^2)"
    )
)

# dist plot
ggplot(rv_normal_transform, aes(x_axis, y, color = random_variables)) +
  geom_line(size = 2) +
  labs(color = "X,Y ~ Normal", y = "Density", x = "x")

rv_norm_transform = tibble(
  rv_norm_X = rnorm(n = 1e6),
  rv_norm_Y = rnorm(n = 1e6),
  rv_norm_Z = rnorm(n = 1e6),
```

B. Common probability distributions

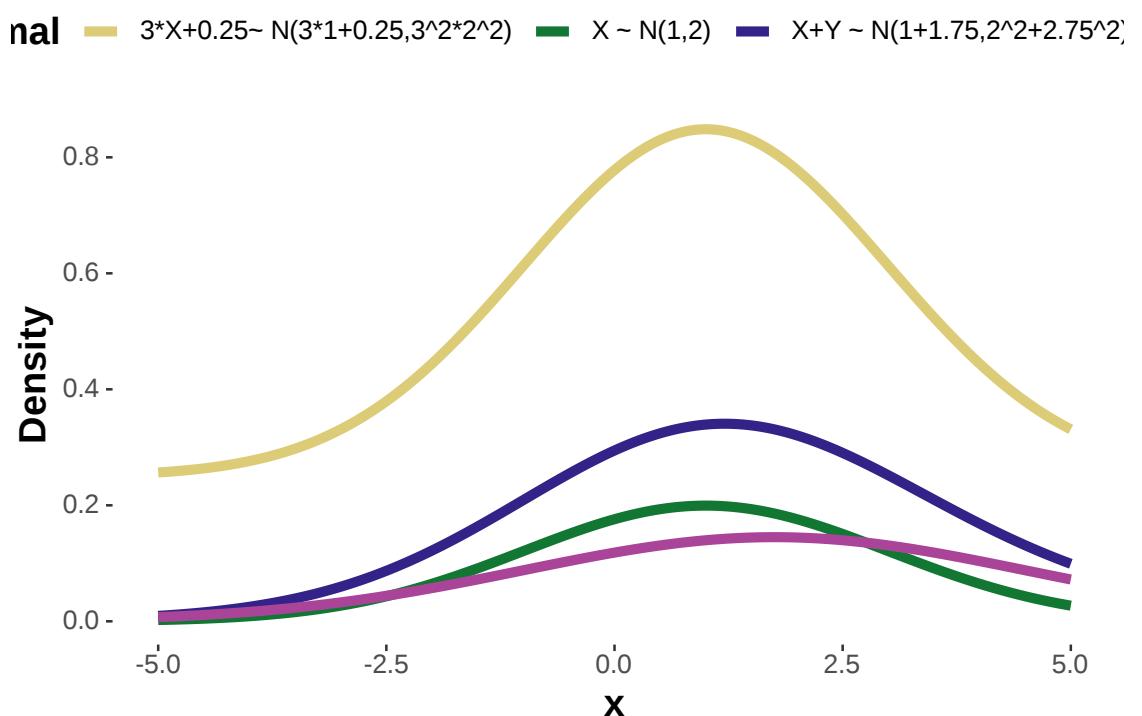


Figure B.25.: Transformation of a normal distributed random variable by addition and linear transformation.

```

rv_summ_transform_1 = rv_norm_X^2+rv_norm_Y^2,
rv_summ_transform_2 = rv_norm_X^2+rv_norm_Y^2+rv_norm_Z^2
) %>%
pivot_longer(cols = starts_with("rv_summ"),
              names_to = "random_variables",
              values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_summ_transform_1" ~ "X^2+Y^2",
                  random_variables == "rv_summ_transform_2" ~ "X^2+Y^2+Z^2"
  )
)

rv_chi_transform = tibble(
  x_axis = seq(from = 0, to = 10, by = .01),
  rv_chi_1 = dchisq(x = x_axis, df = 2),
  rv_chi_2 = dchisq(x = x_axis, df = 3)
) %>%
pivot_longer(cols = starts_with("rv"),
              names_to = "random_variables",
              values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_chi_1" ~ "Chi(2)",
                  random_variables == "rv_chi_2" ~ "Chi(3)"
  )
)

# dist plots
p1 <- ggplot() +
  geom_line(rv_chi_transform, mapping = aes(x_axis, y, color = RVs), size = 2) +
  labs(y = "Density", x = "x")

p2 <- ggplot() +
  geom_line(rv_norm_transform, mapping = aes(y, color = RVs), size = 2, stat = "density") +
  xlim(0,10) +
  ylim(0,0.5) +
  labs(y = "Density", x = "x")

grid.arrange(p1,p2, ncol = 2)

rv_norm_chi_transform = tibble(
  rv_norm_X = rnorm(n = 1e6),
  rv_norm_Y = rnorm(n = 1e6),
  rv_norm_Z = rnorm(n = 1e6),

```

B. Common probability distributions

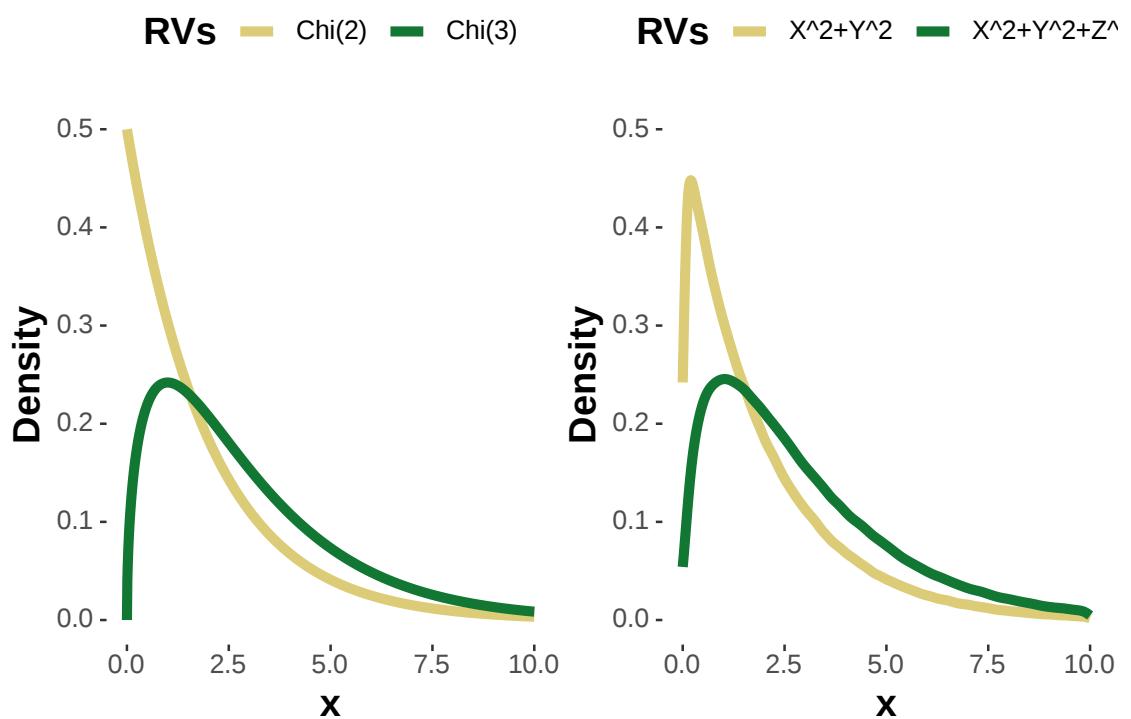


Figure B.26.: Create chi-distributed random variable by summation of standard normal random variables

```

rv_chi = rv_norm_Y^2+rv_norm_Z^2,
rv_transform_t = rv_norm_X/sqrt(rv_chi/2)
) %>%
pivot_longer(cols = starts_with("rv_transform"),
  names_to = "random_variables",
  values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_transform_t" ~ "X/sqrt(chi/df)")
)

rv_student = tibble(
  x_axis = seq(from = -5, to = 5, by = .01),
  rv_t_1 = dt(x = x_axis, df = 2)
) %>%
pivot_longer(cols = starts_with("rv"),
  names_to = "random_variables",
  values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_t" ~ "t(2)")
)

# dist plots
p1 <- ggplot() +
  geom_line(rv_student, mapping = aes(x_axis, y), size = 2) +
  ylim(0,0.4) +
  labs(y = "Density", x = "x")

p2 <- ggplot() +
  geom_line(rv_norm_chi_transform, mapping = aes(y), size = 2, stat = "density") +
  xlim(-5,5) +
  ylim(0,0.4) +
  labs(y = "Density", x = "x")

grid.arrange(arrangeGrob(p1, top = "RV ~ t(2)", arrangeGrob(p2, top = "RV ~ std.norm/sqrt(chi

rv_norm_chi_transform = tibble(
  rv_norm_V = rnorm(n = 1e6),
  rv_norm_W = rnorm(n = 1e6),
  rv_norm_X = rnorm(n = 1e6),
  rv_norm_Y = rnorm(n = 1e6),
  rv_norm_Z = rnorm(n = 1e6),
  rv_chi_1 = rv_norm_V^2+rv_norm_W^2,

```

B. Common probability distributions

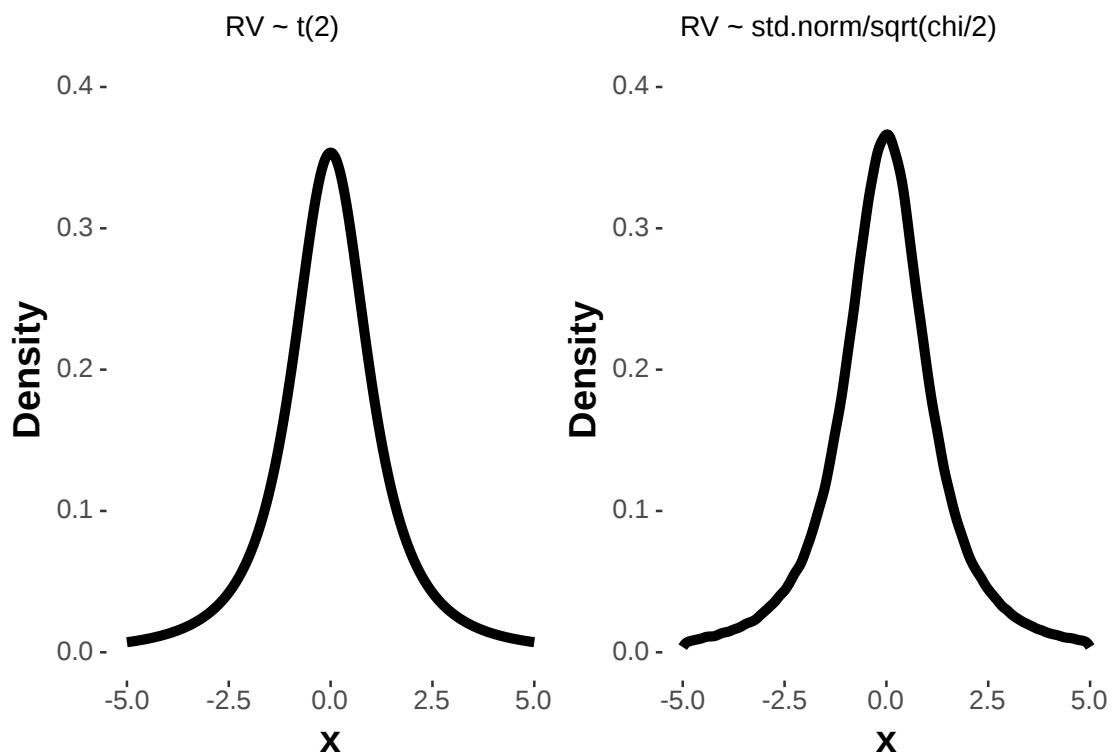


Figure B.27.: Create t-distributed random variable by division of standard normal RV by chi-squared RV

```

rv_chi_2 = rv_norm_X^2+rv_norm_Y^2+rv_norm_Z^2,
rv_transform_F = (rv_chi_1/2)/(rv_chi_2/3)
) %>%
pivot_longer(cols = starts_with("rv_transform"),
  names_to = "random_variables",
  values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_transform_F" ~ "(rv_chi_1/2)/(rv_chi_2/3)")
)

rv_fisher = tibble(
  x_axis = seq(from = 0, to = 20, by = .01),
  rv_F = df(x = x_axis, df1 = 2, df2 = 3)
) %>%
pivot_longer(cols = starts_with("rv"),
  names_to = "random_variables",
  values_to = "y") %>%
mutate(
  RVs = case_when(random_variables == "rv_F" ~ "F(2,3)")
)

# dist plots
p1 <- ggplot() +
  geom_line(rv_fisher, mapping = aes(x_axis, y), size = 2) +
  ylim(0,1) +
  labs(y = "Density", x = "x")

p2 <- ggplot() +
  geom_line(rv_norm_chi_transform, mapping = aes(y), size = 2, stat = "density") +
  xlim(0,20) +
  ylim(0,1) +
  labs(y = "Density", x = "x")

grid.arrange(arrangeGrob(p1, top = "RV ~ F(2,3)", arrangeGrob(p2, top = "RV ~ (rv_chi_1/2)/(rv_chi_2/3)"))
)

```

B. Common probability distributions

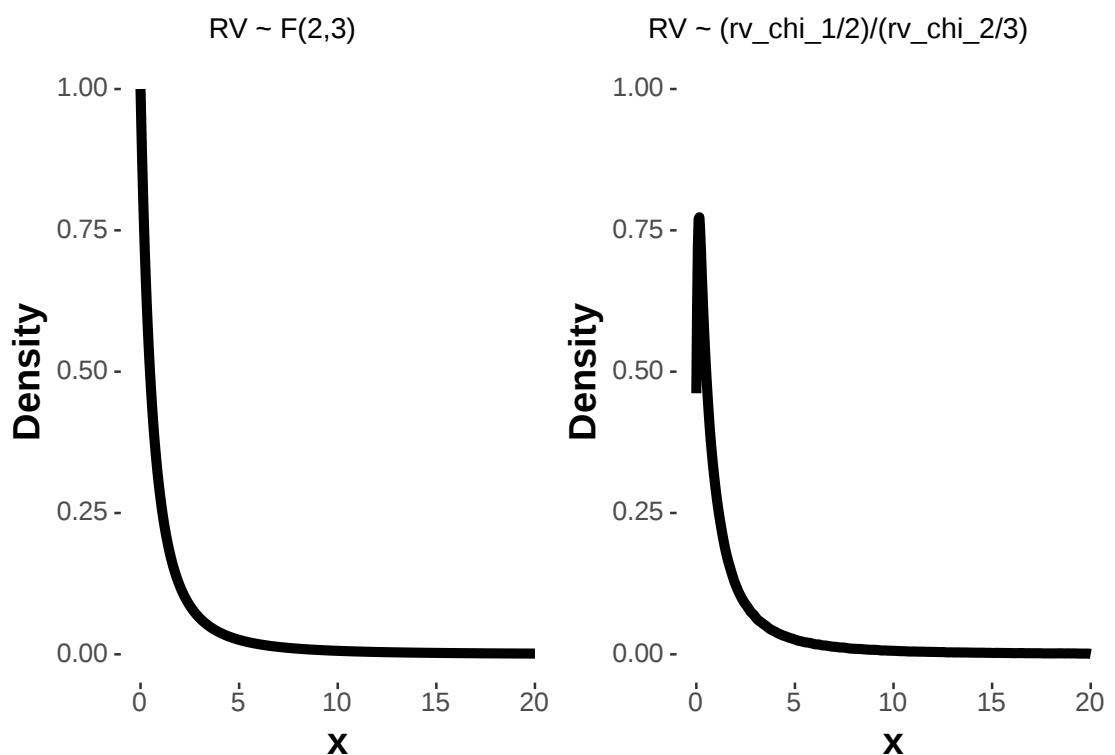


Figure B.28.: Create F-distributed random variable by division of chi-squared RV by (another independent) chi-squared RV

C. Exponential Family and Maximum Entropy

This chapter deals with the Exponential Family of probability distributions.

C.1. An important family: The Exponential Family

Most common distributions used in statistical modeling are members of the exponential family. Among others:

- Poisson distribution,
- Bernoulli distribution,
- Normal distribution,
- Chi-Square distribution, and of course the
- Exponential distribution.

In the upcoming section some of these distributions will be described in more detail. But what makes the *exponential family so special?* On the one hand, distributions of this family have some convenient mathematical properties which makes them attractive to use in *statistical modeling*. In particular for Bayesian Analysis: For example do all these distributions have a *conjugate prior* and the posterior distribution has a simple form. Furthermore the above example distributions are really just examples. The exponential family encompasses a *wide class of distributions* which makes it possible to model various cases.

On the other hand, the use of distributions from the exponential family is also from a *conceptual perspective* attractive. Consider for example the following situation:

Consider we want to infer a probability distribution subject to certain constraints. For example a coin flip experiment can have only a dichotomous outcome $\{0,1\}$ and has a constant probability. *Which distribution should be used in order to model this scenario?*

There are several possible distributions that can be used, according to which *criteria* should a distribution be selected? Often one attempts a *conservative choice*, that is to bring as little subjective information into a model as possible. Or in other terms, one goal could be to select the distribution, among all possible distributions, that is *maximal ignorant* and least biased given the constraints.

Consequently, the question arises how "*ignorance*" can be measured and *distributions compared* according to their "*information content*"? This will be topic of the upcoming excursions, the key words here are "*entropy*", which comes from information theory, and "*Maximum Entropy Principle*".

To briefly anticipate the connection between exponential family and maximum ignorance distributions: The maximum entropy principle starts with constraints that are imposed on a distribution and derives by

C. Exponential Family and Maximum Entropy

maximizing entropy a probability density/mass function. Distributions belonging to the exponential family arise as *solutions to the maximum entropy problem* subject to linear constraints.

In the upcoming section selected continuous and discrete distributions will be described in more detail. Followed by a part which motivation is to strengthen the intuition about understanding *distributions as random variables*.

C.2. Excursions: “Information Entropy” and “Maximum Entropy Principal”

C.2.1. Information Entropy

Entropy is a measure of information content of an outcome of X such that less probable outcomes convey more information than more probable ones. Thus, entropy can be stated as a *measure of uncertainty*. When the goal is to find a distribution that is as ignorant as possible, then, consequently, entropy should be maximal. Formally, entropy is defined as follows: If X is a discrete random variable with distribution $P(X = x_i) = p_i$ then the entropy of X is

$$H(X) = - \sum_i p_i \log p_i.$$

If X is a continuous random variable with probability density $p(x)$ then the differential entropy of X is

$$H(X) = - \int_{-\infty}^{+\infty} p(x) \log p(x) dx.$$

From which considerations is this *entropy* definition derived? There exist various approaches that finally come to the same answer: the above stated definition of entropy. However, the most cited derivation is Shannon's theorem. Another and perhaps more intuitive derivation is Wallis derivation. Jaynes (2003) describes both approaches in detail. The following provides a short insight in both derivations and is taken from (Jaynes 2003).

C.2.1.1. Shannon's theorem

Shannon's approach starts by stating conditions that a measure of the *amount of uncertainty* H_n has to satisfy.

1. It is possible to set up some kind of association between *amount of uncertainty* and real numbers
2. H_n is a continuous function of p_i . Otherwise, an arbitrarily small change in the probability distribution would lead to a big change in the amount of uncertainty.
3. H_n should correspond to common sense in that, when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in case the p_i are all equal, the quantity $h(n)$ is a monotonic increasing function of n .
4. H_n is consistent in that, when there is more than one way of working out its value, we must get the same answer for few possible ways.

C.2. Excursions: "Information Entropy" and "Maximum Entropy Principal"

Under these assumptions the resulting unique measure of uncertainty of a probability distribution p turns out to be just the average *log-probability*:

$$H(p) = - \sum_i p_i \log(p_i).$$

(The interested reader can find a systematic derivation in (Jaynes 2003).) Accepting this interpretation of entropy, it follows that the distribution (p_1, \dots, p_n) which maximizes the above equation, subject to constraints imposed by the available information, will represent the most *honest* description of what the model *knows* about the propositions (A_1, \dots, A_n) (Jaynes 2003).

The function H is called the *entropy*, or the *information entropy* of the distribution $\{p_i\}$.

C.2.1.2. The Wallis derivation

A second and perhaps more intuitive approach of deriving entropy was suggested by G. Wallis. The following description is taken from Jaynes (2003).

We are given information I , which is to be used in assigning probabilities $\{p_1, \dots, p_m\}$ to m different probabilities. We have a total amount of probability

$$\sum_{i=1}^m p_i = 1$$

to allocate among them.

The problem can be stated as follows. Choose some integer $n \gg m$, and imagine that we have n little *quanta* of probabilities, each of magnitude $\delta = \frac{1}{n}$, to distribute in an way we see fit.

Suppose we were to scatter these quanta at random among the m choices (penny-pitch game into m equal boxes). If we simply toss these quanta of probability at random, so that each box has an equal probability of getting them, nobody can claim that any box is being unfairly favoured over any other.

If we do this and the first box receives exactly n_1 quanta, the second n_2 quanta etc. we will say the random experiment has generated the probability assignment:

$$p_i = n_i \delta = \frac{n_i}{n}, \text{ with } i = 1, 2, \dots, m.$$

The probability that this will happen is the multinomial distribution:

$$m^{-n} \frac{n!}{n_1! \cdot \dots \cdot n_m!}.$$

C. Exponential Family and Maximum Entropy

Now imagine that we repeatedly scatter the n quanta at random among the m boxes. Each time we do this we examine the resulting probability assignment. If it happens to conform to the information I , we accept it; otherwise we reject it and try again. We continue until some probability assignment $\{p_1, \dots, p_m\}$ is accepted.

What is the most likely probability distribution to result from this game? It is the one which maximizes

$$W = \frac{n!}{n_1! \cdot \dots \cdot n_m!}$$

subject whatever constraints are imposed by the information I .

We can refine this procedure by using smaller quanta, i.e. large n . By using *Stirlings approximation*

$$n! \sim \sqrt{(2\pi n)} \left(\frac{n}{e}\right)^n,$$

and taking the logarithm from it:

$$\log(n!) \sim \sqrt{(2\pi n)} + n \log\left(\frac{n}{e}\right),$$

we have

$$\log(n!) \sim \sqrt{(2\pi n)} + n \log(n) - n.$$

Taking furthermore, also the logarithm from W and substituting $\log(n!)$ by Sterlings approximation, finally gives the definition of information entropy, as derived by Shannon's theorem:

$$\frac{1}{n} \log(W) \rightarrow - \sum_{i=1}^m p_i \log(p_i) = H(p_1, \dots, p_m).$$

To sum it up: Entropy is a measure of uncertainty. The higher the entropy of a random variable X the more uncertainty it incorporates. When the goal is to find a maximal ignorance distribution, this goal can be consequently translated into a maximization problem: Find the distribution with maximal entropy subject to existing constraints. This will be topic of the next part of our excusos.

C.2.2. Deriving Probability Distributions using the Maximum Entropy Principle

The maximum entropy principle is a means of deriving probability distributions given certain constraints and the assumption of maximizing entropy. One technique for solving this maximization problem is the *Lagrange multiplier technique*.

C.2.2.1. Lagrangian multiplier technique

Given a multivariable function $f(x, y, \dots)$ and constraints of the form $g(x, y, \dots) = c$, where g is another multivariable function with the same input space as f and c is a constant.

In order to minimize (or maximize) the function f consider the following steps, assuming f to be $f(x)$:

1. Introduce a new variable λ , called *Lagrange multiplier*, and define a new function \mathcal{L} with the form:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda(g(x) - c).$$

2. Set the derivative of the function \mathcal{L} equal to the zero:

$$\mathcal{L}'(x, \lambda) = 0,$$

in order to find the critical points of \mathcal{L} .

3. Consider each resulting solution within the limits of the made constraints and derive the resulting distribution f , which gives the minimum (or maximum) one is searching for.

For more details see (Academy 2019)

C.2.2.2. Example 1: Derivation of maximum entropy pdf with no other constraints

For more details see (Finlayson 2017, @keng2017)

Suppose a random variable for which we have absolutely no information on its probability distribution, beside the fact that it should be a pdf and thus, integrate to 1. We ask for the following:

What type of probability density distribution gives maximum entropy when the random variable is bounded by a finite interval, say $a \leq X \leq b$? (Reza 1994)

We assume that the maximum ignorance distribution is the one with maximum entropy. It minimizes the prior information in a distribution and is therefore the most conservative choice.

For the continuous case entropy, the measure of uncertainty, is defined as

$$H(x) = - \int_a^b p(x) \log(p(x)) dx,$$

with subject to the mentioned constraint that the sum of all probabilities is one (as it is a pdf):

$$\int_a^b p(x) dx = 1.$$

C. Exponential Family and Maximum Entropy

Rewrite this into the form of *Lagrangian* equation gives

$$\mathcal{L} = - \int_a^b p(x) \log(p(x)) dx + \lambda \left(\int_a^b p(x) dx - 1 \right).$$

The next step is to *minimize* the Lagrangian function. To solve this, we have to use the *calculus of variations*(Keng 2017).

First differentiating \mathcal{L} with respect to $p(x)$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p(x)} &= 0, \\ -1 - \log(p(x)) + \lambda &= 0, \\ p(x) &= e^{(\lambda-1)}. \end{aligned}$$

Second, the result of $p(x)$ has to satisfy the stated constraint

$$\int_a^b p(x) dx = 1,$$

$$\int_a^b e^{1-\lambda} dx = 1.$$

Solving this equation with respect to λ gives:

$$\lambda = 1 - \log\left(\frac{1}{b-a}\right).$$

Taking both solutions together we get the following probability density function:

$$p(x) = e^{(1-\lambda)} = e^{\left(1 - \left(1 - \log\left(\frac{1}{b-a}\right)\right)\right)},$$

$$p(x) = \frac{1}{b-a}.$$

And this is the *uniform distribution* on the interval $[a, b]$. Such that, the answer of the above question is:

The maximum entropy distribution is associated with a random variable , that is distributed as uniform probability density distribution between a and b.

This should not be too unexpected. As it is quite intuitive that a uniform distribution is the maximal ignorance distribution (when no other constraints were made). The next example will be more exciting.

C.2.2.3. Example 2: Derivation of maximum entropy pdf with given mean μ and variance σ^2

Suppose a random variable X with a preassigned standard deviation σ and mean μ . Again the question is:

Which function $p(x)$ gives the maximum of the entropy $H(x)$?

The Maximum Entropy is defined for the current case as

$$H(X) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx,$$

is subject to the constraint that it should be a pdf

$$\int_{-\infty}^{\infty} p(x) dx = 1,$$

and that μ and σ are given (whereby only one constrained is needed, as the μ is already included in the definition of σ):

$$\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx = \sigma^2.$$

Accordingly to the above mentioned technique the formulas are summarized in form of the *Lagrangian* equation:

$$\mathcal{L} = - \int_{-\infty}^{\infty} p(x) \log p(x) dx + \lambda_0 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right) + \lambda_1 \left(\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right).$$

Next, \mathcal{L} will be partially differentiated with respect to $p(x)$:

$$\frac{\partial \mathcal{L}}{\partial p(x)} = 0,$$

$$-(1 + \log p(x)) + \lambda_0 + \lambda_1(x - \mu)^2 = 0,$$

$$p(x) = e^{\lambda_0 + \lambda_1(x - \mu)^2 - 1}.$$

Further we have to make sure that the result holds for the stated constraints:

$$\int_{-\infty}^{\infty} e^{\lambda_0 + \lambda_1(x - \mu)^2 - 1} - 1 dx = 1,$$

C. Exponential Family and Maximum Entropy

and

$$\int_{-\infty}^{\infty} (x - \mu)^2 e^{\lambda_0 + \lambda_1(x - \mu)^2 - 1} dx = \sigma^2.$$

For the first constraint we get

$$e^{\lambda_0 - 1} \sqrt{-\frac{\pi}{\lambda_1}} = 1,$$

and for the second constraint

$$e^{\lambda_0 - 1} = \sqrt{\frac{1}{2\pi}} \frac{1}{\sigma},$$

Thus

$$\lambda_1 = \frac{-1}{2\sigma^2}$$

Taking all together we can write:

$$p(x) = e^{\lambda_0 + \lambda_1(x - \mu)^2 - 1} = e^{\lambda_0 - 1} e^{\lambda_1(x - \mu)^2},$$

substituting the solutions for $e^{\lambda_0 - 1}$ and λ_1 :

$$p(x) = \sqrt{\frac{1}{2\pi}} \frac{1}{\sigma} e^{\frac{-1}{2\sigma^2}(x - \mu)^2},$$

finally we can rearrange the terms a bit and get:

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{-1}{2} \left(\frac{(x - \mu)^2}{\sigma^2}\right)\right),$$

the *Gaussian probability density distribution*.

To sum it up:

If one is to infer a probability distribution given certain constraints, out of all distributions $\{p_i\}$ compatible with them, one should pick the distribution $\{p_i^*\}$ having the largest value of H (De Martino and De Martino 2018). In other terms, a Maximum Entropy distribution is completely undetermined by features that do not appear explicitly in the constraints subject to which it has been computed.

An **overview of Maximum Entropy distributions** can be found on Wikipedia.

D. Data sets used in the book

Several data sets are used throughout the book as ‘running examples’. They occur in different places to illustrate different things. This chapter centrally describes each data set, together with the most important visualizations and analyses.

D.1. Mental Chronometry

D.1.1. Nature, origin and rationale of the data

Francis Donders is remembered as one of, if not the first experimental cognitive psychologists. He famously introduced the **subtraction logic** which looks at difference in reaction times across different tasks to infer difference in the complexity of the mental processes involved in these tasks. The Mental Chronometry data set presents the results of an online replication of one such subtraction-experiment.

D.1.1.1. The experiment

50 participants were recruited using the crowd-sourcing platform Prolific and paid for their participation. In each experiment trial, participants see either a blue square or a blue circle appear on the screen and are asked to respond as quickly as possible. The experiment consists of three parts, presented to all participants in the same order (see below). The parts differ in the adequate response to the visual stimuli.

1. Reaction task

The participant presses the space bar whenever there is a stimulus (square or circle)

Recorded: reaction time

2. Go/No-Go task

The participant presses the space bar whenever their target (one of the two stimuli) is on the screen

Recorded: the reaction time and the response

3. Discrimination task

The participant presses the **F** key on the keyboard when there is one of the stimuli and the **J** key when there is the other one of the stimuli on the screen.

Recorded: the reaction time and the response

D. Data sets used in the book

The **reaction time** measurement starts from the onset of the visual stimuli to the button press. The **response** variable records whether the reaction was correct or incorrect.

For each participant, the experiment randomly allocates one shape (circle or square) as the target to be used in both the second and the third task.

The experiment was realized using _magpie and can be tried out here.

D.1.1.2. Theoretical motivation & hypotheses

We expect that reaction times of correct responses are lowest in the reaction task, higher in the Go/No-Go task, and highest in the discrimination task.

D.1.2. Loading and preprocessing the data

The raw data produced by the online experiment is not particularly tidy. It needs substantial massages before plotting and analysis.

```
## Observations: 3,750
## Variables: 32
## $ submission_id <dbl> 8554, 8554, 8554, 8554, 8554, 8554, 8554, 8554, 8554, ...
## $ QUD <chr> "Press SPACE when you see a shape on the screen", "Pr...
## $ RT <dbl> 376, 311, 329, 270, 284, 311, 269, 317, 325, 240, 262...
## $ age <dbl> 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 2...
## $ comments <chr> NA, N...
## $ correctness <chr> "correct", "correct", "correct", "correct", "correct"...
## $ education <chr> "high school / college", "high school / college", "hi...
## $ elemSize <dbl> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100...
## $ endTime <dbl> 1.570374e+12, 1.570374e+12, 1.570374e+12, 1.570374e+1...
## $ expected <chr> NA, N...
## $ experiment_id <dbl> 68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 6...
## $ f <chr> NA, N...
## $ focalColor <chr> "blue", "blue", "blue", "blue", "blue", "blue...
## $ focalNumber <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ focalShape <chr> "square", "square", "circle", "square", "circle", "ci...
## $ gender <chr> "female", "female", "female", "female", "fe...
## $ j <chr> NA, N...
## $ key1 <lgl> NA, N...
## $ key2 <chr> NA, N...
## $ key_pressed <chr> NA, N...
## $ languages <chr> "Right", "Right", "Right", "Right", "Right", ...
## $ pause <dbl> 2631, 1700, 1322, 1787, 1295, 2330, 1620, 2460, 1580, ...
## $ response <chr> "space", "space", "space", "space", "space", "space", ...
## $ sort <chr> "grid", "grid", "grid", "grid", "grid", "grid...
```

```

## $ startDate      <chr> "Sun Oct 06 2019 15:45:19 GMT+0100 (Hora de verão da ...
## $ startTime     <dbl> 1.570373e+12, 1.570373e+12, 1.570373e+1...
## $ stimulus       <chr> "square", "square", "circle", "square", "circle", "ci...
## $ target         <chr> "square", "square", "circle", "square", "circle", "ci...
## $ timeSpent     <dbl> 7.2514, 7.2514, 7.2514, 7.2514, 7.2514, 7.251...
## $ total          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ trial_number   <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...
## $ trial_type     <chr> "reaction_practice", "reaction_practice", "reaction_p...

```

```

mc_data_raw <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-data-analy...
glimpse(mc_data_raw)

```

The most pressing problem is that entries in the column `trial_type` contain two logically separate pieces of information: the block (reaction, go/no-go, discrimination) *and* whether the data comes from a practice trial (which we want to discard) or a main trial (which we want to analyze). We therefore separate this information, and perform some other massages, to finally select a preprocessed data set for further analysis:

```

block_levels <- c("reaction", "goNoGo", "discrimination") # ordering of blocks for plotting, e

mc_data_preprocessed <- mc_data_raw %>%
  separate(trial_type, c("block", "stage"), sep = "_", remove = FALSE) %>%
  mutate(comments = ifelse(is.na(comments), "non given", comments)) %>%
  filter(stage == "main") %>%
  mutate(
    block = factor(block, ordered = T, levels = block_levels),
    response = ifelse(is.na(response), "none", response)
  ) %>%
  filter(response != "wait") %>%
  rename(
    handedness = languages, # variable name is simply wrong
    total_time_spent = timeSpent
  ) %>%
  select(
    submission_id,
    trial_number,
    block,
    stimulus,
    RT,
    handedness,
    gender,
    total_time_spent,
    comments
  )

```

D. Data sets used in the book

```
)  
  
# write_csv(mc_data_preprocessed, 'mental-chrono-data_preprocessed.csv')
```

D.1.3. Cleaning the data

Remeber that the criteria for data exclusion should ideally be defined before data collection (or at least inspection). They should definitely never be chosen in such a way as to maximize the “desirability” of an analysis. Data cleaning is not a way of making sure that your favorite research hypothesis “wins”.

Although we have not preregistered any data cleaning regime or analyses for this data set, we demonstrate a frequently used cleaning scheme for reaction time data, which does depend on the data in some sense, but does not require precise knowledge of the data. In particular, we are going to do this:

1. We remove remove the data from an individual participant X if there is an experimental condition C such that the mean RT of X for condition C is more than 2 standard deviations away from the overall mean RT for condition C .
2. From the remaining data, we then remove any individual trial Y if the RT of Y is more than 2 standard deviations away from the mean of experimental condition C (where C is the condition of Y , of course).

Notice that in the case at hand, the experimental conditions are the three types of tasks.

D.1.3.1. Cleaning by-participant

Our rule for removing data from outlier participants is this:

We remove remove the data from an individual participant X if there is an experimental condition C such that the mean RT of X for condition C is more than 2 standard deviations away from the overall mean RT for condition C . We also remove all trials with reaction times below 100ms.

This procedure is implemented in this code:

```
# summary stats (means) for participants  
d_sum_stats_participants <- mc_data_preprocessed %>%  
  group_by(submission_id, block) %>%  
  summarise(  
    mean_P = mean(RT)  
  )  
  
# summary stats (means and SDs) for conditions  
d_sum_stats_conditions <- mc_data_preprocessed %>%  
  group_by(block) %>%
```

```

summarise(
  mean_C = mean(RT),
  sd_C   = sd(RT)
)

d_sum_stats_participants <-
  full_join(
    d_sum_stats_participants,
    d_sum_stats_conditions,
    by = "block"
  ) %>%
  mutate(
    outlier_P = abs(mean_P - mean_C) > 2 * sd_C
  )

# show outlier participants
d_sum_stats_participants %>% filter(outlier_P == 1) %>% show()

## # A tibble: 1 x 6
## # Groups:   submission_id [1]
##   submission_id block      mean_P  mean_C  sd_C outlier_P
##   <dbl> <ord>      <dbl>  <dbl>  <dbl> <lgl>
## 1       8505 discrimination 1078.   518.  185. TRUE

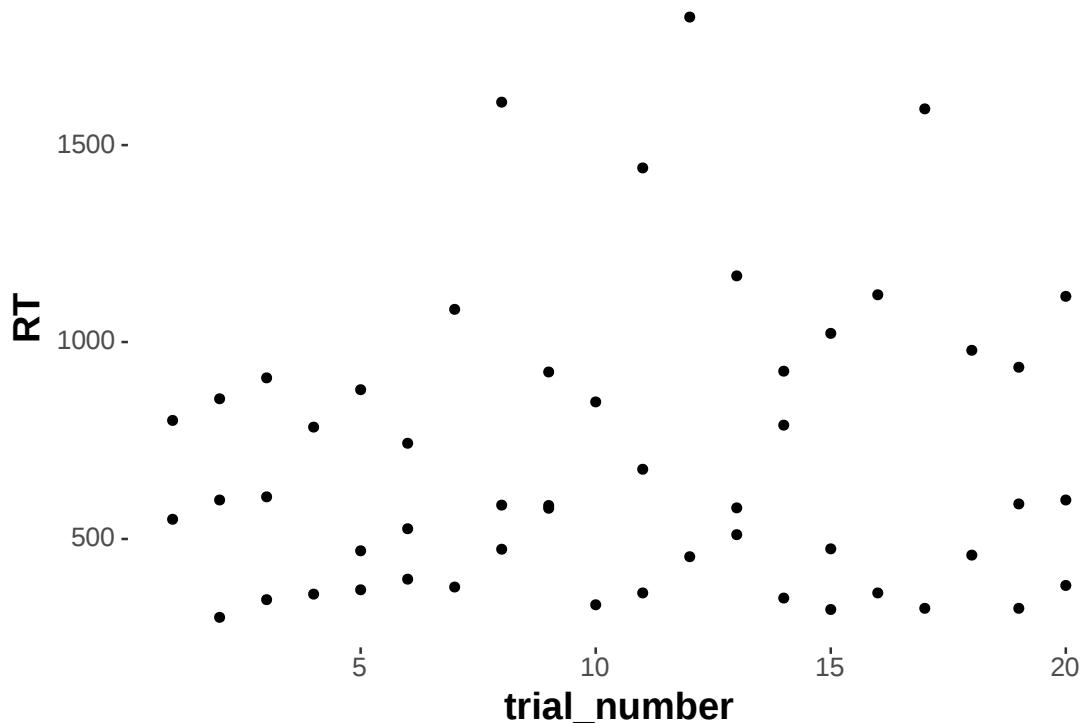
```

When plotting the data for this condition and this participant, we see that the high overall mean is not just caused by a single outlier, but several trials that took longer than 1 second.

```

mc_data_preprocessed %>%
  semi_join(
    d_sum_stats_participants %>% filter(outlier_P == 1),
    by = c("submission_id")
  ) %>%
  ggplot(aes(x = trial_number, y = RT)) +
  geom_point()

```



We are then going to exclude this participant's entire data from all subsequent analysis:¹

```
mc_data_cleaned <- mc_data_preprocessed %>%
  filter(submission_id != d_sum_stats_participants$submission_id[1] )
```

D.1.3.2. Cleaning by-trial

Our rule for excluding data from individual trials is:

From the remaining data, we then remove any individual trial Y if the RT of Y is more than 2 standard deviations away from the mean of experimental condition C (where C is the condition of Y , of course). We also remove all trials with reaction times below 100ms.

The following code implements this:

```
# mark individual trials as outliers
mc_data_cleaned <- mc_data_cleaned %>%
  full_join(
```

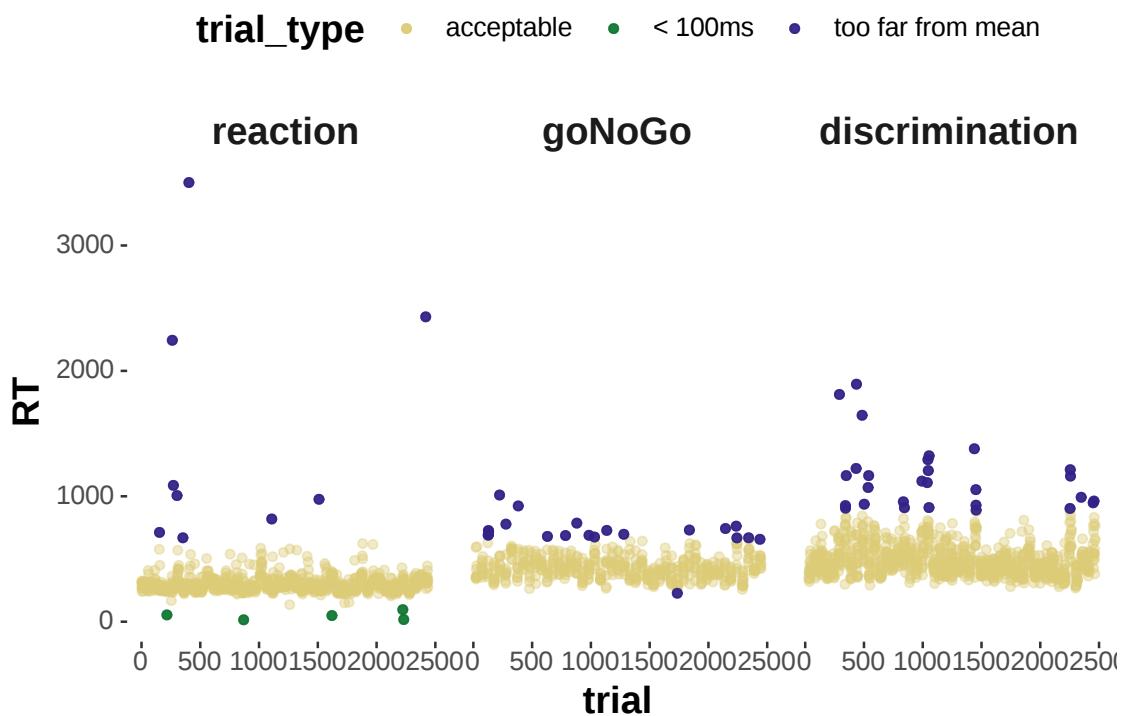
¹This may seem a harsh step, but when data acquisition is cheap, it's generally not a bad strategy to be very strict in exclusion criteria, and to apply rules that are not strongly context-dependent.

```

d_sum_stats_conditions,
by = "block"
) %>%
mutate(
  trial_type = case_when(
    abs(RT - mean_C) > 2 * sd_C ~ "too far from mean",
    RT < 100 ~ "< 100ms",
    TRUE ~ "acceptable"
  ) %>% factor(levels = c("acceptable", "< 100ms", "too far from mean")),
  trial = 1:nrow(mc_data_cleaned)
)
# visualize outlier trials

mc_data_cleaned %>%
  ggplot(aes(x = trial, y = RT, color = trial_type)) +
  geom_point(alpha = 0.4) + facet_grid(~block) +
  geom_point(alpha = 0.9, data = filter(mc_data_cleaned, trial_type != "acceptable"))

```



So, we remove 63 individual trials.

D. Data sets used in the book

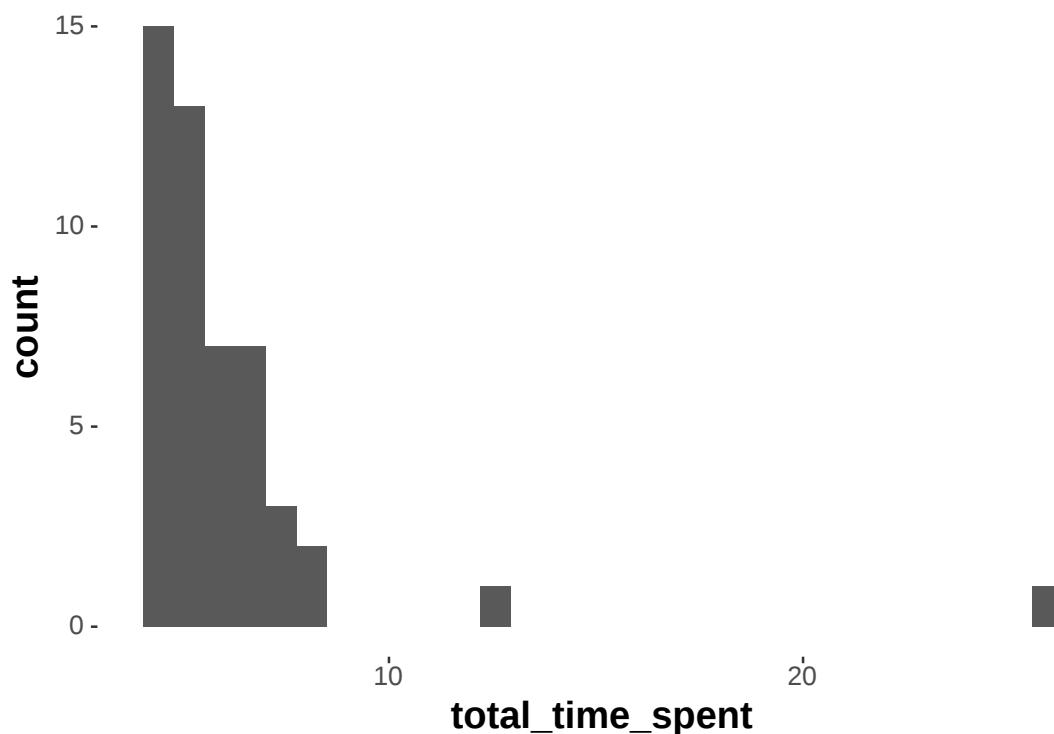
```
mc_data_cleaned <- mc_data_cleaned %>%
  filter(trial_type == "acceptable")

## this version of the data is stored as cleaned
# write_csv(mc_data_cleaned, "data_sets/mental-chrono-data_cleaned.csv")
```

D.1.4. Exploration: summary stats & plots

What's the distribution of `total_time_spent`, i.e., the time each participant took to complete the whole study?

```
mc_data_cleaned %>%
  select(submission_id, total_time_spent) %>%
  unique() %>%
  ggplot(aes(x = total_time_spent)) +
  geom_histogram()
```



There are two participants who took noticeably longer than all the others, but we need not necessarily be concerned about this, because it is not unusual for participants of online experiments to open the experiment and wait before actually starting.

Here are summary statistics for the reaction time measures for each condition (= block).

```
mc_sum_stats_blocks_cleaned <- mc_data_cleaned %>%
  group_by(block) %>%
  nest() %>%
  summarise(
    CIIs = map(data, function(d) bootstrapped_CI(d$RT)))
  ) %>%
  unnest(CIIs)

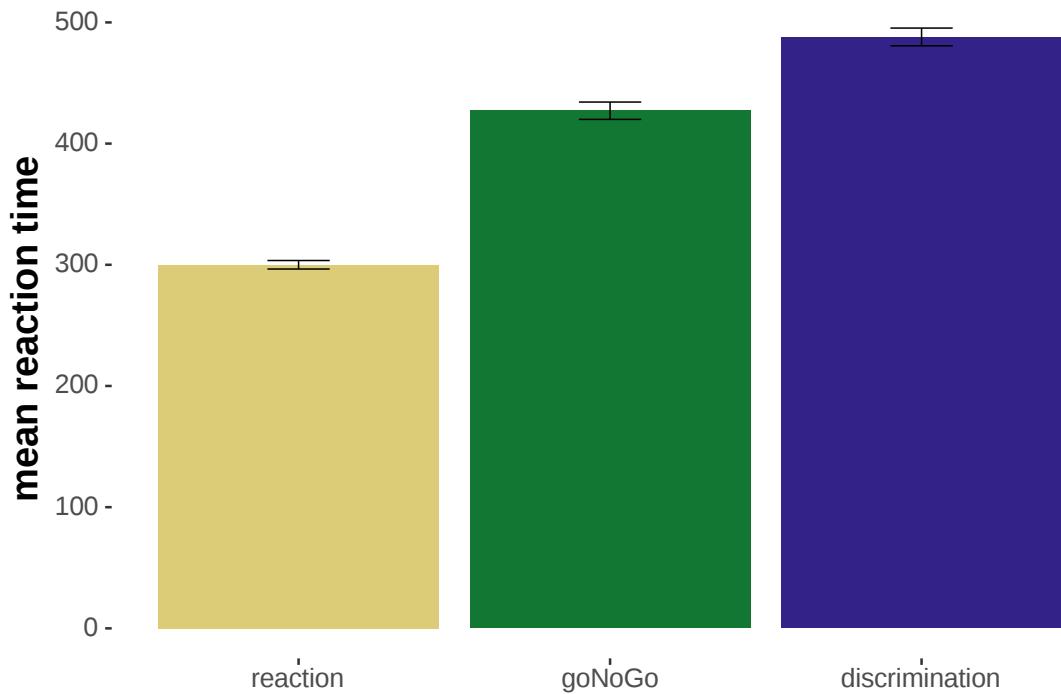
mc_sum_stats_blocks_cleaned

## # A tibble: 3 x 4
##   block      lower   mean upper
##   <ord>     <dbl> <dbl> <dbl>
## 1 reaction  297.  300.  304.
## 2 goNoGo    420.  427.  434.
## 3 discrimination 481. 488. 495.
```

And a plot of the summary:

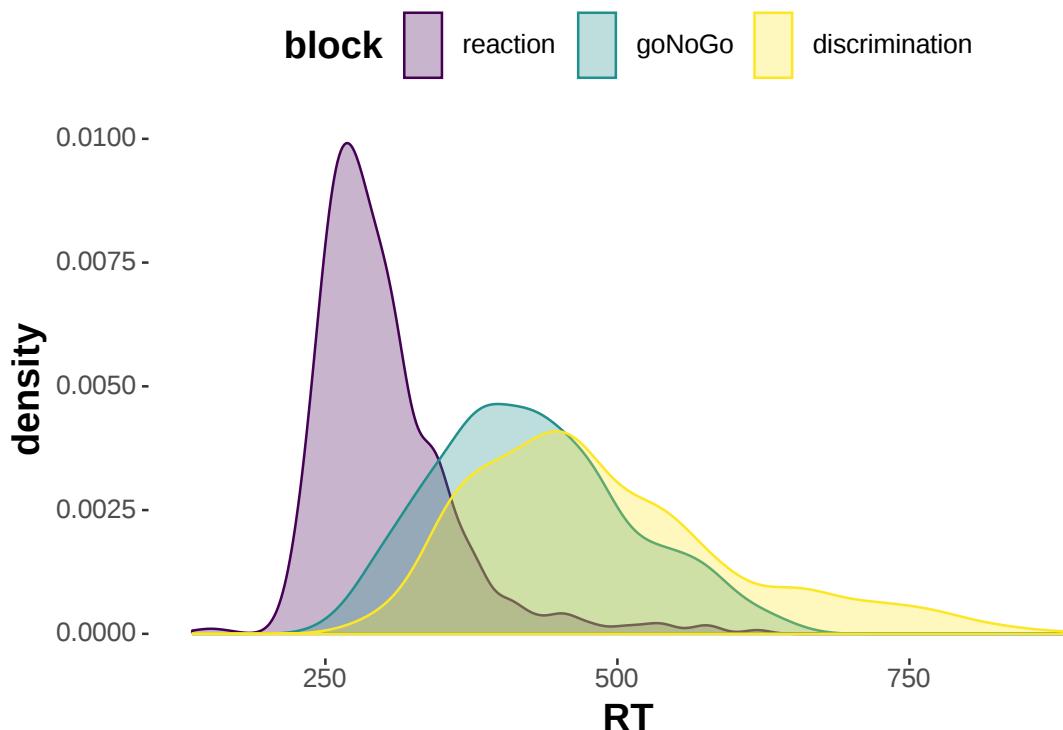
```
mc_sum_stats_blocks_cleaned %>%
  ggplot(aes(x = block, y = mean, fill = block)) +
  geom_col() +
  geom_errorbar(aes(ymin = lower, ymax = upper), size = 0.3, width = 0.2) +
  ylab("mean reaction time") + xlab("") +
  scale_fill_manual(values = project_colors) +
  theme(legend.position = "none")
```

D. Data sets used in the book



We can also plot the data in a manner that is more revealing of the distribution of measurements in each condition:

```
mc_data_cleaned %>%
  ggplot(aes(x = RT, color = block, fill = block)) +
  geom_density(alpha = 0.3)
```



D.1.5. Data analysis

We are interested in seeing whether the mean RTs are smallest for the 'reaction' task, higher for the 'go/no-go' task, and highest for the 'discrimination' task. We test this with a hierarchical Bayesian regression model, taking participant-level variation of intercepts and slopes for factor block into account.

We make 'go/no-go' the default level of the block factor, so that we can directly test our directed hypothesis, using posterior parameter inference.

```
# making 'go/no-go' the reference level
## TODO : write convenience function for this kind of releveling
reflevel <- "goNoGo"
reflevel_index <- which(levels(mc_data_cleaned$block) == reflevel)
contrasts(mc_data_cleaned$block) <- contr.treatment(
  nlevels(mc_data_cleaned$block),
  reflevel_index
)
colnames(contrasts(mc_data_cleaned$block)) <- str_c("_.", levels(mc_data_cleaned$block)[-reflevel_index])

regression_model_ME <- brm(
  formula = RT ~ block + (1 + block | submission_id),
  data = mc_data_cleaned
```

D. Data sets used in the book

```
)  
  
## TODO tidy and concise output  
regression_model_ME
```

We see that the value zero lies clearly outside of the 95% credible interval for block ‘reaction’ and for block ‘discrimination’. The deviation from the intercept (zero point) is in the expected direction. We may conclude that, as hypothesized, reaction times in the ‘reaction’ condition are lowest, higher in the ‘go/no-go’ condition, and highest in the ‘discrimination’ condition.

D.2. Simon Task

CAVEAT: THIS CHAPTER IS A DRAFT; DEFER READING UNTIL LATER

The Simon task is pretty cool. The task is designed to see if responses are faster and/or more accurate when the stimulus to respond to occurs in the same relative location as the response, even if the stimulus location is irrelevant to the task. For example, it is faster to respond to a stimulus presented on the left of the screen with a key that is on the left of the keyboard (e.g. q), than with a key that is on the right of the keyboard (e.g. p).

D.2.1. Experiment

D.2.1.1. Participants

A total of 213 participants took part in an online version of a Simon task. Participants were students enrolled in either “Introduction to Cognitive (Neuro-)Psychology” (N = 166), or “Experimental Psychology Lab Practice” (N = 39) or both (N = 4).

D.2.1.2. Materials & Design

Each trial started by showing a fixation cross for 200 ms in the center of the screen. Then, one of two geometrical shapes was shown for 500 ms. The **target shape** was either a blue square or a blue circle. The target shape appeared either on the left or right of the screen. Each trial determined uniformly at random which shape (square or circle) to show as target and where on the screen to display it (left or right). Participants were instructed to press keys q (left of keyboard) or p (right of keyboard) to identify the kind of shape on the screen. The shape-key allocation happened experiment initially, uniformly at random once for each participant and remained constant throughout the experiment. For example, a participant may have been asked to press q for square and p for circle.

Trials were categorized as either ‘congruent’ or ‘incongruent’. They were congruent if the location of the stimulus was the same relative location as the response key (e.g. square on the right of the screen, and p

key to be pressed for square) and incongruent if the stimulus was not in the same relative location as the response key (e.g. square on the right and q key to be pressed for square).

In each trial, if no key was pressed within 3 seconds after the appearance of the target shape, a message to please respond faster was displayed on screen.

D.2.1.3. Procedure

Participants were first welcomed and made familiar with the experiment. They were told to optimize both speed and accuracy. They then practiced the task for 20 trials before starting the main task, which consisted of 100 trials. Finally, the experiment ended with a post-test survey in which participants were asked for their student IDs and the class they were enrolled in. They were also able to leave any optional comments.

D.2.2. Results

D.2.2.1. Loading and inspecting the data

We load the data into R and show a summary of the variables stored in the tibble:

```
d <- read_csv("data_sets/simon-task.csv")
glimpse(d)

## # Observations: 25,560
## # Variables: 15
## # $ submission_id    <dbl> 7432, 7432, 7432, 7432, 7432, 7432, 7432, 743...
## # $ RT                <dbl> 1239, 938, 744, 528, 706, 547, 591, 652, 627, 485, ...
## # $ condition          <chr> "incongruent", "incongruent", "incongruent", "incon...
## # $ correctness         <chr> "correct", "correct", "correct", "correct", "correct", ...
## # $ class               <chr> "Intro Cogn. Neuro-Psychology", "Intro Cogn. Neuro-...
## # $ experiment_id      <dbl> 52, 52, 52, 52, 52, 52, 52, 52, 52, 52, 52, 52, ...
## # $ key_pressed         <chr> "q", "q", "q", "q", "p", "p", "q", "p", "q", "q", ...
## # $ p                  <chr> "circle", "circle", "circle", "circle", "circle", ...
## # $ pause               <dbl> 1896, 1289, 1705, 2115, 2446, 2289, 2057, 2513, 186...
## # $ q                  <chr> "square", "square", "square", "square", "square", ...
## # $ target_object       <chr> "square", "square", "square", "square", "square", ...
## # $ target_position     <chr> "right", "right", "right", "right", "left", "right", ...
## # $ timeSpent           <dbl> 7.565417, 7.565417, 7.565417, 7.565417, 7.565417, 7...
## # $ trial_number        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## # $ trial_type          <chr> "practice", "practice", "practice", "practice", "pr...
```

It is often useful to check general properties, such as the mean time participants spent on the experiment:

D. Data sets used in the book

```
d %>% pull(timeSpent) %>% mean()
```

```
## [1] 21.61656
```

About 21.62 minutes is quite long, but we know that the mean is very susceptible to outliers, so we may want to look at a more informative set of **summary statistics**:

```
d %>% pull(timeSpent) %>% summary()
```

```
##      Min.   1st Qu.    Median     Mean   3rd Qu.    Max. 
##      5.648    6.905    7.692   21.617   9.113 1158.110
```

D.2.2.2. Summarizing & cleaning the data

We look at outlier-y behavior at the level of individual participants first, then at the level of individual trials.

D.2.2.2.1. Individual-level error rates & reaction times

It is conceivable that some participants did not take the task seriously. They may have just fooled around. We will therefore inspect each individual's response patterns and reaction times. If participants appear to have "misbehaved" we discard all of their data. (**CAVEAT:** Notice the researcher degrees of freedom in the decision of what counts as "misbehavior"! It is therefore that choices like these are best committed to in advance, e.g. via pre-registration!)

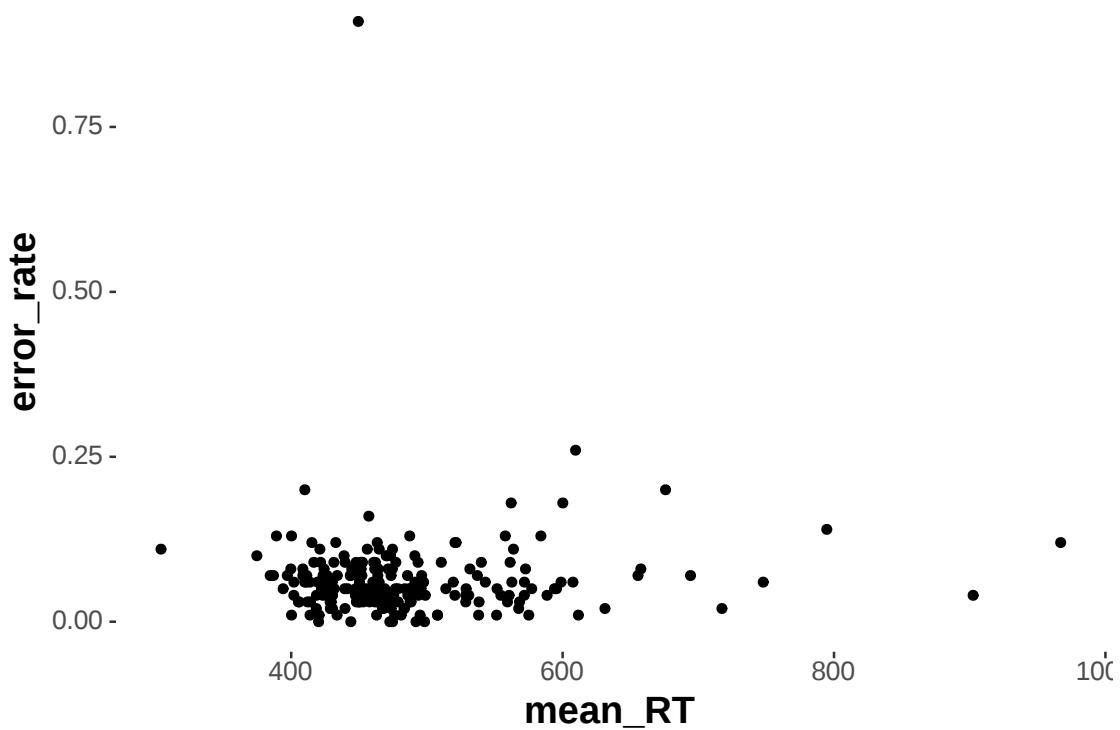
We can calculate the mean reaction times and the error rates for each participant.

```
d_individual_summary <- d %>%
  filter(trial_type == "main") %>%      # look at only data from main trials
  group_by(submission_id) %>%            # calculate the following for each individual
  summarize(mean_RT = mean(RT),
            error_rate = 1 - mean(ifelse(correctness == "correct", 1, 0)))
head(d_individual_summary)

## # A tibble: 6 x 3
##   submission_id  mean_RT  error_rate
##       <dbl>     <dbl>      <dbl>
## 1         7432     595.      0.05
## 2         7433     458.      0.04
## 3         7434     531.      0.04
## 4         7435     433.      0.12
## 5         7436     748.      0.06
## 6         7437     522.      0.12
```

Let's plot this summary information:

```
d_individual_summary %>%
  ggplot(aes(x = mean_RT, y = error_rate)) +
  geom_point()
```

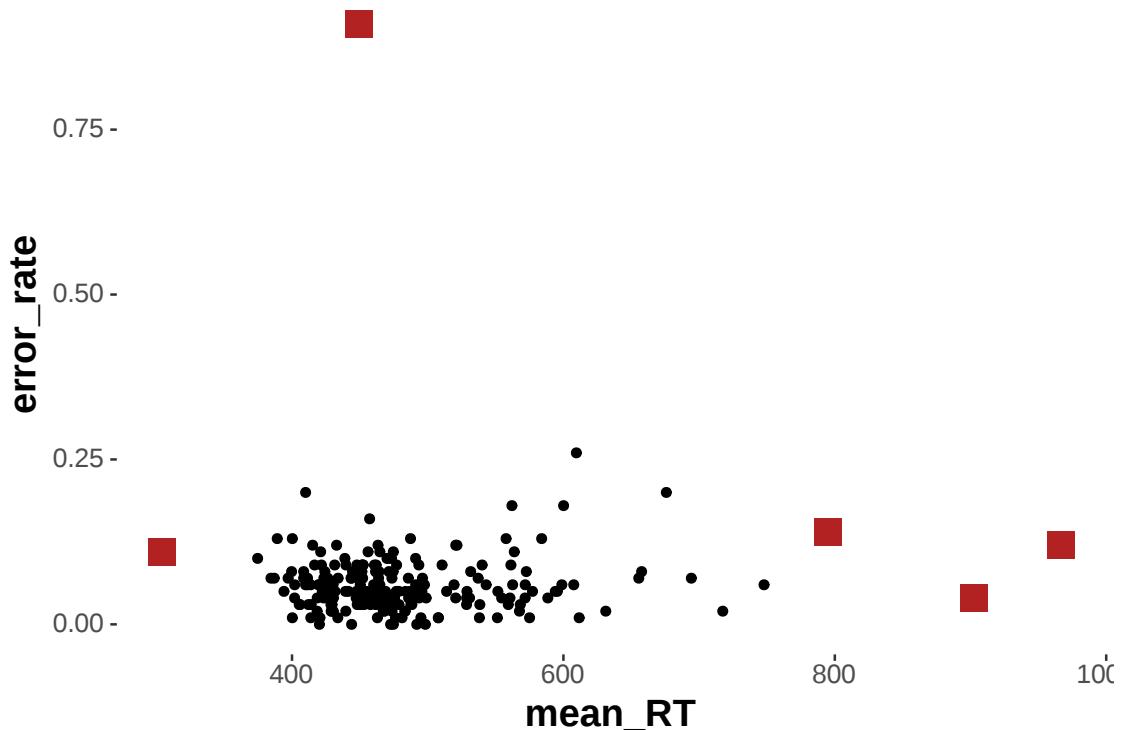


Here's a crude way of branding outlier-participants:

```
d_individual_summary <- d_individual_summary %>%
  mutate(outlier = case_when(mean_RT < 350 ~ TRUE,
                             mean_RT > 750 ~ TRUE,
                             error_rate > 0.5 ~ TRUE,
                             TRUE ~ FALSE))

d_individual_summary %>%
  ggplot(aes(x = mean_RT, y = error_rate)) +
  geom_point() +
  geom_point(data = filter(d_individual_summary, outlier == TRUE),
             color = "firebrick", shape = "square", size = 5)
```

D. Data sets used in the book



We then clean the data set in a first step by removing all participants identified as outlier-y:

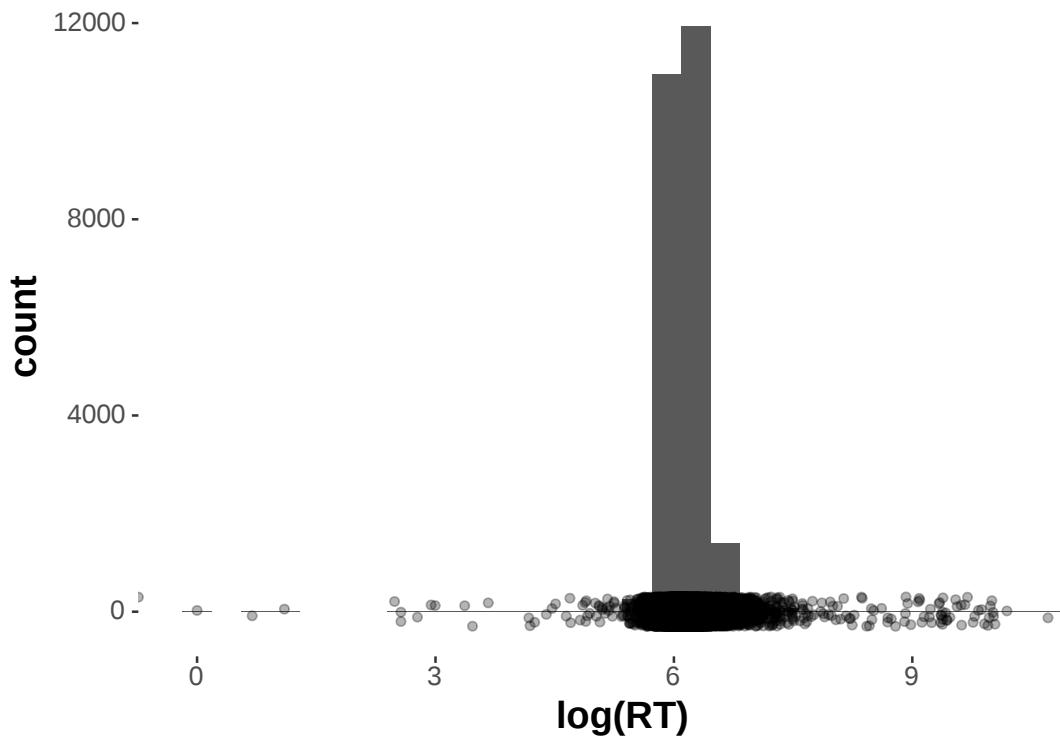
```
d <- full_join(d, d_individual_summary, by = "submission_id") # merge the tibbles
d <- filter(d, outlier == FALSE)
message("We excluded ", sum(d_individual_summary$outlier) , " participants for suspicious mean RTs and higher error rates." )
```

```
## We excluded 5 participants for suspicious mean RTs and higher error rates.
```

D.2.2.2. Trial-level reaction times

It is also conceivable that individual trials resulted in early accidental key presses or were interrupted in some way or another. We therefore look at the overall distribution of RTs and determine (similarly arbitrarily, but once again this should be planned in advance) what to exclude.

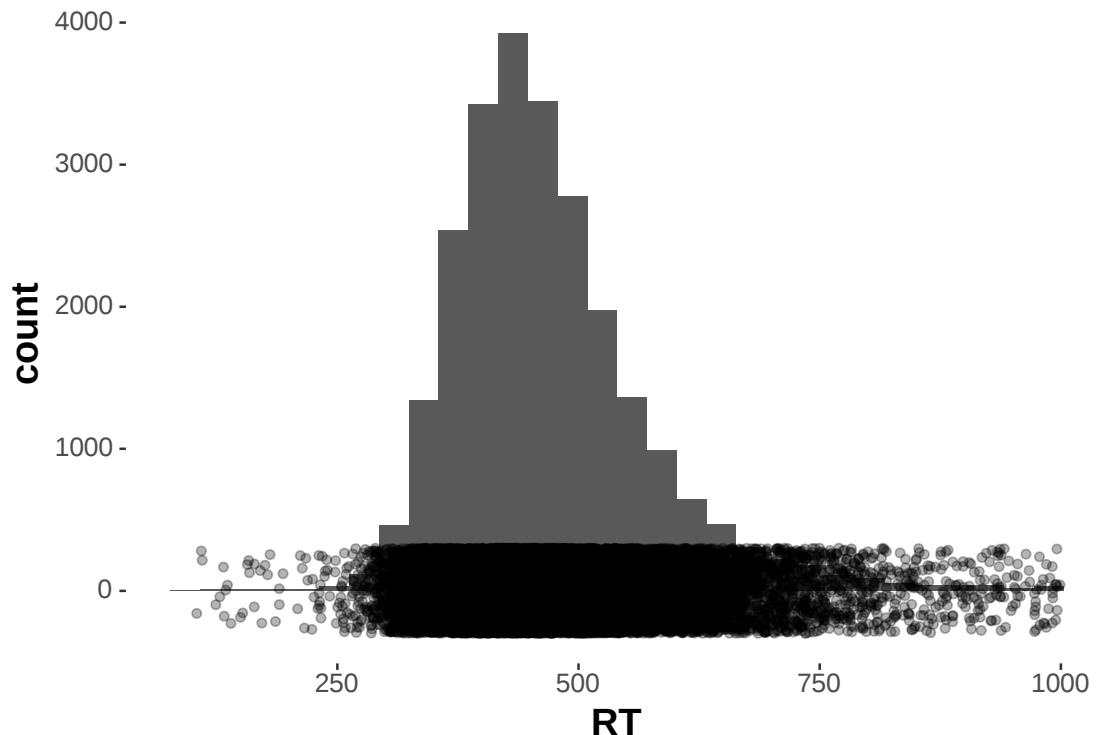
```
d %>% ggplot(aes(x = log(RT))) +
  geom_histogram() +
  geom_jitter(aes(x = log(RT), y = 1), alpha = 0.3, height = 300)
```



Let's decide to exclude all trials that lasted longer than 1 second and also all trials with reaction times under 100 ms.

```
d <- filter(d, RT > 100 & RT < 1000)
d %>% ggplot(aes(x = RT)) +
  geom_histogram() +
  geom_jitter(aes(x = RT, y = 1), alpha = 0.3, height = 300)
```

D. Data sets used in the book



D.2.2.3. Exploring the (main) data

We are mostly interested in the influence of congruency on the reactions times in the trials where participants gave a correct answer. But here we also look at, for comparison, the reaction times for the incongruent trials.

Here is a summary of the means and standard deviations for each condition:

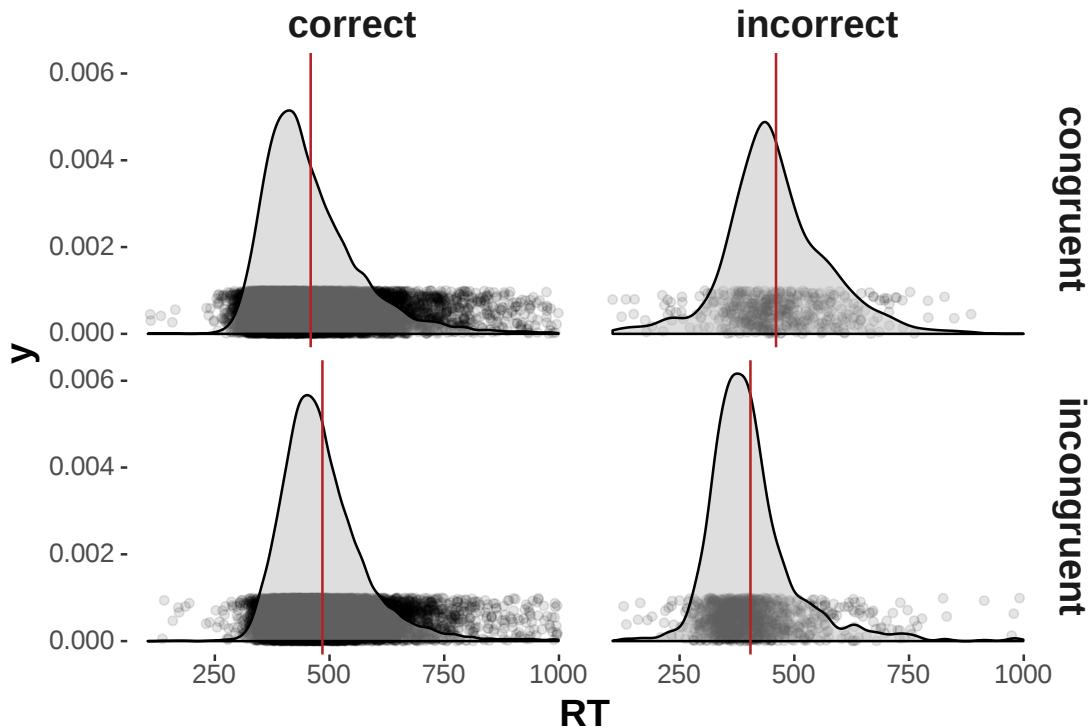
```
d_sum <- d %>%
  group_by(correctness, condition) %>%
  summarize(mean_RT = mean(RT),
            sd_RT = sd(RT))
d_sum

## # A tibble: 4 x 4
## # Groups:   correctness [2]
##   correctness condition   mean_RT   sd_RT
##   <chr>        <chr>     <dbl>    <dbl>
## 1 correct      congruent    459.    105.
## 2 correct      incongruent  484.    91.9
## 3 incorrect    congruent    460.    111.
```

```
## 4 incorrect    incongruent     404.   95.4
```

Here's a plot of the reaction times split up by whether the answer was correct and whether the trial was congruent or incongruent.

```
d %>% ggplot(aes(x = RT)) +
  geom_jitter(aes(y = 0.0005), alpha = 0.1, height = 0.0005) +
  geom_density(fill = "gray", alpha = 0.5) +
  geom_vline(data = d_sum,
             mapping = aes(xintercept = mean_RT),
             color = "firebrick") +
  facet_grid(condition ~ correctness)
```



D.2.3. Analysis

We are interested in comparing the RTs of correct answers in the congruent and incongruent conditions. We saw a difference in mean reaction times, but we'd like to know if this difference is meaningful. One way of testing this is by running a regression model, which tries to predict RTs as a function of congruency.

In the simplest case we would therefore do this:

D. Data sets used in the book

```
model_ST_simple = brm(RT ~ condition, filter(d, correctness == "correct"))
summary(model_ST_simple)
```

According to this analysis, there is reason to believe in a difference in RTs between congruent and incongruent groups. The coefficient estimated for the incongruent group is on average ca. 25 ms higher than that of the congruent group.

However, we can also look at the interaction between correctness and condition. As shown in the above graph, there are four different cells in a 2x2 grid.

In the below model, this is coded with 'dummy coding' such that the top-left cell (congruent-correct) is the intercept, and each other cell is calculated by the addition of offsets.

```
model_ST_complex <- brm(RT ~ condition * correctness, d)
```

We may want to ask the question: are reaction times to correct-congruent responses shorter than reaction times to incorrect-incongruent responses?

To do this, we first need to extract the posterior samples from our model.

```
post_samples <- posterior_samples(model_ST_complex) %>%
  as_tibble()
```

Then we need to determine the correct offsets to match the correct-congruent and incorrect-incongruent cells in the design matrix.

```
# correct-congruent is the reference cell
correct_congruent <- post_samples$b_Intercept

# incorrect_incongruent is the bottom-right cell
incorrect_incongruent <- post_samples$b_Intercept +
  post_samples$b_conditionincongruent +
  post_samples$b_correctnessincorrect +
  post_samples$b_conditionincongruent:correctnessincorrect`
```

Once we know these, we can calculate the probability that the comparison is in the correct direction.

```
mean(correct_congruent < incorrect_incongruent)
```

D.3. World Values Survey (wave 6 | 2010-2014)

D.3.1. Nature, origin and rationale of the data

The World Values Survey (WVS) aims to study *changing values and their impact on social and political life*. The WVS consists of nationally representative surveys conducted in almost 100 countries which contain

almost 90 percent of the world's population, using a common questionnaire. The WVS is the largest non-commercial, cross-national, time series investigation of human beliefs and values.

It currently includes interviews with almost 400,000 respondents. Respondents are people in the age 18 and older residing within private households in each country, regardless of their nationality, citizenship or language.

The main method of data collection in the WVS survey is *face-to-face interview* at respondent's home / place of residence.

```

->
-> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> -> ->
->
->
->
->
-> -> ->
-> -> -> -> -> -> -> -> ->
-> ->
-> ->

```

D.4. King of France

D.4.1. Nature, origin and rationale of the data

A **presupposition** of a sentence is a piece of information that is necessary for the sentence to make sense, but which is not communicated explicitly. If I say "Jones chained my camel to a tree", this sentence presupposes, somewhat incredibly, that I own a camel. If it is false that I own a camel, the sentence makes no sense. Yet, if I say it and you say: "I disagree" you take issue with my claim about chaining, not about me owning a camel. In this sense, the presupposition is not part of the explicitly contributed content (it is "not at issue content", as the linguists would say).

We here partially replicate a previous study by Abrusán and Szendrői (2013) investigating how sentences with false presuppositions are perceived. The main question of interest for us is whether sentences with a false presupposition are rather regarded as true or rather as false. We therefore present participants with sentences (see below) and have them rate these as 'true' or 'false', a so-called **truth-value judgement task**, a common paradigm in experimental semantics and pragmatics. (The original study by Abrusán and Szendrői (2013) also included a third option 'cannot tell', which we do not use, since this data set is mainly used for toying around with binary choice data.)

Abrusán and Szendrői (2013) presented their participants with 11 different types of sentences, of which we here only focus on five. Here are examples of the five conditions we test, using the corresponding condition numbers from the experiment by Abrusán and Szendrői (2013).

D. Data sets used in the book

C0. The king of France is bald.

C1. France has a king, and he is bald.

C6. The King of France isn't bald.

C9. The King of France, he did not call Emmanuel Macron last night.

C10. Emmanuel Macron, he did not call the King of France last night.

The presupposition in question is “France has a king”. C0 and C1 differ only with respect to whether this piece of information is presupposed (C1) or explicitly asserted (C0). The variants C0 and C6 differ only with respect to negation in the main (asserted) proposition. Finally, the contrast pair C9 and C10 is interesting because of a particular topic-focus structure and the placement of negation. In C9 the topic is “the King of France” which introduces the presupposition in question. In C10 the topic is “Emmanuel Macron”, but it introduces the presupposition under a negation.

Figure D.1 shows the results reported by Abrusán and Szendrői (2013).

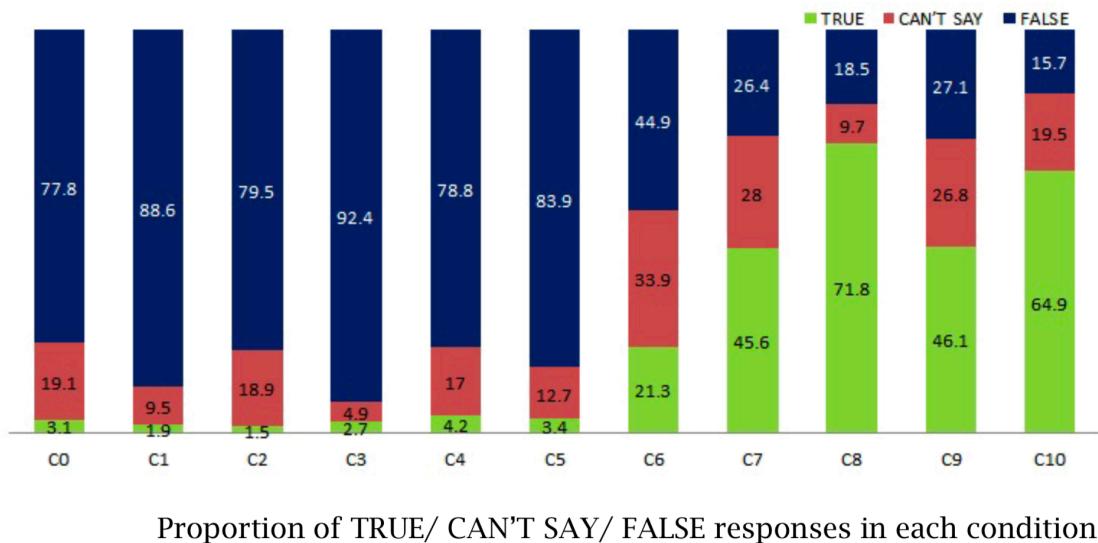


Figure D.1.: Results of @AbrusanSzendoi2013:Experimenting-w .

D.4.1.1. The experiment

D.4.1.1.1. Participants

We obtained data from 97 participants via the online crowd-sourcing platform Prolific.² All participants were native speakers of English.

²We recruited 100 participants, but the data from three participants was not recorded due to technical problems.

D.4.1.1.2. Material

The sentence material consisted of five vignettes. Here are the sentences that constitute “condition 1” of each of the five vignettes:

V1. The King of France is bald.

V2. The Emperor of Canada is fond of sushi.

V3. The Pope’s wife is a lawyer.

V4. The Belgian rainforest provides a habitat for many species.

V5. The volcanoes of Germany dominate the landscape.

As every vignette occurred in each of the five conditions, there are a total of 25 critical sentences.

Additionally, for each vignette, there is a “background check” sentence which is intended to find out whether participants know whether the relevant presuppositions are true. The “background check” sentences are:

BC1. France has a king.

BC2. The Pope is currently not married.

BC3. Canada is a democracy.

BC4. Belgium has rainforests.

BC5. Germany has volcanoes.

Finally, there are also 110 filler sentences, which do not have a presupposition, but also require common world knowledge for a correct answer. As each filler has an uncontroversially correct answer, these fillers also serve as a general attention check, to probe into whether participants are reading the sentences carefully enough. Example filler sentences are:

F1. William Shakespeare was a famous Italian painter in Rome.

F2. There were two world wars in the 20th century.

D.4.1.1.3. Procedure

Each experimental run started with five practice trials, which used the five additional sentences, which were like the filler material and the same for each participant, presented in random order.

The main part of the experiment presented each participant with five critical sentences, exactly one from each vignette and exactly one from each condition, allocated completely at random. Each participant also saw all of the five “background check” sentences. Each “background check” sentence was presented after the corresponding vignette’s critical sentence. All of these test trials were interspersed with 14 random filler sentences.

D.4.1.1.4. Realization

The experiment was realized using _magpie and can be tried out here.

D. Data sets used in the book

D.4.1.2. Theoretical motivation & hypotheses

We will be concerned with the following two research questions.³

1. Is the overall rate (= aggregating over all vignettes & conditions) of "TRUE" judgements for sentences with presupposition failure different from pure guessing chance of 0.5?
2. Is there a difference in (binary) truth-value judgements (aggregated over all vignettes) between C0 (with presupposition) and C1 (where the presupposition is part of the at-issue / asserted content)?
3. Is there a difference in (binary) truth-value judgements (aggregated over all vignettes) between C0 (the positive sentence) and C6 (the negative sentence)?
4. Is there a difference in (binary) truth-value judgements (aggregated over all vignettes) between C9 (where the presupposition is topical) and C10 (where the presupposition is not topical and occurs under negation)?

D.4.2. Loading and preprocessing the data

First, load the data:

```
## Observations: 2,813
## Variables: 16
## $ submission_id  <dbl> 192, 192, 192, 192, 192, 192, 192, 192, 192, 19...
## $ RT              <dbl> 8110, 35557, 3647, 16037, 11816, 6024, 4986, 13019, ...
## $ age              <dbl> 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, ...
## $ comments         <chr> NA, ...
## $ item_version    <chr> "none", "none", "none", "none", "none", "none", "non...
## $ correct_answer   <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, ...
## $ education        <chr> "Graduated College", "Graduated College", "Graduated...
## $ gender            <chr> "female", "female", "female", "female", "f...
## $ languages         <chr> "English", "English", "English", "English", "English...
## $ question          <chr> "World War II was a global war that lasted from 1914...
## $ response          <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, ...
## $ timeSpent         <dbl> 39.48995, 39.48995, 39.48995, 39.48995, 39.48995, 39...
## $ trial_name        <chr> "practice_trials", "practice_trials", "practice_tria...
## $ trial_number      <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ trial_type        <chr> "practice", "practice", "practice", "practice", "pra...
## $ vignette          <chr> "undefined", "undefined", "undefined", "undefined", ...
```



```
data_KoF_raw <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intr...'))
```

And then have a glimpse:

³These research questions are a compromise between actual theoretical relevance and practical (= educational) considerations.

```

glimpse(data_KoF_raw)

## Observations: 2,813
## Variables: 16
## $ submission_id <dbl> 192, 192, 192, 192, 192, 192, 192, 192, 192, 19...
## $ RT <dbl> 8110, 35557, 3647, 16037, 11816, 6024, 4986, 13019, ...
## $ age <dbl> 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, 57, ...
## $ comments <chr> NA, ...
## $ item_version <chr> "none", "none", "none", "none", "none", "none", "none", ...
## $ correct_answer <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, ...
## $ education <chr> "Graduated College", "Graduated College", "Graduated...
## $ gender <chr> "female", "female", "female", "female", "female", "f...
## $ languages <chr> "English", "English", "English", "English", "English...
## $ question <chr> "World War II was a global war that lasted from 1914...
## $ response <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, ...
## $ timeSpent <dbl> 39.48995, 39.48995, 39.48995, 39.48995, 39.48995, 39...
## $ trial_name <chr> "practice_trials", "practice_trials", "practice_tria...
## $ trial_number <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ trial_type <chr> "practice", "practice", "practice", "practice", "pra...
## $ vignette <chr> "undefined", "undefined", "undefined", "undefined", ...

```

The most important variables in this data set are:

- `submission_id`: unique identifier for each participant
- `trial_type`: whether the trial was of the category `filler`, `main`, `practice` or `special`, where the latter encodes the “background checks”
- `item_version`: the condition which the test sentence belongs to (only given for trials of type `main` and `special`)
- `response`: the answer (“TRUE” or “FALSE”) on each trial
- `vignette`: the current item’s vignette number (applies only to trials of type `main` and `special`)

As the variable names used in the raw data are not ideal, we will pre-process the raw data a bit for easier analysis.

```

data_KoF_processed <- data_KoF_raw %>%
  # discard practice trials
  filter(trial_type != "practice") %>%
  mutate(
    # add a 'condition' variable
    condition = case_when(
      trial_type == "special" ~ "background check",
      trial_type == "main" ~ str_c("Condition ", item_version),
      TRUE ~ "filler"
    )
  )

```

D. Data sets used in the book

```
) %>%
  factor(
    ordered = T,
    levels = c(str_c("Condition ", c(0, 1, 6, 9, 10)), "background check", "filler")
  )
# write_csv(data_KoF_processed, "data_sets/king-of-france_data_processed.csv")
```

D.4.3. Cleaning the data

We clean the data in two consecutive steps:

1. Remove all data from any participant who got more than 50% of the answer to filler material wrong.
2. Remove individual main trials if the corresponding “background check” question was answered wrongly.

D.4.3.1. Cleaning by-participant

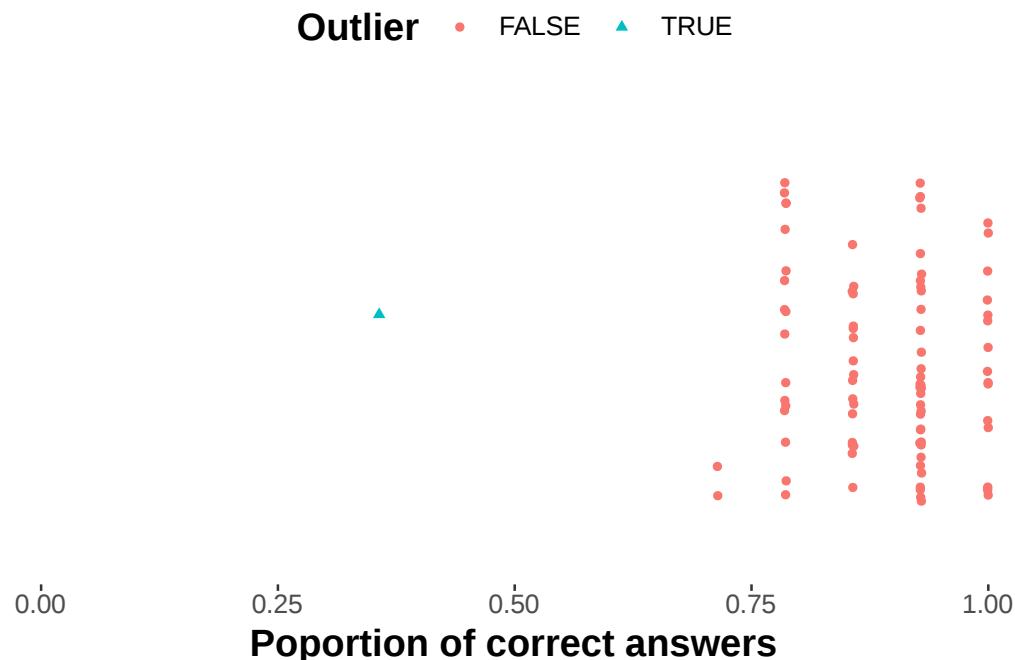
```
# look at error rates for filler sentences by subject
# mark every subject with < 0.5 proportion correct

subject_error_rate <- data_KoF_processed %>%
  filter(trial_type == "filler") %>%
  group_by(submission_id) %>%
  summarise(
    proportion_correct = mean(correct_answer == response),
    outlier_subject = proportion_correct < 0.5
  ) %>%
  arrange(proportion_correct)
```

Plot the results:

```
# plot by-subject error rates
subject_error_rate %>%
  ggplot(aes(x = proportion_correct, color = outlier_subject, shape = outlier_subject),
  geom_jitter(aes(y = ""), width = 0.001) +
  xlab("Poportion of correct answers") + ylab("") +
  ggtitle("Distribution of proportion of correct answers on filler trials") +
  xlim(0,1) +
  scale_color_discrete(name = "Outlier") +
  scale_shape_discrete(name = "Outlier")
```

Distribution of proportion of correct answers on 1



Apply the cleaning step:

```
# add info about error rates and exclude outlier subject(s)
d_cleaned <-
  full_join(data_KoF_processed, subject_error_rate, by = "submission_id") %>%
  filter(outlier_subject == FALSE)
```

D.4.3.2. Cleaning by-trial

```
# exclude every critical trial whose 'background' test question was answered wrongly

d_cleaned <-
  d_cleaned %>%
  # select only the 'background question' trials
  filter(trial_type == "special") %>%
  # is the background question answered correctly?
  mutate(
    background_correct = correct_answer == response
  ) %>%
```

D. Data sets used in the book

```
# select only the relevant columns
select(submission_id, vignette, background_correct) %>%
# right join lines to original data set
right_join(d_cleaned, by = c("submission_id", "vignette")) %>%
# remove all special trials, as well as main trials with incorrect background che
filter(trial_type == "main" & background_correct == TRUE)

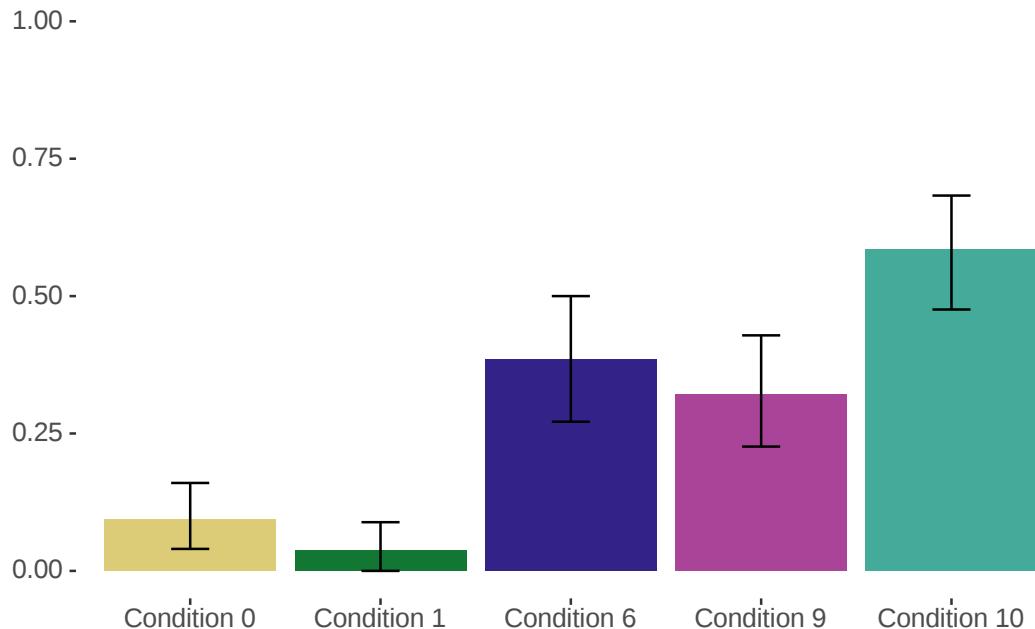
# write_csv(d_cleaned, "data_sets/king-of-france_data_cleaned.csv")
```

D.4.4. Exploration: summary stats & plots

Plot for ratings by condition:

```
d_cleaned %>%
# drop unused factor levels
droplevels() %>%
# get means and 95% bootstrapped CIs for each condition
group_by(condition) %>%
nest() %>%
summarise(
  CIs = map(data, function(d) bootstrapped_CI(d$response == "TRUE")))
) %>%
unnest(CIs) %>%
# plot means and CIs
ggplot(aes(x = condition, y = mean, fill = condition)) +
geom_bar(stat = "identity") +
geom_errorbar(aes(ymin = lower, ymax = upper, width = 0.2)) +
ylim(0,1) +
ylab("") + xlab("") + ggtitle("Proportion of 'TRUE' responses per condition") +
theme(legend.position = "none") +
scale_fill_manual(values = project_colors)
```

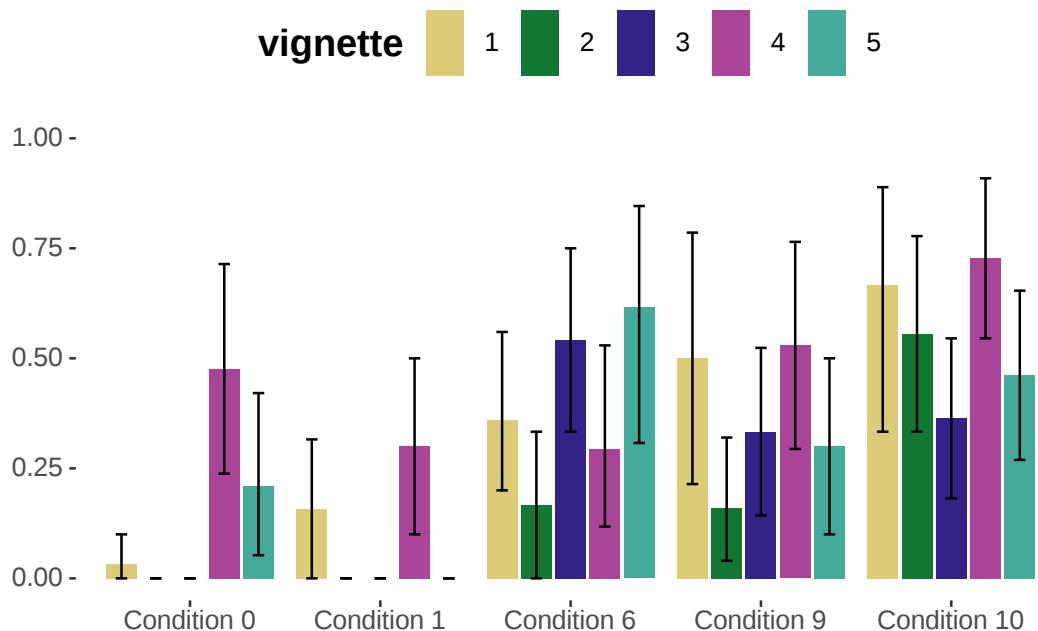
Proportion of 'TRUE' responses per condition



Plot for each condition & vignette:

```
data_KoF_processed %>%
  filter(trial_type == "main") %>%
  droplevels() %>%
  group_by(condition, vignette) %>%
  nest() %>%
  summarise(
    CIs = map(data, function(d) bootstrapped_CI(d$response == "TRUE")))
  ) %>%
  unnest(CIs) %>%
  ggplot(aes(x = condition, y = mean, fill = vignette)) +
  geom_bar(stat = "identity", position = "dodge2") +
  geom_errorbar(
    aes(ymin = lower, ymax = upper),
    width = 0.3,
    position = position_dodge(width = 0.9)
  ) +
  ylim(0,1) +
  ylab("") + xlab("") + ggtitle("Proportion of 'TRUE' responses per condition & vignette")
```

Proportion of 'TRUE' responses per condition



D.4.5. Data analysis

D.5. Bio-Logic Jazz-Metal (and where to consume it)

D.5.1. Nature, origin and rationale of the data

This is a very short and non-serious experiment that asks for just three binary decisions from each participant, namely their spontaneous preference for one of two presented options (biology vs logic, jazz vs metal and mountains vs beach). The data from this experiment will be analyzed and plotted. This is supposed to be a useful and hopefully entertaining self-generated data set with which to practice making contingency tables and to apply binomial tests and fun stuff like that.

D.5.1.1. The experiment

D.5.1.1.1. Participants

We obtained data from 102 participants of this very course.

D.5.1.1.2. Material

There were three critical trials (and nothing else). All trials had the same trailing question:

If you have to choose between the following two options, which one do you prefer?

Each critical trial then presented two options as buttons, one of which had to be clicked.

1. Biology vs Logic
2. Jazz vs Metal
3. Mountains vs Beach

D.5.1.1.3. Procedure

Each participant saw all three critical trials (and no other trials) in random order.

D.5.1.1.4. Realization

The experiment was realized using `_magpie` and can be tried out here.

D.5.1.2. Theoretical motivation & hypotheses

This is a bogus experiment, and no sane person would advance a serious hypothesis about this. Except that, actually, the lecturer is careless enough to conjecture that appreciators of Metal music like logic more than Jazz-enthusiasts would (because Metal is cleaner and more mechanic, while Jazz is fuzzy and organic, obviously).⁴

D.5.2. Loading and preprocessing the data

First, load the data:

```
## Observations: 306
## Variables: 19
## $ submission_id <dbl> 379, 379, 379, 378, 378, 378, 377, 377, 377, 376, 376...
## $ QUD <lgl> NA, N...
## $ RT <dbl> 9230, 9330, 5248, 5570, 2896, 36236, 5906, 4767, 1042...
## $ age <dbl> 30, 30, 30, 29, 29, 29, 20, 20, 20, 21, 21, 21, 23, 2...
## $ comments <chr> NA, N...
## $ education <chr> "Graduated High School", "Graduated High School", "Gr...
## $ endTime <dbl> 1.573751e+12, 1.573751e+12, 1.573751e+12, 1.573738e+1...
## $ experiment_id <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, ...
## $ gender <chr> "male", "male", "male", "male", "male", "fema...
## $ languages <chr> "German", "German", "German", "German", "German", "Ge...
## $ option1 <chr> "Mountains", "Biology", "Metal", "Metal", "Biology", ...
## $ option2 <chr> "Beach", "Logic", "Jazz", "Jazz", "Logic", "Beach", "...
```

⁴Notice how easy it is to motivate any-old psychological theory. Some other scientific disciplines are much better at smothering nonsensical ideas from the start.

D. Data sets used in the book

```
## $ question      <chr> "If you have to choose between the following two opti...
## $ response       <chr> "Beach", "Logic", "Metal", "Metal", "Logic", "Beach",...
## $ startDate      <chr> "Thu Nov 14 2019 18:01:24 GMT+0100 (CET)", "Thu Nov 1...
## $ startTime      <dbl> 1.573751e+12, 1.573751e+12, 1.573751e+12, 1.573738e+1...
## $ timeSpent     <dbl> 2.3601500, 2.3601500, 2.3601500, 2.1552667, 2.1552667...
## $ trial_name     <chr> "forced_choice", "forced_choice", "forced_choice", "f...
## $ trial_number   <dbl> 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3,...
```

```
data_BLJM_raw <- read_csv(url('https://raw.githubusercontent.com/michael-franke/int...'))
```

Take a peak:

```
glimpse(data_BLJM_raw)
```

```
## Observations: 306
## Variables: 19
## $ submission_id <dbl> 379, 379, 379, 378, 378, 378, 377, 377, 377, 376, 376...
## $ QUD            <lgl> NA, N...
## $ RT             <dbl> 9230, 9330, 5248, 5570, 2896, 36236, 5906, 4767, 1042...
## $ age            <dbl> 30, 30, 30, 29, 29, 29, 20, 20, 20, 21, 21, 21, 23, 2...
## $ comments        <chr> NA, N...
## $ education       <chr> "Graduated High School", "Graduated High School", "Gr...
## $ endTime         <dbl> 1.573751e+12, 1.573751e+12, 1.573751e+12, 1.573738e+1...
## $ experiment_id   <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, ...
## $ gender          <chr> "male", "male", "male", "male", "male", "fema...
## $ languages        <chr> "German", "German", "German", "German", "German", "Ge...
## $ option1          <chr> "Mountains", "Biology", "Metal", "Metal", "Biology", ...
## $ option2          <chr> "Beach", "Logic", "Jazz", "Jazz", "Logic", "Beach", ...
## $ question         <chr> "If you have to choose between the following two opti...
## $ response         <chr> "Beach", "Logic", "Metal", "Metal", "Logic", "Beach",...
## $ startDate        <chr> "Thu Nov 14 2019 18:01:24 GMT+0100 (CET)", "Thu Nov 1...
## $ startTime        <dbl> 1.573751e+12, 1.573751e+12, 1.573751e+12, 1.573738e+1...
## $ timeSpent       <dbl> 2.3601500, 2.3601500, 2.3601500, 2.1552667, 2.1552667...
## $ trial_name       <chr> "forced_choice", "forced_choice", "forced_choice", "f...
## $ trial_number     <dbl> 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3,...
```

The most important variables in this data set are:

- `submission_id`: unique identifier for each participant
- `option1` and `option2`: what the choice options were
- `response`: which of the two options was chosen

Notice that there is no convenient column indicating which of the three critical conditions we are dealing with, so we create that information from the information in columns option1 and option2, while also discarding everything we will not need:

```
data_BLJM_processed <-
  data_BLJM_raw %>%
  mutate(
    condition = str_c(str_sub(option2, 1, 1), str_sub(option1, 1, 1))
  ) %>%
  select(submission_id, condition, response)
# write_csv(data_BLJM_processed, "../data_sets/bio-logic-jazz-metal-data-processed.csv")
data_BLJM_processed

## # A tibble: 306 x 3
##   submission_id condition response
##       <dbl> <chr>     <chr>
## 1         379 BM        Beach
## 2         379 LB        Logic
## 3         379 JM        Metal
## 4         378 JM        Metal
## 5         378 LB        Logic
## 6         378 BM        Beach
## 7         377 BM        Mountains
## 8         377 LB        Biology
## 9         377 JM        Jazz
## 10        376 BM        Beach
## # ... with 296 more rows
```

D.5.3. Exploration: counts & plots

We are interest in some counts. First, let's look at the overal choice rates in each condition:

```
data_BLJM_processed %>%
  # function `count` is masked by another package, must call explicitly
  dplyr::count(condition, response)

## # A tibble: 6 x 3
##   condition response     n
##   <chr>     <chr>   <int>
## 1 BM        Beach      44
## 2 BM        Mountains  58
## 3 JM        Jazz       64
```

D. Data sets used in the book

```
## 4 JM      Metal      38
## 5 LB      Biology    58
## 6 LB      Logic     44
```

Overall it seems that mountains are preferred over beaches, Jazz is preferred over Metal and Biology is preferred over Logic.

The overall counts, however, do not tell us anything about any potentially interesting relationship between preferences. So, let's have a closer look at the lecturer's conjecture that a preference for logic tends to go with a stronger preference for metal than a preference for biology does. To check this, we need to look at different counts, namely the number of people who selected which music-subject pair. We collect these counts in variable `BLJM_associated_counts`:

```
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts

## # A tibble: 4 x 3
##   JM     LB       n
##   <chr> <chr>   <int>
## 1 Jazz   Biology  38
## 2 Jazz   Logic    26
## 3 Metal  Biology  20
## 4 Metal  Logic    18
```

Notice that this representation is tidy, but not as ideal for visual inspection. A more commonly seen format can be obtained by pivoting to a wider representation:

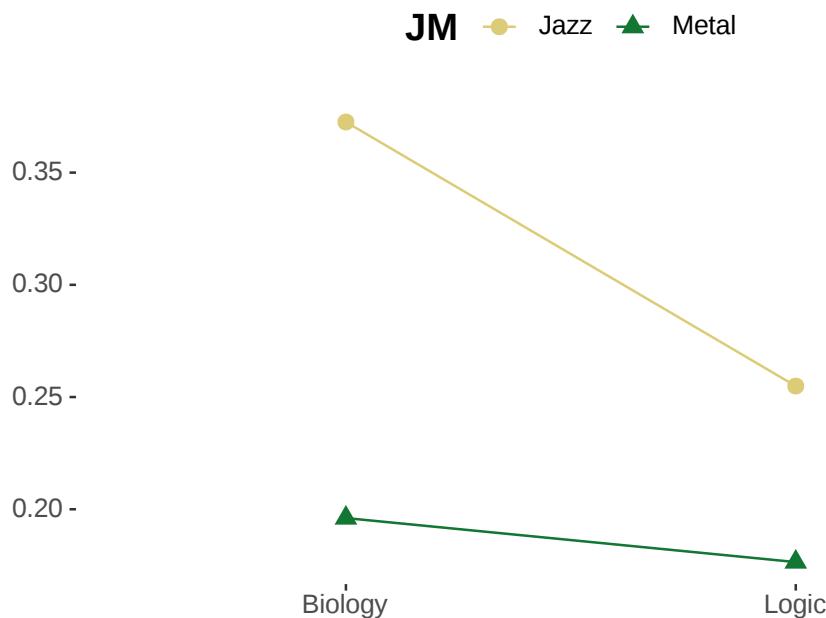
```
# visually attractive table representation
BLJM_associated_counts %>%
  pivot_wider(names_from = LB, values_from = n)

## # A tibble: 2 x 3
##   JM     Biology Logic
##   <chr>   <int> <int>
## 1 Jazz      38    26
## 2 Metal     20    18
```

The tidy representation is ideal for plotting, though. Notice, however, that the code below plots proportions of choices, not raw counts:

```
BLJM_associated_counts %>%
  ggplot(aes(x = LB, y = n/sum(n), color = JM, shape = JM, group = JM)) +
  geom_point(size = 3) + geom_line() +
  labs(
    title = "Proportion of choices of each music+subject pair",
    x = "",
    y = ""
)
```

Proportion of choices of each music+subject p



The lecturer's conjecture might be correct. This does look like there could be an interaction. While Jazz is preferred more generally, the preference for Jazz over Metal seems more pronounced for those participants who preferred Biology than for those who preferred Logic.

D.6. Avocado prices

D.6.1. Nature, origin and rationale of the data

This data set has been plucked from kaggle. More information on the origin and composition of this data set can be found on kaggle's website covering the avocado data. The data set includes information about

D. Data sets used in the book

the prices of (Hass) avocados and the amount sold (of different kinds) at different points in time. The data is originally from the Hass Avocado Board where the data is described as follows:

The [data] represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers' cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU's) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

Columns of interest are:

- Date - date of the observation
- AveragePrice - average price of a single avocado
- Total Volume - Total number of avocados sold
- type - whether the price/amount is for conventional or organic
- 4046 - Total number of small avocados sold (PLU 4046)
- 4225 - Total number of medium avocados sold (PLU 4225)
- 4770 - Total number of large avocados sold (PLU 4770)

D.6.2. Loading and preprocessing the data

We load the data into a variable named `avocado_data` but also immediately rename some of the columns to have more convenient handles:

```
avocado_data <- read_csv(url('https://raw.githubusercontent.com/michael-franke/intro-to-data-science/master/datasets/avocado.csv'))  
# remove currently irrelevant columns  
select( -X1 , - contains("Bags") , - year , - region) %>%  
# rename variables of interest for convenience  
rename(  
  total_volume_sold = `Total Volume` ,  
  average_price = `AveragePrice` ,  
  small = '4046' ,  
  medium = '4225' ,  
  large = '4770' ,  
)
```

We can then take a glimpse:

```
glimpse(avocado_data)
```

```
## Observations: 18,249
## Variables: 7
## $ Date              <date> 2015-12-27, 2015-12-20, 2015-12-13, 2015-12-06, ...
## $ average_price     <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1...
## $ total_volume_sold <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60...
## $ small             <dbl> 1036.74, 674.28, 794.70, 1132.00, 941.48, 1184.27...
## $ medium            <dbl> 54454.85, 44638.81, 109149.67, 71976.41, 43838.39...
## $ large              <dbl> 48.16, 58.33, 130.50, 72.58, 75.78, 43.61, 93.26, ...
## $ type               <chr> "conventional", "conventional", "conventional", "...
```

D.6.3. Summary statistics

We are interested in the following summary statistics for the variables `total_amount_sold` and `average_price` for the whole data and for each type of avocado separately:

- mean
- median
- variance
- the bootstrapped 95% confidence interval of the mean

To get these results we define a convenience function that calculates exactly these measures:

```
summary_stats_convenience_fct <- function(numeric_data_vector) {
  bootstrap_results <- bootstrapped_CI(numeric_data_vector)
  tibble(
    CI_lower = bootstrap_results$lower,
    mean = bootstrap_results$mean,
    CI_upper = bootstrap_results$upper,
    median = median(numeric_data_vector),
    var = var(numeric_data_vector)
  )
}
```

We then apply this function once for the whole data set and once for each type of avocado (conventional or organic). We do this using a nested tibble in order to record the joint output of the convenience function (so that we only need to calculate the bootstrapped 95% confidence interval twice).

```
# summary stats for the whole data taken together
avocado_sum_stats_total <- avocado_data %>%
  select(type, average_price, total_volume_sold) %>%
  pivot_longer(
    cols = c(total_volume_sold, average_price),
    names_to = 'variable',
```

D. Data sets used in the book

```

    values_to = 'value'
) %>%
group_by(variable) %>%
nest() %>%
summarise(
  summary_stats = map(data, function(d) summary_stats_convenience_fct(d$value))
) %>%
unnest(summary_stats) %>%
mutate(type = "both_together") %>%
# reorder columns: moving `type` to second position
select(1,type,everything())

# summary stats for each type of avocado
avocado_sum_stats_by_type <- avocado_data %>%
  select(type, average_price, total_volume_sold) %>%
  pivot_longer(
    cols = c(total_volume_sold, average_price),
    names_to = 'variable',
    values_to = 'value'
) %>%
group_by(type, variable) %>%
nest() %>%
summarise(
  summary_stats = map(data, function(d) summary_stats_convenience_fct(d$value))
) %>%
unnest(summary_stats)

# joining the summary stats in a single tibble
avocado_sum_stats <-
  full_join(avocado_sum_stats_total, avocado_sum_stats_by_type)

# inspect the results
avocado_sum_stats

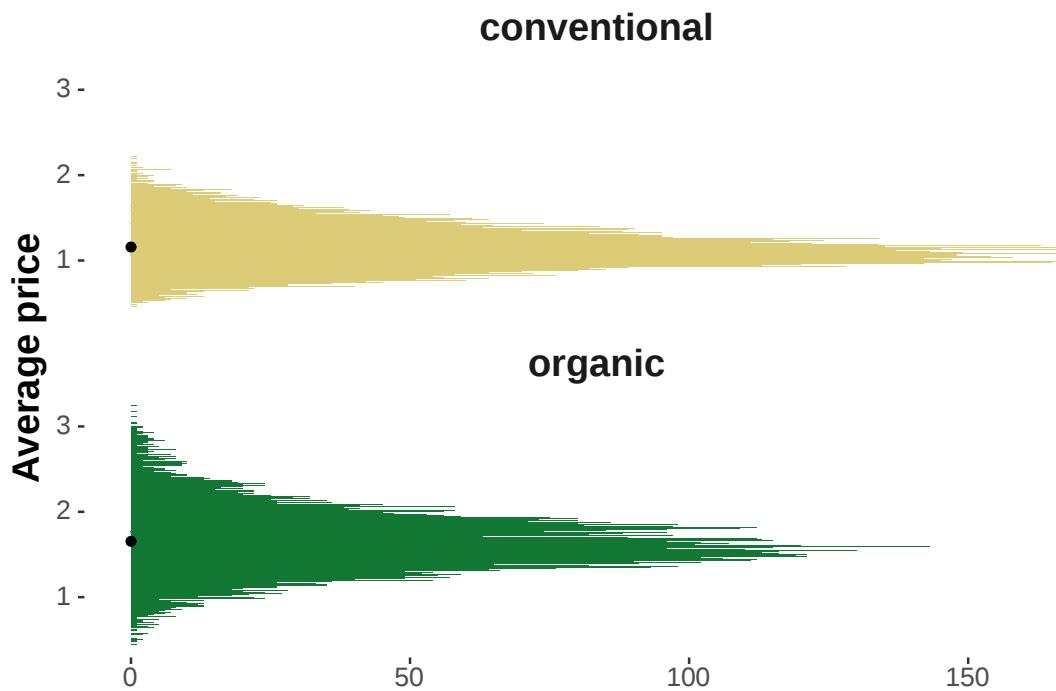
## # A tibble: 6 x 7
##   variable     type      CI_lower      mean      CI_upper      median      var
##   <chr>       <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 average_price both_together  1.40  1.41e0  1.41e0  1.37e0  1.62e- 1
## 2 total_volume_sold both_together 799701.  8.51e5  9.01e5  1.07e5  1.19e+13
## 3 average_price conventional  1.15  1.16e0  1.16e0  1.13e0  6.92e- 2
## 4 total_volume_sold conventional 1558132.  1.65e6  1.76e6  4.08e5  2.25e+13
## 5 average_price organic     1.65  1.65e0  1.66e0  1.63e0  1.32e- 1
## 6 total_volume_sold organic    44771.   4.78e4  5.07e4  1.08e4  2.03e+10

```

D.6.4. Plots

Here are plots of the distributions of average_price for different types of avocados:

```
avocado_data %>%
  ggplot(aes(x = average_price, fill = type)) +
  geom_histogram(binwidth = 0.01) +
  facet_wrap(type ~ ., ncol = 1) +
  coord_flip() +
  geom_point(
    data = avocado_sum_stats_by_type %>% filter(variable == "average_price"),
    aes(y = 0, x = mean)
  ) +
  ylab('') +
  xlab('Average price') +
  theme(legend.position = "none")
```

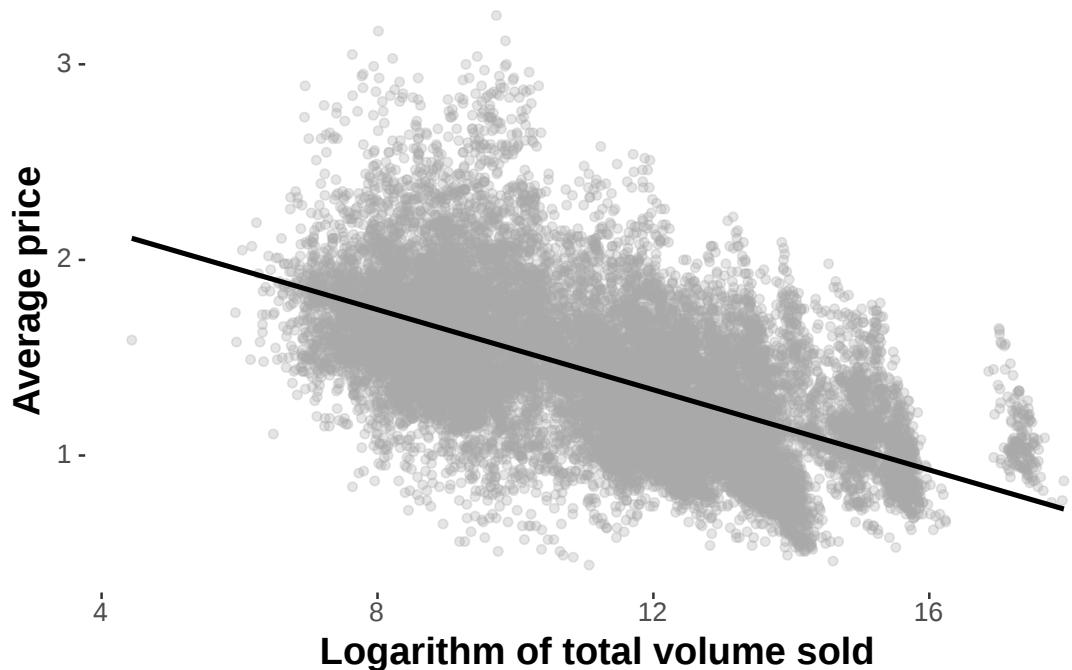


Here is a scatter plot of the logarithm of total_volume_sold against average_price:

D. Data sets used in the book

```
avocado_data %>%
  ggplot(aes(x = log(total_volume_sold), y = average_price)) +
  geom_point(color = "darkgray", alpha = 0.3) +
  geom_smooth(color = "black", method = "lm") +
  xlab('Logarithm of total volume sold') +
  ylab('Average price') +
  gtitle("Avocado prices plotted against the (log) amount sold")
```

Avocado prices plotted against the (log) amount



And another scatter plot, using a log-scaled *x*-axis and distinguishing different types of avocados are:

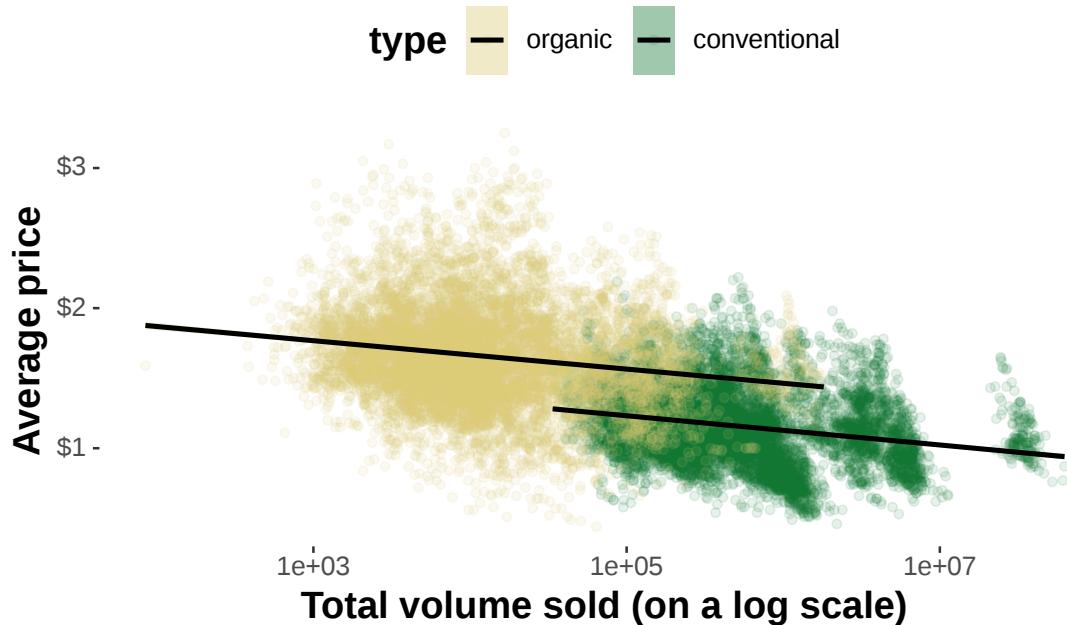
```
# pipe data set into function `ggplot`
avocado_data %>%
  # reverse factor level so that horizontal legend entries align with
  # the majority of observations of each group in the plot
  mutate(
    type = fct_rev(type)
  ) %>%
  # initialize the plot
  ggplot(
    # defined mapping
```

```

mapping = aes(
    # which variable goes on the x-axis
    x = total_volume_sold,
    # which variable goes on the y-axis
    y = average_price,
    # which groups of variables to distinguish
    group = type,
    # color and fill to change by grouping variable
    fill = type,
    color = type
)
) +
# declare that we want a scatter plot
geom_point(
    # set low opacity for each point
    alpha = 0.1
) +
# add a linear model fit (for each group)
geom_smooth(
    color = "black",
    method = "lm"
) +
# change the default (normal) of x-axis to log-scale
scale_x_log10() +
# add dollar signs to y-axis labels
scale_y_continuous(labels = scales::dollar) +
# change axis labels and plot title & subtitle
labs(
    x = 'Total volume sold (on a log scale)',
    y = 'Average price',
    title = "Avocado prices plotted against the amount sold per type",
    subtitle = "With linear regression lines"
)
)

```

Avocado prices plotted against the amount sold With linear regression lines



- Abrusán, Márta, and Kriszta Szendrői. 2013. "Experimenting with the King of France: Topics, Verifiability and Definite Descriptions." *Semantics & Pragmatics* 6 (1): 1–43.
- Academy, Khan. 2019. "Lagrange multipliers, introduction." 2019. <https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/constrained-optimization/a/lagrange-multipliers-single-constraint>.
- Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1). [American Statistical Association, Taylor & Francis, Ltd.]: 17–21. <https://doi.org/10.2307/2682899>.
- Blitzstein, Joseph K., and Jessica Hwang. 2014. *Introduction to Probability*. Chapman; Hall/CRC.
- Box, George E. P. 1979. "Robustness in the Strategy of Scientific Model Building." In *Robustness in Statistics*, edited by R. L. Launer and G. N. Wilkinson, 201–36. Cambridge, MA: Academic Press.
- Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer.
- Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. <https://doi.org/10.18637/jss.v080.i01>.
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1). <https://doi.org/10.18637/jss.v076.i01>.

- De Martino, Andrea, and Daniele De Martino. 2018. "An Introduction to the Maximum Entropy Approach and its Application to Inference Problems in Biology." *Heliyon* 4 (4). Elsevier.
- Finlayson, Sam. 2017. "Deriving probability distributions using the Principal of Maximum Entropy." 2017. <https://sgfin.github.io/2017/03/16/Deriving-probability-distributions-using-the-Principle-of-Maximum-Entropy/#introduction>.
- Franke, Michael, and Timo B. Roettger. 2019. "Bayesian Regression Modeling (for Factorial Designs): A Tutorial."
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Chapman; Hall/CRC.
- Gigerenzer, Gerd. 2004. "Mindless Statistics." *Journal of Social Economy* 33: 587–606.
- Golding, Nick. 2019. *greta: Simple and Scalable Statistical Modelling in R*. <https://CRAN.R-project.org/package=greta>.
- Goodman, Noah D, and Andreas Stuhlmüller. 2014. "The Design and Implementation of Probabilistic Programming Languages." <http://dippl.org>.
- Gronau, Quentin F., Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers, and Helen Steingrover. 2017. "A Tutorial on Bridge Sampling." *Journal of Mathematical Psychology* 81: 80–97.
- Haller, Heiko, and Stefan Krauss. 2002. "Misinterpretations of Significance: A Problem Students Share with Their Teachers?" *Methods of Psychological Research Online* 7 (1): 1–20.
- Halpern, Joseph Y. 2003. *Reasoning about Uncertainty*. MIT Press.
- Healy, Kieran. 2018. *Data Visualization*. New Jersey, USA: Princeton University Press.
- Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge university press.
- Keng, Brian. 2017. "Maximum Entropy Distributions." 2017. <http://bjlkeng.github.io/posts/maximum-entropy-distributions/>.
- Klingenberg, Bernhard. n.d. "Art of Stat." <http://www.artofstat.com/home.html>.
- Klugkist, Irene, Bernet Kato, and Herbert Hoijtink. 2005. "Bayesian Model Selection Using Encompassing Priors." *Statistica Neerlandica* 59 (1): 57–69.
- Kruschke, John. 2015. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- Lee, Michael D., and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.
- McElreath, Richard. 2015. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman; Hall/CRC.
- Myung, In Jae. 2003. "Tutorial on Maximum Likelihood Estimation." *Journal of Mathematical Psychology* 47: 90–100.
- Oh, Man-Suk. 2014. "Bayesian Comparison of Models with Inequality and Equality Constraints." *Statistics and Probability Letters* 84: 176–82.

D. Data sets used in the book

- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Reza, Fazlollah M. 1994. *An Introduction to Information Theory*. Courier Corporation.
- Rouder, Jeffrey N., Paul I. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. "Bayesian T Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin & Review* 16 (2): 225–37.
- Tufte, Edward. 1983. *The Visual Display of Quantitative Information*. Graphics Press.
- Vallverdú, Jordi. 2016. *Bayesians Versus Frequentists: A Philosophical Debate on Statistical Reasoning*. Heidelberg: Springer.
- Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57 (2): 307–33.
- Wagenmakers, Eric-Jan, and Simon Farrell. 2004. "AIC Model Selection Using Akaike Weights." *Psychonomic Bulletin & Review* 11 (1): 192–96.
- Wetzels, Ruud, Raoul P.P.P. Grasman, and Eric-Jan Wagenmakers. 2010. "An Encompassing Prior Generalization of the Savage–Dickey Density Ratio." *Computational Statistics and Data Analysis* 54: 2094–2102.
- Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28.
- . 2014. "Tidy Data." *Journal of Statistical Software* 59 (10).
- . 2017. *tidyverse: Easily Install and Load the 'Tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc.
- Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using R*. Routledge.