Universiteit Leiden
Science

# Statistics for Astronomy and Physics students

*Fall 2020*

Thomas Nagler

Version: February 22, 2021

# Contents

# 2
# Probability basics

> *Having observed some data $X_1, \dots, X_n$, what can we say about the mechanism that generated them?*

This question is the basic problem of statistics. In statistics, we assume that the mechanism involves some form of randomness. This is helpful even if the mechanism is genuinely deterministic. Here, we use 'randomness' to reflect our lack of knowledge. In many cases, this lack of knowledge cannot be resolved — because the mechanism is too complex for us to understand (e.g., interaction of all particles in the universe) or because a critical piece of information is not observed.

Before we can talk properly about statistics, we need a suitable language to talk about randomness. *Probability theory* is a mathematical field concerned with random phenomena and provides just that. In the following, we learn about basic concepts, definitions, and properties that build the foundation for statistical theory.

## 2.1 Sample spaces and events

In probability and statistics, random phenomena are often called *experiment*.

**Definition 2.1.** *(i) The* ***sample space*** *$\Omega$ is the set of all possible outcomes of an experiment.*

*(ii) Elements $\omega \in \Omega$ are called* ***outcomes*** *or* ***realizations****.*

*(iii) Subsets of $A \subset \Omega$ are called* ***events****.*

A more philosophical interpretation of the above is the following: The sample space $\Omega$ can be seen as all possible (hypothetical) states of our world. An element $\omega \in \Omega$ is a particular state of our world. An event is a collections of states that share some property.

Let's make this more concrete.

**Example 2.2.** *For the experiment consisting in tossing a coin twice,*

$$\Omega = \{HH, HT, TH, TT\},$$

*where $H$ stands for heads and $T$ for tails. The event that the first toss is heads is $A = \{HH, HT\}$.*

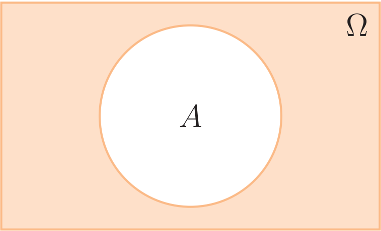
**Example 2.3.** *Let $\omega$ be the outcome of a measurement of some quantity, for instance stellar mass. Then $\Omega = (0, \infty)$. The event that the measurement is larger than $0.5$, but less than $1.2$ is $A = (0.5, 1.2)$.*

**Example 2.4.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \ldots, Dec\}$. The event that he was born in spring is $A = \{Mar, Apr, May\}$.*

Now consider two events: an earthquake ($A$) and a flood ($B$) defined on the sample space $\Omega$. The set $A$ contains all states of our world $\omega$, in which a earthquake happens. The set $B$ contains all $\omega$, for which a flood happens.

What is the event of no earthquake happening?

**Definition 2.5.** *For a given event $A$, let $A^c = \{\omega \in \Omega : \omega \notin A\}$. This is called the **complement** of $A$ and is thought of as the event 'not $A$'.*



To understand events, visualizations like the *Venn diagram* are often helpful. Think of the rectangle as the sample space $\Omega$. That is, all points $\omega$ inside the rectangle together form the set $\Omega$. The disk represents the event $A$. All points $\omega$ inside the disk form the set $A$. The event $A^c$ ('a does not happen') is defined as all outcomes in $\Omega$ that are not part of the set $A$ (the shaded area).

**Remark 2.1.** *The complement of $\Omega$ is an empty set $\emptyset$. Conversely, the complement of an empty set $\emptyset$ is $\Omega$. Convince yourself visually that this is true.*

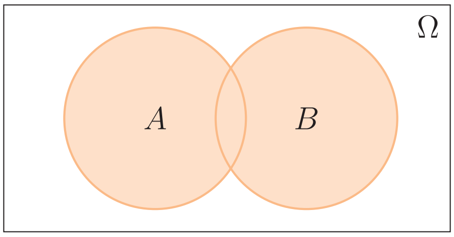
**Example 2.6.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \ldots, Dec\}$. The event that he was born in spring is $A = \{Mar, Apr, May\}$. The event that he was not born in spring is*

$$\begin{aligned} A^c &= \{\omega \in \Omega : \omega \notin A\} \\ &= \{Jan, Feb, Jun, Jul, \ldots, Dec\}. \end{aligned}$$

There are several ways to connect or disentangle two different events $A$ and $B$. First, what is the event that there is an earthquake *or* a flood?

**Definition 2.7.** *The **union** of the events $A$ and $B$, which can be thought of as an event 'A or B', is defined as*

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or both}\}.$$



The 'or' here is non-exclusive: $A$ and $B$ may both happen simultaneously, but we also accept it if just one of them does.

Similarly, or a possibly infinite sequence of events $A_1, A_2, A_3, \ldots$

$$\bigcup_i A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\}.$$

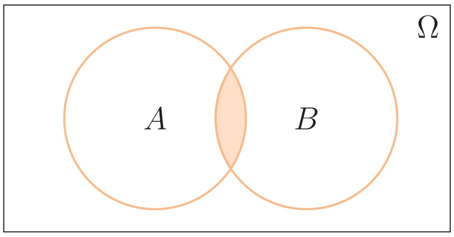
**Example 2.8.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \ldots, Dec\}$. The event that he was born in spring is $A = \{Mar, Apr, May\}$. The event that he was born in summer is $B = \{Jun, Jul, Aug\}$. The event that he was born in spring or summer is*

$$\begin{aligned} A \cup B &= \{Mar, Apr, May\} \cup \{Jun, Jul, Aug\} \\ &= \{Mar, Apr, May, Jun, Jul, Aug\} \end{aligned}$$

The second way to connect the two is: what is the event that there is both an earth quake *and* a flood?

**Definition 2.9.** *The **intersection** of $A$ and $B$ (an event 'A and B') is*

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B \text{ simultaneously}\}.$$



Similarly, for an infinite sequence of events $A_1, A_2, A_3, \ldots$

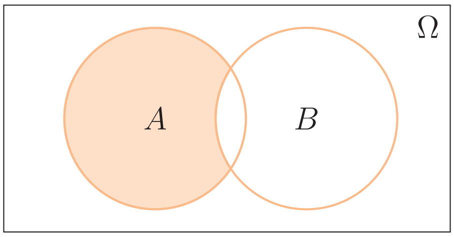
$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}.$$

**Example 2.10.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \ldots, Dec\}$. The event that he was born in winter is $A = \{Dec, Jan, Feb\}$. The event that he was born in the first half of a year is $B = \{Jan, \ldots, Jun\}$. The event that he was born both in the winter and in the first half of the year is $\{Jan, Feb\}$.*

$$\begin{aligned} A \cap B &= \{Dec, Jan, Feb\} \cap \{Jan, \ldots, Jun\} \\ &= \{Jan, Feb\} \end{aligned}$$

Next, what is the event that an earthquake, bot no flood happens?

**Definition 2.11.** *The set* ***difference*** *is defined as $A \setminus B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$. It can be thought of as the event 'A but not B'.*



**Example 2.12.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \ldots, Dec\}$. The event that he was born in summer is $A = \{Jun, Jul, Aug\}$. The event that he was born in August is $B = \{Aug\}$. The event that he was born in Summer but not in August is*

$$A \setminus B = \{Jun, Jul, Aug\} \setminus \{Aug\} = \{Jun, Jul\}$$

## 2.2 Probabilities

Having set a mathematical framework to speak about events, we can move on to probabilities of events.

### 2.2.1 The axioms of probability

In 1933, Andrey Kolmogorov defined the concept 'probability' formally as a function satisfying three axioms. Loosely speaking, an axiom is a property that is taken to be true with no questions asked. Axioms build the foundation of all of mathematics: real numbers, sets, logic — and probability. Axioms were a big topic around 100 years ago.[1] The idea is to find a minimal set of principles that everybody can agree on. Everything else needs to be deduced from these facts.

We need some more notation. We call sets $A_1, A_2, A_3, \ldots$ are *disjoint*, if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. For example, $A_1 = [0, 1)$, $A_2 = [1, 2)$, $A_3 = [2, 3), \ldots$ are disjoint sets.

[1] Considering the long history of mathematics, axioms are a modern concept.

**Definition 2.13.** *A function $\mathbb{P}$ that assigns a number $\mathbb{P}(A)$ to each event $A \subseteq \Omega$ is a **probability distribution** or a **probability measure**, if it satisfies:*

*(i)* $\mathbb{P}(A) \geq 0$ *for every* $A$,

*(ii)* $\mathbb{P}(\Omega) = 1$,

*(iii) If* $A_1, A_2, \ldots$ *are disjoint, then*

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

The axioms are supposed to reflect what we mean by the concept 'probability'. A (i) non-negative number, such that (ii) the set of all possible outcomes has probability 1, and (ii) the probabilities of exclusive events (only one of them can happen) sum up.[2]

There are two common interpretations of probability: frequencies and degrees of belief.

- In the frequency interpretation, we think of repeating the exact same experiment over and over again. Then $\mathbb{P}(A)$ is the frequency of outcomes where the event $A$ happens. For example, if we say that the probability of heads is $1/2$, we mean that if we flip a (fair) coin many times, the proportion of times we get heads will tend to $1/2$.
- In the degree-of-belief interpretation, $\mathbb{P}(A)$ is the observer's strength of belief that $A$ is true. Like in "I'm 100% sure the Dutch national team wins the world cup."

In either case we require Axioms 1–3 to hold.

### 2.2.2 Properties of probabilities

The axioms imply many intuitive properties of the probability. You can verify the following as an exercise:

- $\mathbb{P}(\emptyset) = 0$.
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
- $0 \leq \mathbb{P}(A) \leq 1$.

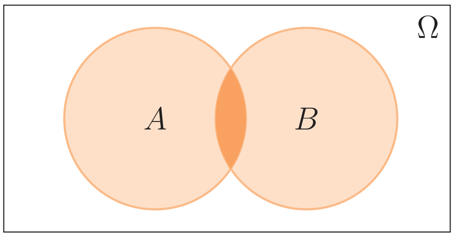
[2] Admittedly, the third axiom is not that intuitive for non-mathematicians.

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

The following result is less trivial:

**Lemma 2.14** (Inclusion-exclusion principle)**.** *For any events $A$ and $B$,*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\,.$$



We shall follow an old mathematical tradition, *proof by picture*.[3] Visually, we can think of the probability as areas (in the sense of 2d-volume). The area enclosed by the rectangle is $\mathbb{P}(\Omega) = 1$. The probability of an event $E$, is the area it covers in proportion to the area of the rectangle. For example, $\mathbb{P}(A)$ is the area enclosed by the circle around $A$. We want to compute the area of the shaded region in the figure above. Add $\mathbb{P}(A)$ (the area of the circle around $A$) to $\mathbb{P}(B)$ (the area of the circle around $B$). Then we counted the area dark region in the middle ($A \cap B$) twice. Hence, we need to subtract $\mathbb{P}(A \cap B)$.

Let's do this formally. If you have trouble understanding the proof, try to visualize every step.

*Proof of Lemma 2.14.* Note that $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. The sets $(A \setminus B)$, $(A \cap B)$, and $(B \setminus A)$ are disjoint. Hence,

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}\Big((A \setminus B) \cup (A \cap B) \cup (B \setminus A)\Big) \\ \text{(third axiom)} \quad &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) + \underbrace{\mathbb{P}(A \cap B) - \mathbb{P}(A \cap B)}_{=0}. \end{aligned}$$

We now use the third axiom again, but in the other direction. The sets $(A \setminus B)$ and $(A \cap B)$ are disjoint and $(A \setminus B) \cup (A \cap B) = A$. Hence,

$$\mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) = \mathbb{P}(A),$$

and similarly,

$$\mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B) = \mathbb{P}(B).$$

[3] That's actually a classical mathematician's joke. Mathematicians are funny.

Using this in the equation above yields

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \qquad \square\end{aligned}$$

Let's end this section with a few examples.

**Example 2.15.** *Toss a fair coin twice. Let $H_1$ be the event that head occurs on toss $1$ and let $H_2$ be the event that head occurs on toss $2$, then*

$$\begin{aligned}\mathbb{P}\left(H_1 \cup H_2\right) &= \mathbb{P}(H_1) + \mathbb{P}(H_2) - \mathbb{P}\left(H_1 \cap H_2\right) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.\end{aligned}$$

In the example, we used a hidden assumption to see that $\mathbb{P}\left(H_1 \cap H_2\right) = 1/4$. More on that later.

Now suppose $\Omega = \{\omega_1, \dots, \omega_n\}$. Such sample spaces are called *finite*, because they contain a finite number of possible outcomes.

**Example 2.16.** *If we throw a die twice, then $\Omega$ has $36$ elements:*

$$\Omega = \{(i, j) : i, j \in \{1, 2, \dots, 6\}\}.$$

*If each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$. For instance, the probability that the sum of the dice is $11$ is equal to $2/36$.*

For a finite set $A$, let $|A|$ denote the number of elements in $A$. If $\Omega$ is finite and each outcome is equally likely, then

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

This is called the *uniform distribution* on the set $\Omega$.

**Example 2.17.** *We ask a random person in the street what his month of birth is. The sample space is $\Omega = \{Jan, Feb, \dots, Dec\}$. The probability that he was born in the winter is $3/12 = 1/4$, if being born in each month is equally probable.*

## 2.3 Independence

Colloquially, we speak of two independent events, when they have nothing to do with each other. For example, the events 'it will be raining tomorrow' and 'you read an article on mice yesterday' are entirely unrelated. There is a formal, probabilistic definition of such events.

**Definition 2.18.** *Events $A$ and $B$ are called* ***independent,*** *if*

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

*We shall write $A \perp B$ in this case.*

To understand this definition, reconsider Example 2.15. Suppose we flip a fair coin twice. From the many coins we have flipped in our lives, we think of the two flips as independent. The outcome of the first flip tells us nothing about the outcome of the second. Now what's the probability that both flips turn heads? Our intuition says 1/4 (1/2 for the first flip times 1/2 for the second). This is exactly what Definition 2.18 says. If you are not convinced, we will see another, equivalent definition of independence in the next section.

Independence does not only apply to a pair of events.

**Definition 2.19.** *A collection of events $\{A_i : i \in I\}$ is called independent, if*

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

*for any finite subset $J \subseteq I$.*

**Example 2.20.** *Toss a fair coin* 10 *times. Let $A =$ 'at least one head' and $T_j$ be the event that tail occurs on the $j^{th}$ coin toss. Then*

$$\begin{aligned} \mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\ &= 1 - \mathbb{P}(\textit{all tails}) \\ &= 1 - \mathbb{P}\left(\bigcap_{i=1}^{10} T_i\right) \\ \textit{(using independence)} \qquad &= 1 - \prod_{i=1}^{10} \mathbb{P}(T_i) \\ &= 1 - \left(\frac{1}{2}\right)^{10} \approx 0.999. \end{aligned}$$

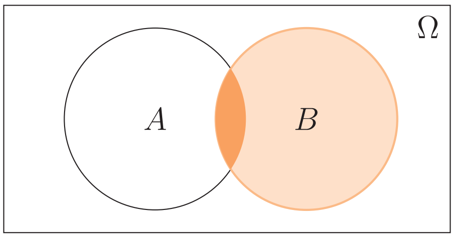
## 2.4 Conditional probability

Conditional probabilities are about if–then statements: if we know that $B$ happened, what is the probability of an event $A$? As in: if we found someones fingerprints one a dead body, what's the probability we caught the murder? We call this $\mathbb{P}(A \mid B)$, the conditional probability of $A$ given $B$.

**Definition 2.21.** *If $\mathbb{P}(B) > 0$, then the* ***conditional probability*** *of $A$ given $B$ is defined by*

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

The restriction $\mathbb{P}(B) > 0$ is necessary for the fraction to be well defined. This aligns with common sense: if something cannot possibly happen, it's foolish to talk about it's consequences. Let's get a visual intuition for the formula.



We see the conventional probability $\mathbb{P}(A)$ as the area of the disk around $A$ divided by the area of the rectangle ($\Omega$). Now suppose we know that the realization $\omega$ lies in the disk $B$ and forget about all the other cases. Then $B$ acts as our new $\Omega$. Now we see the conditional probability $\mathbb{P}(A \mid B)$ is the area covered by $A$ (which, after forgetting everything else, is $A \cap B$) relative to the area of the disk around $B$. We therefore think $\mathbb{P}(A \mid B)$ as the fraction of times $A$ occurs among those in which $B$ occurs.

For fixed $B$, $\mathbb{P}(A \mid B)$ is a proper probability measure – it satisfies all three axioms:

$$\begin{aligned} \mathbb{P}(A \mid B) &\geq 0, \quad \text{for any } A \subseteq \Omega, \\ \mathbb{P}(\Omega \mid B) &= 1, \\ \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \;\middle|\; B\right) &= \sum_{i=1}^{\infty} \mathbb{P}\left(A_i \mid B\right), \quad \text{for disjoint } A_1 A_2, \dots. \end{aligned}$$

It does not behave like that as a function of $B$ though. In general it is not true that

$$\mathbb{P}(A \mid B \cup C) = \mathbb{P}(A \mid B) + \mathbb{P}(A \mid C).$$

Neither is $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)$ true in general (related to the 'prosecutor's fallacy'[4]).

Our intuition often fails us when conditional probabilities are involved. There's a huge number of 'fallacies' and 'paradoxes'. That's why understanding the concept of conditional probabilities is even more important.

**Example 2.22.** *A medical test for COVID-19 has outcomes $+$ and $-$. The joint probabilities are*

[4] https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy

| | *COVID* | *healthy* |
|---|---|---|
| + | *1%* | *1%* |
| − | *0.2%* | *97.6%* |

*The test appears to be pretty accurate:*[5]

$$\begin{aligned}\mathbb{P}(+ \mid \mathit{COVID}) &= \frac{\mathbb{P}(+ \cap \mathit{COVID})}{\mathbb{P}(\mathit{COVID})} = \frac{1\%}{1\% + 0.2\%} \approx 83\%,\\ \mathbb{P}(- \mid \mathit{healthy}) &= \frac{\mathbb{P}(- \cap \mathit{healthy})}{\mathbb{P}(\mathit{healthy})} = \frac{97.6\%}{97.6\% + 1\%} \approx 99\%.\end{aligned}$$

*The first equation states that, if someone has the disease, the test detects the disease in 83% of the cases. The second equation states that, if someone does not has the disease, the test will correctly diagnose 'no disease' in 99% of the cases. These conditional probabilities are the two common quality measures for medical tests (called* sensitivity *and* specificity*). If you read 'the test is 99% correct' in a newspaper, it refers to one of these conditional probabilities (or both). The above accuracy numbers are in line with what we know about the COVID tests in use to day.*

*Now suppose you go for a test and the test is positive. What is the probability you have the disease?*

$$\mathbb{P}(\mathit{COVID} \mid +) = \frac{\mathbb{P}(\mathit{COVID} \cap +)}{\mathbb{P}(+)} = \frac{1\%}{1\% + 1\%} = 50\%.$$

*So if you get a positive test, you have a 50% chance to be healthy anyway. That's not intuitive at all, the test is correct at least 83% of the time!*

*Indeed it is correct on 83% of diseased patients and correct on 99% of healthy patients. However, there are way more healthy people (*$P(\mathit{healthy}) = 98\%$*) than diseased ones (*$P(\mathit{COVID}) = 2\%$*). Out of the huge number of healthy people,* $1\%$ *are incorrectly diagnosed with the disease. If the whole population would get tested,* $98\% \times 1\% \approx 1\%$ *of the entire population would be tested positive despite being healthy. Contrast this with the total number of diseased people:* $1\%$ *of the population. Hence, most of the people with a positive test are healthy.*

The previous example is not just a brain teaser but important for medical practice. False positives often have severe consequences (think: quarantine, mental health issues, expensive medicine, unnecessary surgery). Testing a large proportion of a population can therefore cause more harm than good if a) the test is not reliable enough, or b) the disease is too rare. Especially b) is rarely part of the public discourse, but at least as important as a).

Let's continue with some useful results on conditional probabilities.

[5] Here we use that for two events $A, B$, it holds $P(A) = P(A \cap B) + P(A \cap B^c)$. This follows from the third axiom because $A$ can be partitioned in the two disjoint sets $A \cap B$ and $A \cap B^c$.

**Lemma 2.23.**

*(i) If $A$ and $B$ are independent, then*

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \textit{and} \quad \mathbb{P}(B \mid A) = \mathbb{P}(B).$$

*(ii) For any pair of events $A$ and $B$,*

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

*Proof.*

(i) Recall that independence of $A$ and $B$ is equivalent to $\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B)$. By the definition of conditional probabilities (Definition 2.21), it holds

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

Repeat the same argument by interchanging the roles of $A$ and $B$.

(ii) For the first equality, multiply both sides in Definition 2.21 with $\mathbb{P}(B)$. Repeat the same argument by interchanging the roles of $A$ and $B$. □

The first part has an intuitive interpretation. Independence of $A$ and $B$ means that the events are completely unrelated. Knowing about the outcome of $B$ cannot provide any additional information on the event $A$. Therefore, the conditional probability $\mathbb{P}(A \mid B)$ must be the same as the unconditional probability $\mathbb{P}(A)$. The second part is just a useful trick in computations.

**Example 2.24.** *Draw two cards from a deck (of 52 cards), without replacement. Let $A$ be the event that the first draw is the Ace of Clubs and let $B$ be the event that the second draw is the Queen of Diamonds. Then*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \frac{1}{52} \times \frac{1}{51}.$$

*Note that $A$ and $B$ are not independent, because if $A$ happens, the second draw cannot be the Ace of Clubs (this card was removed from the deck).*

There are two more useful formulas related to conditional probabilities. The first, *Bayes' theorem*[6], is a direct consequence of Lemma 2.23.

**Theorem 2.25** (Bayes' theorem)**.** *Let $A, B$ be events with $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then*

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

[6] Named after Reverend Thomas Bayes, who invented the concept of conditional probability in 1763. (Yes, that long ago!)

*Proof.* We have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \overset{[Lemma\ 2.23]}{=} \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

□

This also tells us that $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)$ *only if* $\mathbb{P}(A) = \mathbb{P}(B)$. Otherwise, the two conditional probabilities are only proportional up to factor determined by the relative probabilities.

The last result relates to partitions of the sample space. A *partition* of $\Omega$ is a (finite or infinite) sequence of disjoint events $A_i$, such that $\bigcup_i A_i = \Omega$. The next result shows that any unconditional probability can be computed from a sum of weighted conditional probabilities.

**Theorem 2.26** (Law of total probability)**.** *Let $A_1, \dots, A_k$ be a partition of $\Omega$. Then for any event $B$,*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

*Proof.* Because $A_1, \dots, A_k$ is a partition of $\Omega$, it holds

$$B = B \cap \Omega = B \cap \big(\bigcup_{i=1}^{k} A_i\big) = \bigcup_{i=1}^{k}(B \cap A_i).$$

Furthermore, because the events $A_1, \dots, A_k$ are disjoint, also the events $(B \cap A_1), \dots, (B \cap A_k)$ must be disjoint. The third axiom then yields

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}\Big(\bigcup_{i=1}^{k}(B \cap A_i)\Big) \\ &= \sum_{i=1}^{k} \mathbb{P}(B \cap A_i) \\ [Lemma\ 2.23] \quad &= \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i). \end{aligned}$$

□

Let's see these two results in action:

**Example 2.27.** *Suppose I divide my email into three categories:*

- $A_1 =$ *'spam',*
- $A_2 =$ *'low priority'*
- $A_3 =$ *'high priority'.*

*From previous experience I know that*

$$\mathbb{P}(A_1) = 0.7, \qquad \mathbb{P}(A_2) = 0.2, \qquad \mathbb{P}(A_3) = 0.1.$$

*Let $B$ be the event that the email contains the word 'free'. From previous experience,*

$$\mathbb{P}(B \mid A_1) = 0.9, \qquad \mathbb{P}(B \mid A_2) = 0.01, \qquad \mathbb{P}(B \mid A_3) = 0.01.$$

*Now suppose I receive a new email containing the word free. What is the probability, that it is spam? Bayes' theorem and the law of total probability yield*

$$\begin{aligned}\mathbb{P}(A_1 \mid B) &= \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\sum_{i=1}^{3} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)} \\ &= \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} \\ &= 0.995.\end{aligned}$$

## 2.5 Random variables

Recall the basic problem of statistics:

> Having observed some data $X_1, \dots, X_n$, what can we say about the mechanism that generated them?

This says nothing about sample spaces and events, so what was all this fuzz about? The key is the concept of a *random variable.*

**Definition 2.28** (Random variable)**.** *A random variable is a mapping*

$$X : \Omega \to \mathbb{R}$$

*that assigns a real number $X(\omega)$ to each $\omega \in \Omega$.*

**Example 2.29.** *Flip a coin 5 times. Let $X(\omega)$ be the number of heads in the sequence $\omega$. For example, if $\omega = HHTTH$, then $X(\omega) = 3$.*

**Example 2.30.** *Let $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ be the unit disk. Consider drawing a point at random from $\Omega$. A typical outcome is of the form $\omega = (x, y)$. Some examples of random variables are $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$.*

Random variables provide the link between sample spaces and events and the data. In general, a random variable is any quantity whose actual value is random, i.e., dependent on a realization $\omega \in \Omega$. We then view the data $X_1, \dots, X_n$ as $n$ realizations of a random variable $X(\omega)$.

In fact, we rarely mention the sample space $\Omega$ in statistics and work with random variables directly. For example, we define $X = X(\omega)$ to be the luminosity of a randomly chosen star in a galaxy. There are several choices we could make for the underlying sample space ($\omega$ is the ID of a star, or $\omega$ is the luminosity itself, or ...). If we're just interested in the luminosity, the exact choice of the sample space becomes irrelevant. For working with probabilities, we just need to know that there is *some* sample space underlying the experiment.

## 2.6 Distribution functions

Random variables take on real ($\mathbb{R}$) values by definition. That makes our life a lot easier, because we can characterize probability distributions by more traditional mathematical functions taking real numbers as arguments (as opposed to sets as in Definition 2.13).

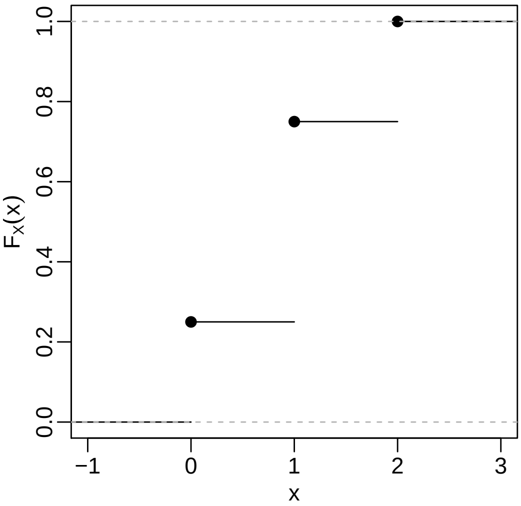
**Definition 2.31.** *The* ***cumulative distribution function*** *(CDF) of the random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$ defined by*

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}).$$

**Example 2.32.** *Suppose we flip a fair coin (so $\mathbb{P}(H) = \mathbb{P}(T) = 1/2$) twice and let $X$ be the number of heads. Then $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ and $\mathbb{P}(X = 1) = 1/2$. The corresponding CDF is*

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2. \end{cases}$$

*The graph of this function is shown below:*



*It is a step function that jumps at 0, 1, 2, and 3 (the possible outcomes). The size of the jump at $x$ is equal to the probability $\mathbb{P}(X = x)$. The dots indicate that at any jump point $x$, $F(x)$ is equal to the value* after *the jump.*

CDFs have a few useful properties. While Definition 2.31 gives a *probabilistic* definition of CDFs, such functions can also be characterized *analytically.*

**Theorem 2.33.** *A function $F : \mathbb{R} \to [0, 1]$ is a CDF for some probability $\mathbb{P}$* ***if and only if*** *$F$ satisfies the following conditions:*

*(i)* *$F$ is non-decreasing: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$,*

*(ii)* $\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1,$

*(iii)* *$F$ is right-continuous: $F(x) = F(x^+)$, where $F(x^+) = \lim_{y\downarrow x} F(y)$.*

The proof isn't hard but quite boring. Let's turn to more interesting things.

## 2.7 Probability mass and density functions

CDFs are a bit hard to interpret, since we're accumulating probabilities up to a value $x$. It would be much easier if we would just know how probable a specific value of $x$ is. Probability mass and density functions do just that. To define them, we need to discern discrete and continuous random variables.

**Definition 2.34.** *A random variable $X$ is called* ***discrete****, if it can take only countably*[7]*many different values, say $\{x_1, x_2, \ldots\}$.*

[7] A set is countable if we can assign each of its elements with a natural number. Examples are finite sets, e.g., $\{0, 1\}$, or countably infinite sets like $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{Q}$.

**Definition 2.35.** *We define the* ***probability mass function*** *(PMF) of a discrete variable $X$ by*

$$f_X(x) = \mathbb{P}(X = x).$$

The following properties follow directly from the definition of probabilities and the CDF:

- For all $x$ we have
$$f_X(x) \geq 0, \qquad \sum_i f_X(x_i) = 1.$$
- The CDF is related to $f_X$ by
$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i\colon x_i \leq x} f_X(x_i).$$

In fact, we implicitly derived the CDF in Example 2.32 from the PMF $f_X$. Here's another common example:

**Example 2.36** (Bernoulli distribution)**.** *Consider a coin flipping experiment and define $X$ by $X(H) = 1$, $X(T) = 0$. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. This is called the* Bernoulli$(p)$ *distribution and has PMF*

$$f(x) = \begin{cases} 1 - p & \textit{for } x = 0 \\ p & \textit{for } x = 1 \\ 0 & \textit{otherwise}, \end{cases}$$

*The CDF is*

$$F(x) = \begin{cases} 0 & \textit{for } x < 0 \\ 1 - p & \textit{for } 0 \leq x < 1 \\ 1 & \textit{for } x \geq 1. \end{cases}$$

The Bernoulli distribution is super common in statistics because it applies to any random variable that only has two possible outcomes (win/lose, pass/fail, cat/dog, ...). Later in this course, we'll see more examples of common discrete distributions.

**Definition 2.37.** *A random variable $X$ is called* ***continuous****, if the cdf $F_X(x)$ is continuous.*

Maybe it's not obvious, but this definition already implies that $X$ can take uncountably many values. (Otherwise the CDF *must* jump somewhere.) In this case, the concept of a PMF is meaningless, because $\mathbb{P}(X = x) = 0$ for all $x$.[8] Instead, we use a slightly different concept, a *density* function.

[8] The sum of uncountably many strictly positive numbers is always infinite. Hence, $\mathbb{P}(X = x) > 0$ is only possible for countably many $x$.

**Definition 2.38.** *If $F_X(x)$ is differentiable, the function $f_X(x) = F'_X(x)$ is called the* ***probability density function*** *(PDF).*

From this definition it follows that

- $f_X(x) \geq 0$,
- $\int_{-\infty}^{\infty} f_X(t)dt = 1$,
- $F_X(x) = \displaystyle\int_{-\infty}^{x} f_X(t)dt$.

The density $f_X(x)$ is *not* a probability. Instead it is interpreted as a relative likelihood. Regions where $f_X$ is large are more likely; regions where $f_X$ is small are less likely. (You've probably seen the concept a density as mass per unit volume. The idea here is similar.)

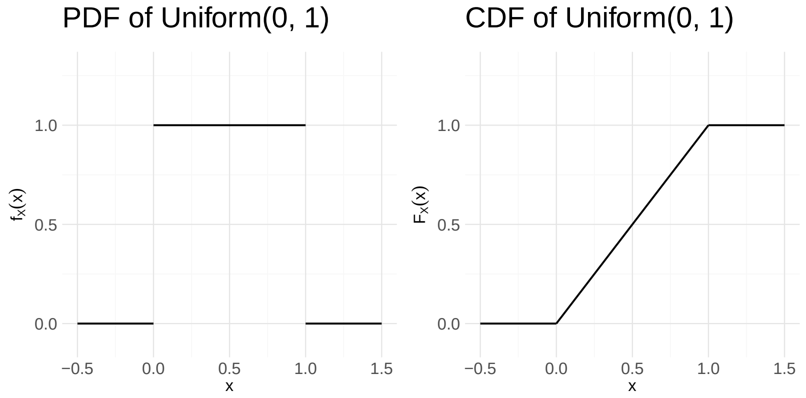
**Example 2.39** (Uniform distribution)**.** *Suppose that $a < b$ and $X$ has PDF*

$$f_X(x) = \begin{cases} 1/(b-a) & \text{for } a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

*Clearly, $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(x)dx = 1$. A random variable with such a PDF is said to have a* Uniform$(a, b)$ *distribution. The CDF is given by*

$$F_X(x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leq x \leq b \\ 1 & x > b. \end{cases}$$

*The PDF and CDF of the* Uniform$(0, 1)$ *distribution are shown below:*



*The graph of the PDF indicates that all values in the interval* $(0, 1)$ *are equally likely, which explains the distribution's name.*

Note that, unlike a PMF, a PDF *can* be larger than 1 (and even unbounded!).

**Example 2.40.** *Let* $f(x) = (2/3)x^{-1/3}$ *for* $0 < x < 1$ *and* $f(x) = 0$ *otherwise. Then* $f(x) \geq 0$ *and* $\int_{-\infty}^{\infty} f(x)dx = 1$, *but* $f$ *is unbounded.*

This reminds us to not misinterpret a PDF as a probability, but a relative likelihood.

Now that we know about continuous random variables, we can state a few more properties of the CDF.

**Lemma 2.41.** *Let* $F$ *be the CDF of a random variable* $X$. *Then*

*(i)* $\mathbb{P}(X = x) = F(x) - F(x^-)$, *where* $F(x^-) = \lim_{y\uparrow x} F(y)$;

*(ii)* $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;

*(iii)* $\mathbb{P}(X > x) = 1 - F(x)$;

*(iv) If* $X$ *is continuous, then*

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) \\ &= \mathbb{P}(a < X \leq b) \\ &= \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a \leq X \leq b). \end{aligned}$$

## 2.8 Bivariate distributions

Similar concepts apply to the *joint distribution* of multiple random variables. Joint distributions come into play when we are interested in events that depend on several variables. For example, the probability that tomorrow there is no rain ($X$) and the temperature ($Y$) is more than 20 degrees. For simplicity, we shall only consider joint distributions of two random variables here, but the concept naturally extends to higher dimensions.

**Definition 2.42.** *We define the* ***joint CDF*** *as*

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x \text{ and } Y \leq y) = \mathbb{P}(X \leq x, Y \leq y).$$

We must again differentiate discrete and continuous random variables.

### 2.8.1 Discrete variables

**Definition 2.43.** *Let* $X$ *and* $Y$ *be discrete random variables. The* ***joint PMF*** *is*

$$f(x, y) = \mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x, Y = y).$$

**Example 2.44.** *Here is a bivariate distribution for random variables $X$ and $Y$ both taking values $0$ or $1$:*

| | $Y=0$ | $Y=1$ | |
|---|---|---|---|
| $X=0$ | $1/9$ | $2/9$ | $1/3$ |
| $X=1$ | $2/9$ | $4/9$ | $2/3$ |
| | $1/3$ | $2/3$ | $1$ |

*For instance, $f_{X,Y}(1,1) = \mathbb{P}(X=1, Y=1) = 4/9$.*

From a joint distribution, we can also extract the distributions of the individual variables. The latter are called *marginal distributions*.

**Definition 2.45.** *If $(X,Y)$ has the joint mass function $f_{X,Y}$, the **marginal mass function** for $X$ is*

$$f_X(x) = \mathbb{P}(X=x) = \sum_y \mathbb{P}(X=x, Y=y) = \sum_y f(x,y).$$

*Likewise, the marginal mass function for $Y$ is*

$$f_Y(y) = \mathbb{P}(Y=y) = \sum_x \mathbb{P}(X=x, Y=y) = \sum_x f(x,y).$$

As you can see, we extract the marginal PDFs by summing the joint PMF over all possible values of the other variable. For example, in Example 2.44 we have $f_X(0) = 1/3$ and $f_Y(1) = 2/3$.

### 2.8.2 Continuous variables

Similarly, we define a joint PDF for continuous random variables.

**Definition 2.46.** *Let $X$ and $Y$ be continuous random variables. A function $f(x,y)$ is called the **joint PDF** of $(X,Y)$, if*

*(i)* $f(x,y) \geq 0$ *for all* $(x,y)$;

*(ii)* $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1$;

*(iii) for any set* $A \subset \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}\big((X,Y) \in A\big) = \int \int_A f(x,y) dx dy.$$

(Compare this to the properties stated after Definition 2.38.)

**Example 2.47.** *Let*

$$f(x,y) = \begin{cases} 1 & \textit{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \textit{otherwise.} \end{cases}$$

*This is the PDF of the uniform distribution on the unit square. Suppose we want to compute*

$$\mathbb{P}(X \leq 1/2, Y \leq 1/2) = F_{X,Y}(1/2, 1/2) = \mathbb{P}(X < 1/2, Y < 1/2).$$

*Integration gives that* $F_{X,Y}(1/2, 1/2) = 1/4$.

Marginal densities are constructed like in the discrete case, just replacing the sum by an integral.

**Definition 2.48.** *For continuous random variables* $(X, Y)$ *with a joint PDF* $f(x, y)$ *the* ***marginal PDFs*** *are*

$$f_X(x) = \int f(x,y)dy, \quad f_Y(y) = \int f(x,y)dx.$$

*The corresponding marginal CDFs are denoted by* $F_X$ *and* $F_Y$*.*

**Example 2.49.** *Let*

$$f_{X,Y}(x,y) = \begin{cases} e^{-(x+y)} & \textit{for } x \geq 0 \textit{ and } y \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

*Integration gives that* $f_X(x) = e^{-x}$.

**Example 2.50.** *Let*

$$f_{X,Y}(x,y) = \begin{cases} x + y & \textit{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \textit{otherwise.} \end{cases}$$

*Integration gives that* $f_Y(y) = 1/2 + y$ *for* $0 \leq y \leq 1$ *and* $f_Y(y) = 0$ *otherwise.*

## 2.9 Conditional distributions

We can also apply the concept of conditional probability to random variables and associated functions. The **conditional PMF** of a discrete random variable $X$, given that a discrete random variable $Y$ takes the value $y$, is

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

provided $f_Y(y) > 0$. The **conditional PDF** of a continuous random variable $X$, given that a continuous random variable takes the value $y$ is

$$f_{X|Y}(x \mid y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)}, & \text{for } f_Y(y) > 0 \\ 0, & \text{for } f_Y(y) = 0. \end{cases}$$

Then

$$\mathbb{P}(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y) dx,$$

so $f_{X|Y}(\cdot \mid y)$ is a proper density function for the conditional probability.

**Example 2.51.** *Let*

$$f_{X,Y}(x,y) = \begin{cases} x + y & \textit{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \textit{otherwise.} \end{cases}$$

*Since* $f_Y(y) = y + 1/2$ *for* $0 \le y \le 1$,

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}$$

*for* $0 \le x \le 1, 0 \le y \le 1$ *and* $f_{X|Y}(x|y) = 0$ *otherwise. Thus,*

$$\mathbb{P}\left(X < \frac{1}{4} \mid Y = \frac{1}{3}\right) = \int_0^{1/4} f_{X|Y}\left(x \mid \frac{1}{3}\right) dx = \frac{11}{80}.$$

## 2.10 Independence (ctd')

We call two random variables independent, if all events associated with the two variables are independent.

**Definition 2.52.** *Two random variables $X$ and $Y$ are* ***independent*** *(denoted $X \perp Y$), if for every $A$ and $B$*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

The interpretation is the same as for events: two variables are independent, if their outcomes are entirely unrelated (do not influence each other).

Independence can also be characterized using densities (this follows immediately from their definitions).

**Theorem 2.53.** *Let $X$ and $Y$ have joint PDF (or PMF) $f_{X,Y}$. Then $X \perp Y$ if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $(x,y)$.*

**Example 2.54.** *Let $X$ and $Y$ have the joint distribution as in the following table:*

| | $Y=0$ | $Y=1$ | |
|---|---|---|---|
| $X=0$ | $1/4$ | $1/4$ | $1/2$ |
| $X=1$ | $1/4$ | $1/4$ | $1/2$ |
| | $1/2$ | $1/2$ | $1$ |

*Then $X$ and $Y$ are independent, which can be verified by the previous theorem. For example, $f(0,0) = 1/4 = f_X(0)f_Y(0)$, and similarly for other cases.*

**Example 2.55.** *Let $X$ and $Y$ be independent and both have the same PDF*

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

*Suppose we want to find $\mathbb{P}(X+Y \le 1)$. Thanks to independence,*

$$f(x,y) = f_X(x)f_Y(y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

*So*

$$\mathbb{P}(X+Y \le 1) = \int\int_{x+y\le 1} f(x,y)dxdy = \frac{1}{6}.$$

Let's back up for a second and think about the bigger picture. Our goal in this course is to learn from data. These data are modeled as random variables $X_1, \dots, X_n$. In many circumstances it is reasonable to assume these random variables are independent. For example, if $X_j$ is the $j$th flip of a coin, we have no reason to assume that the outcome of one flip is affecting another. The situation is similar in other repeated experiments or some measurements taken on distinct objects (e.g., stars, galaxies, ...).

We need to be a little bit more precise here. Independence of $X_1, dots, X_n$ is more than just *pairwise* independence of all $X_i, X_j$:

**Definition 2.56.** *$X_1, \dots, X_n$ are **independent**, if for every $A_1, \dots, A_n$*

$$\mathbb{P}(X_1 \in A_1, \dots X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

This definition is stronger than pairwise independence. The interpretation is as follows: the joint outcome of $(X_{i_1}, \dots, X_{i_k})$ is entirely unrelated to the joint outcome of $(X_{j_1}, \dots, X_{j_\ell})$, whenever no variable appears twice, i.e., $\{i_1, \dots, i_k\} \cap \{j_1, \dots, j_\ell\} = \emptyset$.

There is another assumption commonly made in statistics: that each of the variables $X_1, \dots, X_n$ has the same distribution, i.e. $F_{X_1} = \dots, F_{X_n}$.

**Definition 2.57.** *If $X_1 \sim F, \dots, X_n \sim F$ are independent, we say that $X_1, \dots, X_n$ are **iid** (independent and identically distributed).*

If $X_1, \dots, X_n$ are *iid*, we also write $X_1, \dots, X_n \sim F$. If $F$ has density $f$ we also write $X_1, \dots, X_n \sim f$.

## 2.11 Transforms

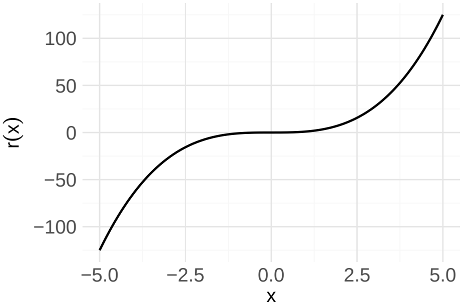
Sometimes we know the distribution of random variable $X$, but are interested in the distribution of a transformation $r(X)$. The following result comes in handy.

**Theorem 2.58.** *Suppose $X$ is a continuous random variable, $r$ is a invertible[9]and differentiable[10]function and $Y = r(X)$. Then*

$$f_Y(y) = \frac{f_X(r^{-1}(y))}{|r'(r^{-1}(y))|}.$$

Similar results exist for discrete variables and joint densities. You can look them up yourself[11] whenever there's a need. The above result will be good enough for this course.

**Example 2.59.** *Let $Y = X^3$. The function $r(x) = x^3$ is strictly increasing and, hence, invertible.*



*Now $r^{-1}(y) = y^{1/3}$ and $\frac{dr(y)}{dy} = 3y^2$. So*

$$f_Y(y) = \frac{f_X(y^{1/3})}{|3(y^{1/3})^2|} = \frac{f_X(y^{1/3})}{3\,|y^{2/3}|}.$$

## 2.12 Expectation

Distributions are functions and, thus, fairly complex objects. Instead we can also look at *summaries* of the distribution. The most important summary is the *expect value* $\mathbb{E}[X]$. It tells us what value a random variable $X$ we can expect to see *on average*. For the formal definition, we once again need to discern discrete and continuous cases.

[10] A function is invertible if and only if it is strictly monotone (increasing or decreasing).

[10] Whenever you can compute the derivative, it is differentiable. This will rarely be an issue.

[11] https://en.wikibooks.org/wiki/Probability/Transformation_of_Probability_Densities

**Definition 2.60.** *The* ***expected value*** *(also* mean *or* first moment*) of a random variable $X$ is defined as*

$$\mathbb{E}[X] = \int_\Omega x dF(x) = \begin{cases} \displaystyle\sum_{x \in \Omega} x f(x) & \text{if } X \text{ is discrete} \\ \displaystyle\int_\Omega x f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

The integral $\int_\Omega x dF(x)$ has a precise measure theoretic meaning that you don't need to worry about. You may just treat it as short hand notation for one of the two cases on the right.

In both cases, the expected value is an average over all possible outcomes of $X$, weighted by the likelihood of occurrence. Another way to think about it is the following approximation: repeat the same experiment many times and average all outcomes, then $\mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^n X_i$, for a large number of *iid* draws $X_1, \dots, X_n$ with the same distribution as $X$. While this is only an approximation, it helps to understand the meaning of the number $\mathbb{E}[X]$. Let's see some examples.

**Example 2.61.** *Let $X \sim \text{Bernoulli}(p)$. Then $\mathbb{E}[X] = 1 \times p + 0 \times (1 - p) = p$. So if we flip a fair coin many times, count heads as 1 and tails as 0, we expect to see a value of 0.5 on average.*

**Example 2.62.** *Let $X$ denote the outcome of a single throw of a fair die. Then $\mathbb{E}[X] = (1 + 2 + .. + 6) \times 1/6 = 3.5$.*

**Example 2.63.** *Let $X \sim \text{Uniform}(a, b)$. Then*

$$\mathbb{E}[X] = \frac{1}{(b-a)} \int_a^b x dx = \frac{b^2 - a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{a+b}{2}.$$

*Hence, if we draw uniformly from the interval $(1, 2)$ many times, we expect to see a value of 1.5 on average.*

As you can see, computing the expectation of a random variable can be fairly easy. It's similarly easy to compute the expectation of a transformed random variable.

**Theorem 2.64.** *Let $Y = r(X)$. Then*

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x) dF_X(x).$$

**Example 2.65.** *Let $X \sim \text{Bernoulli}(p)$ and $Y = X^2$. Then*

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = 1^2 \times p + 0^2 \times (1 - p) = p.$$

**Example 2.66.** *Let $X \sim \text{Uniform}(0,1)$ and let $Y = r(X) = e^X$. Then*

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int_{\mathbb{R}} e^x f(x)dx = \int_0^1 e^x dx = e - 1.$$

The same also works when more than one random variable is involved, just that we need to use the joint PMF/PDF as a weight in the sum/integral. For example, for two continuous random variables $X_1, X_2$ and $Y = r(X_1, X_2)$,

$$\mathbb{E}[Y] = \mathbb{E}[r(X_1, X_2)] = \int r(x_1, x_2) f_{X_1,X_2}(x_1, x_2) dx_1 dx_2.$$

The expectation has the nice property of being *linear*, which just means that you can pull the sum and any constants out of it. If you think about our approximation of $\mathbb{E}[X]$ as averaging over many draws from an experiments, this makes sense (and it follows immediately from Definition 2.60).

**Theorem 2.67.** *If $X_1, \dots, X_n$ are random variables and $a_0, a_1, \dots, a_n$ are constants, then*

$$\mathbb{E}\left[a_0 + \sum_{i=1}^n a_i X_i\right] = a_0 + \sum_{i=1}^n a_i \mathbb{E}[X_i].$$

*(In particular, $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.)*

**Example 2.68.** *Let $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p)$ and $X = \sum_{i=1}^n Y_i$. Then we say $X$ has $\text{Binomial}(n,p)$-distribution. By linearity of the expectation, it holds*

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[Y_i] = np.$$

While sums of random variables are easy to handle, this is not in general true for products. A convenient and common special case is when variables are independent.

**Theorem 2.69.** *If $X_1, \dots, X_n$ are independent random variables, then*

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

*This might fail without independence assumption!*

*Proof.* Let's just look at the continuous case and $n = 2$. If $X_1, X_2$ are independent, the joint density $f_{X_1,X_2}$ is just the product of marginal densities $f_{X_1} \times f_{X_2}$. Hence, by Theorem 2.64

$$\begin{aligned}
\mathbb{E}[X_1X_2] &= \int x_1x_2f_{X_1,X_2}(x_1,x_2)dx_1dx_2\\
&= \int x_1x_2f_{X_1}(x_1)f_{X_2}(x_2)dx_1dx_2\\
&= \int x_1f_{X_1}(x_1)dx_1 \times \int x_2f_{X_2}(x_2)dx_2\\
&= \mathbb{E}[X_1] \times \mathbb{E}[X_2],
\end{aligned}$$

where we used independence in the second equality. □

## 2.13 Variance and standard deviation

The expectation is important, but alone it gives a very limited view on the distribution. Imagine a lottery that always pays out the expected value: you pay €10 for your ticket and you immediately get €8 back. Nobody would play that game! What makes lotteries exciting (for some people at least) is the variability of the outcome. The most common measures for variability are the *variance* and *standard deviation.*

**Definition 2.70.** *Let $X$ be a random variable with mean $\mathbb{E}[X] = \mu$.*

- *The* ***variance*** *of $X$ is defined as*

$$\mathbb{V}[X] = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 dF(x).$$

- *The* ***standard deviation*** *of $X$ is* $\mathrm{sd}[X] = \sqrt{\mathbb{V}[X]}$.

Both variance and standard deviation are one-number summaries of the distribution. They answer the question: "how much does $X$ fluctuate around its mean"? In the extreme case $\mathbb{V}[X] = 0$, there is no variability at all and $X$ is just a constant. While the variance is easier to calculate with, the standard deviation is easier to interpret. Because we square what's in the expectation, we need to take the square root of the result to bring it back to the original scale/units. Keep that in mind when reporting or reading about these measures.

**Theorem 2.71.** *The variance has the following properties:*

*(i)* $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

*(ii)* *If* $a$ *and* $b$ *are constants, then* $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$.

*(iii)* *If* $X_1, \dots, X_n$ *are independent (!) and* $a_1, \dots, a_n, b$ *are constants, then*

$$\mathbb{V}\left[\sum_{i=1}^n a_i X_i + b\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i].$$

You can verify this an exercise. The key ingredient is linearity of the expectation Theorem 2.67.

**Example 2.72.** *Let* $Y \sim \text{Bernoulli}(p)$. *Recall that* $\mathbb{E}[Y^2] = p$ *and* $\mathbb{E}[Y] = p$. *Therefore* $\mathbb{V}[Y] = p - p^2 = p(1-p)$.

**Example 2.73.** *Suppose* $X \sim \text{Binomial}(n, p)$. *Recall that* $X = \sum_{i=1}^n Y_i$ *for* $Y_1, \dots, Y_n$ iid $\text{Bernoulli}(p)$ *random variables. Therefore*

$$\mathbb{V}[X] = \sum_{i=1}^n \mathbb{V}[Y_i] = np(1-p).$$

## 2.14 Covariance and correlation

Now we know about the two most important summaries of the distribution of a single random variable. When there's another variable, a new concept comes into play: the dependence between variables. The only thing we know about dependence so far is the concept of independence: two variables are completely unrelated. It's not hard to imagine a situation where variables *are related*. For example, a person's body weight tends to be related to her height. Again, the complete picture of the relationship is captured by the joint distribution (or PMF/PDF), but there are one-number summaries of the dependence.

**Definition 2.74.** *Let* $X$ *and* $Y$ *be random variables with means* $\mu_X$ *and* $\mu_Y$ *and standard deviations* $\sigma_X$ *and* $\sigma_Y$*, respectively.*

- *The* ***covariance*** *between* $X$ *and* $Y$ *is defined as*

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right].$$

- *The* ***correlation*** *between* $X$ *and* $Y$ *is defined as*

$$\rho_{X,Y} = \rho(X, Y) = \frac{\mathbb{C}\text{ov}[X, Y]}{\sigma_X \sigma_Y}.$$

Similar to the variance, the covariance is a bit harder to interpret (just think about its units), mainly because it mixes two things: i) the individual variability of $X$ and $Y$, ii) the dependence between $X$ and $Y$. The correlation is a standardized version that takes out the variability part. That makes it a pure measure of dependence, which is typically what we want.

**Theorem 2.75.** *(i) The covariance satisfies* $\mathbb{C}\mathrm{ov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

*(ii) The correlation satisfies* $\rho(X,Y) \in [-1,1]$.

*(iii) If* $Y = aX + b$ *for some constants* $a$ *and* $b$*, then*

$$\rho(X,Y) = \begin{cases} 1, & \text{if } a > 0 \\ -1, & \text{if } a < 0. \end{cases}.$$

*(iv) If* $X$ *and* $Y$ *are independent, then* $\rho(X,Y) = 0$.
*(The converse is in general false.)*

The first property is useful mainly for calculations. The second tells us that the correlation is standardized to the interval $[-1,1]$. The sign and magnitude of the correlation tell us what kind of dependence we are dealing with. Let's consider the three extreme cases. As shown in (iii), the correlation has absolute magnitude 1 when two variables are perfectly linearly related. The sign tells us how:

- $\rho(X,Y) > 0$: $X$ and $Y$ tend to be both large or both small at the same time.
- $\rho(X,Y) < 0$: large values of $X$ tend to occur with small values of $Y$ and vice versa.

Finally, if $\rho(X,Y) = 0$, we say that $Y$ and $X$ are *uncorrelated*. This is always the case when $X$ and $Y$ are independent.

Keep in mind that correlation is only a measure of *linear dependence*. There are other forms of dependence where $Y$ and $X$ are perfectly related (e.g., $Y = X^2$), but the correlation is not 1. Similarly, there are cases where $\rho(X,Y) = 0$, but the variables are not independent.

Now that we know about covariances, we can drop the independence assumption in Theorem 2.71 (iii):

**Theorem 2.76.** *For random variables* $X_1, \dots, X_n$ *and constants* $a_1, \dots, a_n$,

$$\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \mathbb{C}\mathrm{ov}[X_i, X_j].$$

We see that, in general, we need an extra correction term for the variance of a sum. To understand why, consider the case where $X_1 = -X_2$. Then clearly $X_1 + X_2 = 0$, so there is no variability at all. If we would just sum up the variances of $X_1$ and $X_2$, we would get $2\mathbb{V}[X_1] \neq 0$. The covariance term in the theorem above fixes this.

### 2.14.1 Conditional expectation

The final concept that we need from probability theory is the *conditional expectation*. As you might expect, it is the expectation related to a conditional distribution. Knowing this, its mathematical definition is just what you expect:[12]

**Definition 2.77.** *The* ***conditional expectation*** *of $X$ given $Y = y$ is*

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum_x x f_{X|Y}(x|y) & \textit{discrete case} \\ \int x f_{X|Y}(x|y) dx & \textit{continuous case.} \end{cases}$$

Recall that the unconditional expectation $\mathbb{E}[X]$ is interpreted as "the value of $X$ we expect to see on average". Now suppose we have already observed some other information $Y = y$. Which value do we expect now on average? That's it, the conditional expectation $\mathbb{E}[X \mid Y = y]$. Let's make this a bit more concrete and go back to the example with body height ($X$) and weight ($Y$). If we draw a random person from the Dutch population, we expect his height to be roughly 1.80m ($\mathbb{E}[X] \approx 1.80$). What if I told you this person weighs $50kg$ ($Y =\approx 50$). How tall do you expect this person to be now ($\mathbb{E}[X \mid Y = y]$)?

The concept should be easy to grasp, the mathematically, the conditional expectation is a bit trickier than the unconditional one. The conditional expectation is different if we change the information we condition on. (If I tell you the person weighs 90kg the answer will be different than above.) So while $\mu = \mathbb{E}[X]$ is a number, $\mu(y) = \mathbb{E}[X \mid Y = y]$ is a function.

Now, what happens if we plug the random variable $Y$ into the function $\mu(y)$? Well we get another random variable defined as $Z = \mu(Y)$. This random variable is typically denoted as $\mathbb{E}[Y \mid X]$ (which, admittedly, is a bit confusing but you'll get used to it). Of course, this variable has a distribution and we can compute its expectation.

**Theorem 2.78** (Law of total expectation/Tower rule)**.** *It holds*

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

The Tower rule is a convenient tool for computations as we'll see in a minute. Let's first verify that the formula makes sense. Suppose we know the exact

[12] You see what I did there?

function $\mu(y)$ stating what body height to expect given every possible value of the body weight $y > 0$. Now we start drawing random people from the population, check their weight $Y$, and compute the expected height $Z = \mu(Y)$. On average, our guesses $Z$ should equal the average height in the population $\mathbb{E}[X]$.

**Example 2.79.** *Suppose* $Y \sim \text{Uniform}(0, 1)$. *After we observe* $Y = y$, *we draw* $X|Y = y \sim \text{Uniform}(y, 1)$. *Note that* $f_{X|Y}(x|y) = 1/(1 - y)$ *and hence*

$$\mathbb{E}[X|Y = y] = \int_y^1 x \frac{1}{1-y} dx = \frac{1+y}{2}.$$

*So* $\mathbb{E}[X|Y] = (1+Y)/2$. *This is a random variable, whose value is* $(1+y)/2$ *after* $Y = y$ *is observed. To compute* $\mathbb{E}[X]$, *we write*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}\left[\frac{1+Y}{2}\right] = \frac{1}{2}(1 + \mathbb{E}[Y]) = \frac{3}{4}.$$

More generally, it also holds

$$\mathbb{E}[\mathbb{E}[r(X, Y)|Y]] = \mathbb{E}[r(X, Y)].$$

and there is a similar result for the variance (”law of total variance”). We won’t need it for this course, but it’s good to know it exists.

# 3 Descriptive statistics and exploratory data analysis

The last chapter introduced the fundamentals of probability theory to set the stage for the main objective of this course: doing statistics (or 'analyzing data'). Recall the basic problem of statistics:

> *Having observed some data $X_1, \dots, X_n$, what can we say about the mechanism that generated them?*

By now we know that the data $X_1, \dots, X_n$ are modeled as random variables. We have learned that they are characterized by a probability distribution. And if this distribution is known, we know how to interpret and summarize it.

But how do we know the distribution of the data? Well, we don't — and that's what distinguishes probability theory from statistics. The data we observe gives us some clues how the distribution may look like and we wish to extract as much information as possible. The first step is to summarize and explore the data. Here the aim is to 'get a feeling' for the data before we to do actual modeling and inference. This process is called *exploratory data analysis (EDA)*.

## 3.1 Sample averages and the law of large numbers

Suppose our data $X_1, \dots, X_n$ are *iid* random variables from an *unknown* distribution $F$. We start with the most basic task: learning about their expectation. Because the data are *iid*, there is a number $\mu \in \mathbb{R}$ such that $\mathbb{E}[X_i] = \mu$ for all $i = 1, \dots, n$. Recall that $\mu$ is called *expectation* because that's the value we expect the $X_i$'s to take *on average*. Now the problem is that we can't know $\mu$ because we don't know $F$. But is there an observed quantity that *approximates* $\mu$? The average of the $X_i$'s is an obvious candidate.

**Definition 3.1.** *For data $X_1, \dots, X_n$, the* ***sample average*** *or* ***sample mean*** *is defined as*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Intuitively, we expect $\bar{X}_n$ to be a reasonable approximation of $\mu$. And our intuition is right: we shall see that $\bar{X}_n \to \mu$ as $n \to \infty$ (in a certain sense). In plain words: as we get more data, $\bar{X}_n$ gets closer and closer to $\mu$. There is a little twist to the story though. Because $X_1, \dots, X_n$ are random variables, $\bar{X}_n$ is a random variable, too! So we first need a notion of convergence that accounts for randomness:

**Definition 3.2** (Convergence in probability)**.** *Let $Y_1, \dots, Y_n$, be a sequence of random variables and $Y$ another random variable. We say that $Y_n$ converges to $Y$ in probability or $Y_n \to_p Y$, if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n - Y| > \epsilon\big) = 0.$$

In plain words, $Y_n \to_p Y$ means: as we get more and more data ($n \to \infty$), the probability that $Y_n$ is $\epsilon$ away from $Y$ goes to 0. You could also write the definition the other way around: for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n - Y| \leq \epsilon\big) = 1.$$

So with probability going to 1, $Y_n$ and $Y$ become arbitrarily close to each other.

The general definition above involves a random variable $Y$ as the limit. In most cases of interest, the limit $Y$ is actually a constant (i.e., a random variable with zero variance). Generally speaking, most rules for "usual" limits also apply for limits in probability. For example, $Y_n \to_p Y$ and $X_n \to_p X$ imply $Y_n + X_n \to_p Y + X$, and so on. We won't go into detail here; you may just assume that all the rules you know from analysis apply.

Now we're all set to state what I like to call the *fundamental theorem of statistics.*

**Theorem 3.3** (The law of large numbers, LNN)**.** *Let $X_1, \dots, X_n$ be* iid *random variables with $\mathbb{E}[X_i] = \mu < \infty$ and define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\bar{X}_n \to_p \mu.$$

While we can't know $\mu$, $\bar{X}_n$ is something we observe. The LNN implies that the sample mean $\bar{X}_n$ is a reasonable approximation of $\mu$. Hence, $\bar{X}_n$ gives us a "feeling" what the actual mean $\mu$ might be. The LNN makes this intuition mathematically precise. It allows us to learn about the expected value of an unknown random mechanism just from seeing the data.

**Example 3.4.** *Let's illustrate the LLN with a small experiment: We simulate $X_1, \dots, X_n \sim$ Bernoulli$(0.5)$ and compute $\bar{X}_n$ for each $n$. We repeat this experiment three times. By the law of large numbers, we expect the three resulting sequences to converge to the expected value $\mathbb{E}[X_1] = 0.5$. The results are shown in Figure 3.1. Each line (color) corresponds to a sequence $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots,$ one*

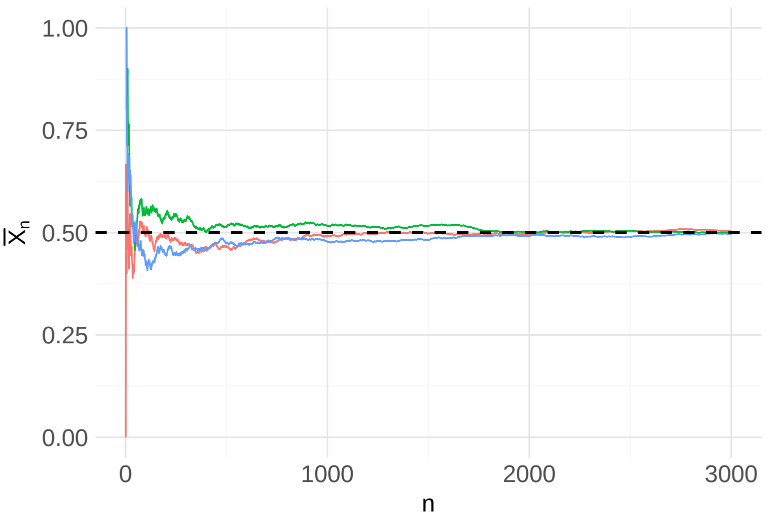


Figure 3.1: The law of large numbers in action (Example 3.4). Each line corresponds to a sequence $\bar{X}_n$ after simulating from $n$ *iid* Bernoulli(0.5) random variables.

*line for each repetition of the experiment. We see that for small $n$, $\bar{X}_n$ can be quite far away from the mean. As we increase the amount of data, all three lines seem to stabilize around 0.5. However, the three lines are different, reflecting the randomness of our sample. The green line lies mainly above 0.5, the others mainly below. The LLN states that, despite this randomness, it becomes less and less likely that one of the lines ends up away from 0.5.*

**Remark 3.1.** *Just as a side note: There is a stronger version of the LLN called the* strong law of large numbers *(SLLN). It involves a different notion of convergence called* almost sure convergence. *It states that the the probability that the sequence $X_n$ converges to $\mu$ is exactly 1: $\mathbb{P}(\lim_{n\to\infty} X_n = \mu) = 1$. Here we're making a probability statement about the limit $\lim_{n\to\infty} X_n$. Convergence in probability is a statement about convergence of probabilities. For us, convergence in probability will be enough, but it's good to have heard about almost sure convergence.*

### 3.1.1 Estimators and consistency

The statement "$\bar{X}_n$ is a good approximation of $\mu$" is made mathematically precise by $X_n \to_p$. In that case, we say that $\bar{X}_n$ is a *consistent estimator* for $\mu$. Let us put this in a slightly more abstract setting.

**Definition 3.5** (Estimator)**.** *If $X_1, \dots, X_n$ is our data, any quantity that can be expressed as $g(X_1, \dots, X_n)$ for some function $g$ is called an **estimator**.*

Less formally, an estimator is any number that you compute from data.

**Definition 3.6** (Consistency)**.** *Let $\theta$ be an unknown quantity that we are interested in. An estimator $\widehat{\theta}_n$ is called* ***consistent*** *for $\theta$ if*

$$\widehat{\theta}_n \to_p \theta.$$

**Example 3.7.** *Let $\theta = \mathbb{E}[X]$. Then $\widehat{\theta}_n = \bar{X}_n$ is a consistent estimator.*

## 3.2 Sample (co)variances

The sample mean is an estimator for the expectation. We can similarly find estimators for other summaries of a distribution.

**Definition 3.8.**

- ***Sample variance***: $S_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}\big(X_i - \bar{X}_n\big)^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$
- ***Sample standard deviation***: $S_n = \sqrt{S_n^2}.$

**Theorem 3.9.** *If $X_1, \ldots, X_n$ are* iid *samples from a distribution $F$, it holds $S_n^2 \to_p \mathbb{V}[X]$ and $S_n \to_p \sqrt{\mathbb{V}[X]}$, for a random variable $X \sim F$.*

*Proof.* By the LLN $\bar{X}_n \to_p \mathbb{E}[X]$ and, thus, $(\bar{X}_n)^2 \to_p \mathbb{E}[X]^2$. Similarly (defining $Y_i = X_i^2$), the LLN gives that $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to_p \mathbb{E}[X^2]$. In combination this yields

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 \to_p \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{V}[X].$$

Finally, because the square root is a continuous function, $S_n = \sqrt{S_n^2} \to_p \sqrt{\mathbb{V}[X]}$. □

Similarly, we can define estimators of the covariance and correlation for two-dimensional *iid* data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

**Definition 3.10.**

- ***Sample covariance:*** $C_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$.
- ***Sample correlation:*** $R_n = \frac{C_n}{\sqrt{S_{n,X}^2}\sqrt{S_{n,Y}^2}}$,

  *where*

$$S_{n,X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \qquad S_{n,Y}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

Again using the LLN, we find that $C_n$ and $R_n$ are consistent estimators for $\mathbb{C}\mathrm{ov}[X, Y]$ and $\rho(X, Y)$, respectively.

### 3.2.1 Using summaries in EDA

The interpretation of the sample quantities is similar to their population versions[1]

- The sample mean is a measure of location.
- Sample variance and standard deviation are measures of variability.
- Sample covariance and correlation are measures of dependence.

Computing these summaries at the very start of a data analysis is a good idea. They give us a "feeling" of the behavior of certain variables: location, variability, and dependence.

Let's see a few examples. Figures 3.2 to 3.4 show *scatterplots* of two variables $X$ and $Y$: each dot represents one sample $(X_i, Y_i)$ in the data set (with $X_i$ drawn on the x-axis and $Y_i$ on the y-axis). The figure captions contain the sample means, standard deviations, and correlations computed from these data sets. As you can see from Figure 3.2, $X$ is located around 0.5, $Y$ around 3. The variability of $X$ is much larger than the variability of $Y$, which is reflected in the sample standard deviations. The correlation is around 0, so there's not much dependence going on. In Figure 3.3, the sample means and standard deviations remain the same, but now we have a correlation of 0.62. We can see this dependence in the scatterplot by the upward trend in the data: when $X$ is small, $Y$ tends to be small; when $X$ is large, $Y$ tends to be large. This is what we call positive dependence. In Figure 3.4 the correlation changes to $-0.82$. So now we have negative dependence which is reflects the downward trend: when $X$ is small, $Y$ tends to be large; when $X$ is large, $Y$ tends to be small.

[1] Everything computed from the actual (unknown) distribution $F$ is called a "population version". Everything computed from data is called a "sample version".

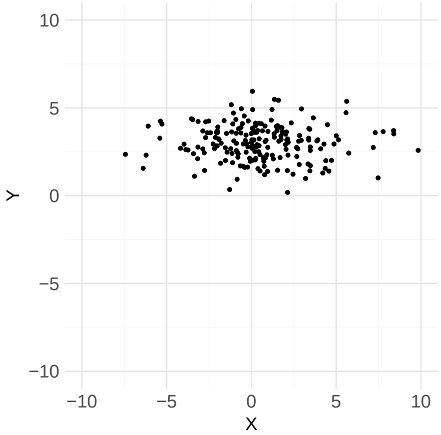


Figure 3.2: Example with $\bar{X}_n = 0.54$, $\bar{Y}_n = 3.02$, $S_{n,X} = 2.85$, $S_{n,Y} = 1.02$, $R_n = -0.04$.

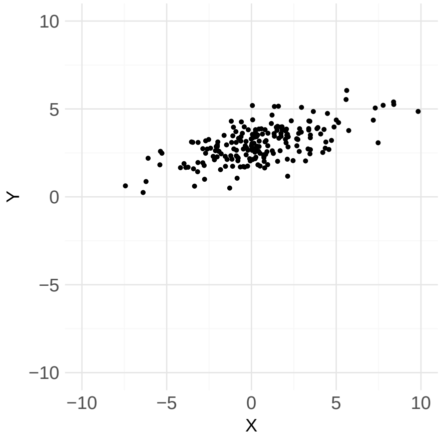


Figure 3.3: Example with $\bar{X}_n = 0.54$, $\bar{Y}_n = 3.02$, $S_{n,X} = 2.85$, $S_{n,Y} = 1.02$, $R_n = 0.62$.

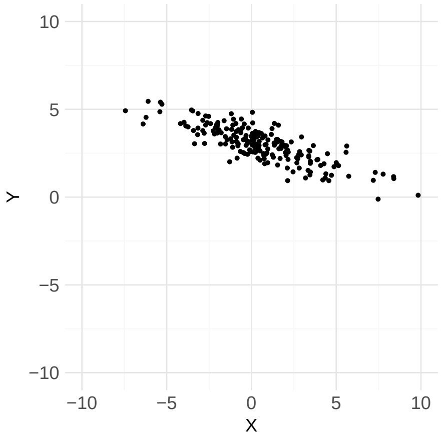


Figure 3.4: Example with $\bar{X}_n = 0.54$, $\bar{Y}_n = 3.02$, $S_{n,X} = 2.85$, $S_{n,Y} = 1.02$, $R_n = -0.82$.

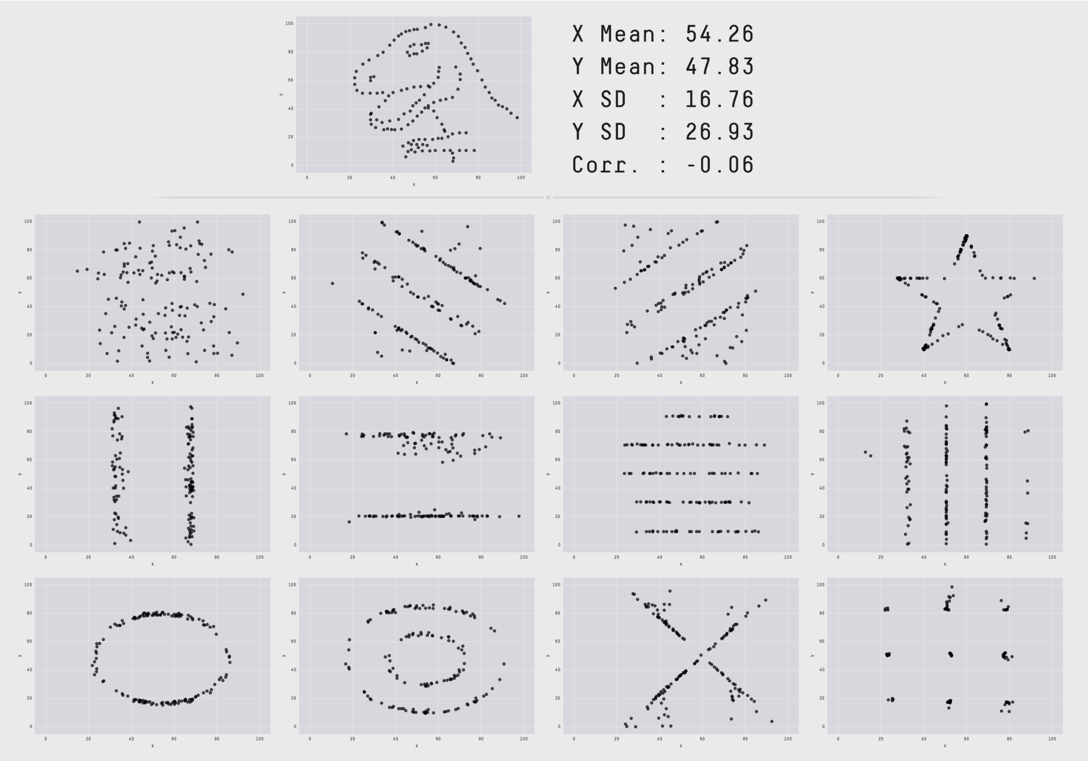


Figure 3.5: The Datasaurus: all scatterplots have the same mean, standard deviation, and correlation.

**Beware of the Datasaurus!**

Summarizing the data into a few numbers is nice, because it gives us a quick overview of what's going on. But never forget that this is a simplification! Two data sets with the same summaries can be wildly different. The *Datasaurus* has established itself as the mascot of this piece of wisdom. All scatterplots in the figure have the same means, standard deviations, and correlations. Yet the data sets couldn't be more different. Take this as a cautionary tale: summarizing your data first is fine, but always make plots to check what's really going on.

## 3.3 The empirical distribution function

Instead of estimating summaries of a distribution $F$, we can also estimate the distribution function itself. To do so, we define the *indicator function*

$$\mathbb{1}(A) = \begin{cases} 1, & A \text{ is true,} \\ 0, & A \text{ is not true.} \end{cases}$$

So how does this help to estimate $F$? You can verify that $F(x) = \mathbb{P}(X \leq x) = \mathbb{E}[\mathbb{1}(X \leq x)]$. So estimating $F$ isn't much different from estimating an expectation (except that we have to estimate one expectation for every $x \in \mathbb{R}$).

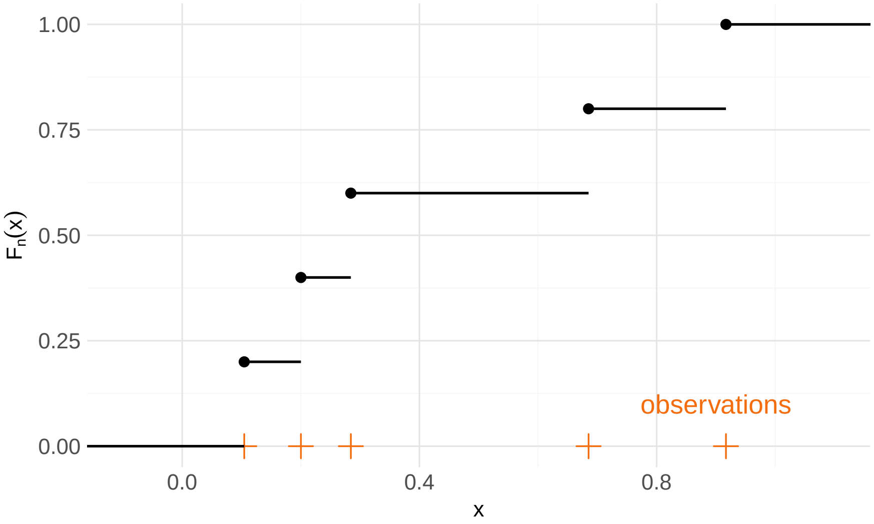


Figure 3.6: The empirical cumulative distribution function of a simulated data set $X_1, \ldots, X_5$. Orange crosses indicate the observations.

**Definition 3.11** (Empirical cumulative distribution function, ECDF)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} F$. *Then the ECDF* $F_n$ *is defined as*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x).$$

*(Similarly, we can define* $F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \leq y)$ *as the ECDF of a bivariate distribution.)*

The formula is very intuitive. Recall that $\mathbb{1}(X_i \leq x) = 1$ for exactly those $X_i$ with $X_i \leq x$. (Otherwise $\mathbb{1}(X_i \leq x) = 0$). Hence, $\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$ is counting how many observations $X_i$ are less than $x$. Dividing by $n$ gives us the proportion of samples that are less than $x$. Intuitively, this proportion should be a good approximation of the probability that $X \leq x$.

The ECDF $F_n$ is a function very similar to the CDF of a discrete random variable. In fact, it is the CDF of a discrete random variable $\tilde{X}$ that puts probability mass $1/n$ on each data point. As a result, the ECDF is right continuous and jumps by $1/n$ at every observation. You can see this in an example with $n = 5$ in Figure 3.6. The crosses represent the location of our five observations $X_1, \ldots, X_5$. The ECDF starts out at 0 until we encounter the first observation (coming from the left). At each of the observations the ECDF jumps up by $1/5 = 0.2$ until it reaches 1, from where on it remains constant.

We can show that the ECDF $F_n(x)$ is a consistent estimator for $F(x)$. In fact it is consistent *uniformly*, i.e. for all values of $x$ at the same time.

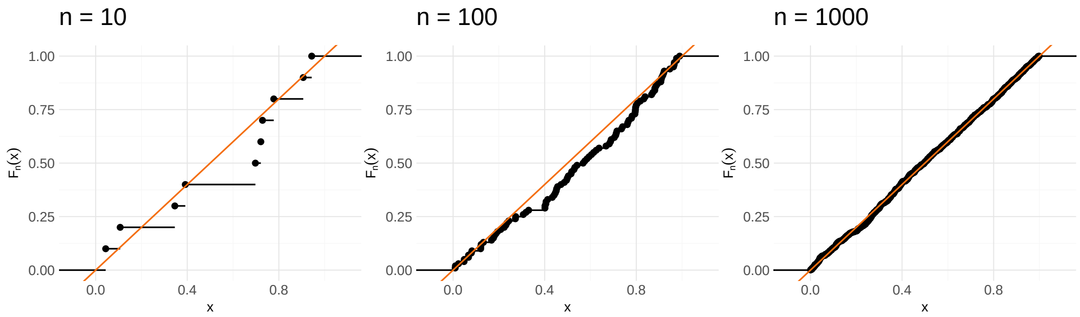


Figure 3.7: Uniform convergence of the ECDF $F_n$ to the CDF $F$ of a Uniform$(0, 1)$ distribution.

**Theorem 3.12** (Glivenko-Cantelli theorem)**.** *Let* $X_1, \dots, X_n \overset{iid}{\sim} F$*. Then*

$$\max_x |F_n(x) - F(x)| \to_p 0.$$

*In particular,* $F_n(x) \to_p F(x)$ *for all* $x \in \mathbb{R}$*.*

This convergence is visualized in Figure 3.7. We simulate data sets from the Uniform$(0, 1)$ distribution of increasing size $n$. The true CDF is $F(x) = x$ which is shown as as the straight diagonal line. For $n = 10$ we are fairly close in some regions , but far away in others (around $x = 0.7$). As $n$ increases we get closer and closer to the true CDF. And we do so in away that is uniform in the sense that there's no region where our approximation remains bad. For $n = 1000$ the true CDF $F$ and the ECDF $F_n$ are hardly distinguishable.

## 3.4 The histogram

As mentioned in the last chapter, the CDF $F$ is mathematically important, but it's hard to interpret the graphs. The same is true for the ECDF. Instead, we would much rather have an estimate of the PDF/PMF, which are easier to interpret. The *histogram* does just that:

(i) For some $x_0 < x_K$, we divide the interval $(x_0, x_K]$ into $K$ *bins* $B_k = (x_{k-1}, x_k]$ of equal size $\delta = x_1 - x_0 = x_2 - x_1 = \dots$

(ii) We compute $N_k = \sum_{i=1}^n \mathbb{1}(X_i \in B_k)$, the number of observations that fall into each bin $B_k$, $k = 1, \dots, K$.

(iii) For $x \in B_k$, the histogram is defined as

$$h_n(x) = \frac{N_k}{n \times \delta}.$$

This process is visualized in Figure 3.8. The data is shown as crosses in the top panel. Then the interval $(0, 3.5]$ is divided into 7 bins of equal size (mid panel). Then we count the number of observations per bin to compute the relative frequencies in step 3 (bottom panel).

The histogram is extremely powerful. With a single glance we get a good feeling for the shape of the entire distribution. Two important characteristics to look out for are *skew* and *modality*. Skew is related to symmetry: a distribution is called right-skewed if the histogram/density is leaning to the left and left-skewed if it is leaning to the right. (I know that's confusing, but it is what it is.) Modality tells us about potential clusters (showing up as bumps in the graph). Each bump is called a *mode*; for example, a density with two distinct bumps is called *bimodal*. You can see some exemplary graphs in Figure 3.9.

**Example 3.13.** *Figure 3.10 shows a histogram for the metallicity of globular clusters in the Milky way (relative to the sun) with $K = 10$ bins. We see that the distribution is slightly right skewed and potentially bimodal (with one bump around -1.5 and a possible second one at -0.9). But why did I choose 10 bins? In Figure 3.11 we see histograms for the same data, but this time with $K = 2$ (left) and $K = 100$ (right). There's not much we can learn from two bins, the graph is hiding most of the information in two huge blocks. On the other hand, the graph with 100 bins is extremely erratic and we wouldn't expect the true density to have a shap with that many peaks and troughs.*

This example illustrates that the number of bins is crucial for getting meaningful information out of a histogram. Choosing this number is more or less guesswork unfortunately. A good *rule of thumb* is $K \approx 2n^{1/3}$. (There's some theory behind this, but every data set is different.) In practice, we usually try a few values and see what works best.

Let's conclude with a theorem on the consistency of the histogram. This (almost) follows from the Glivenko-Cantelli theorem.

**Theorem 3.14.** *Let $X_1, \dots, X_n \overset{iid}{\sim} f$, and $K \to \infty$, $K/n \to 0$ as $n \to \infty$.*

*(i) If $f$ is a PDF, then $h_n(x) \to_p f(x)$ for all $x$.*

*(ii) If $f$ is a PMF, then $\delta h_n(x) \to_p f(x)$ for all $x$.*

## 3.5 Quantiles and the boxplot

The *boxplot* is another visual summary of the distribution and often used to detect *outliers* (very unusual observations). It is based on *quantiles*, which we have to define first.

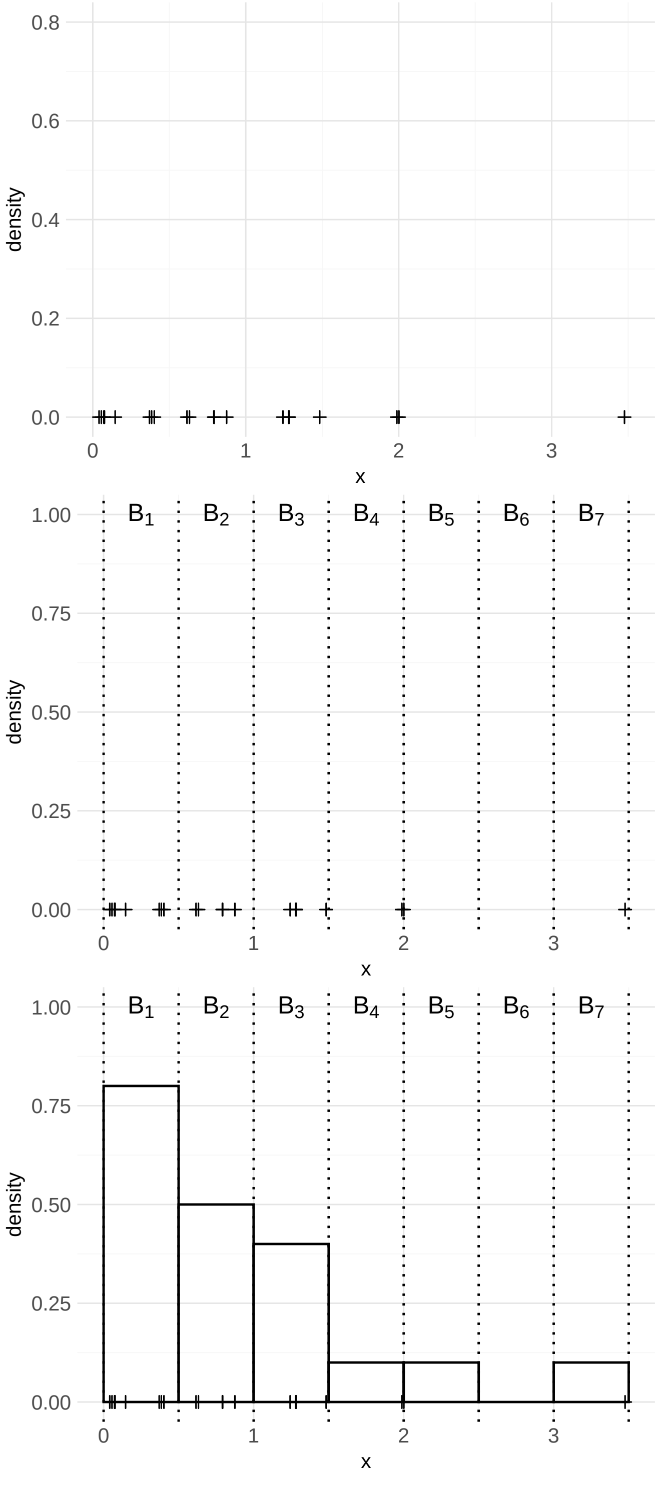


Figure 3.8: Construction of a histogram: the data is shown the top panel, the interval $(0, 3.5]$ is divided into bins (mid panel), relative frequencies drawn in the bottom panel.

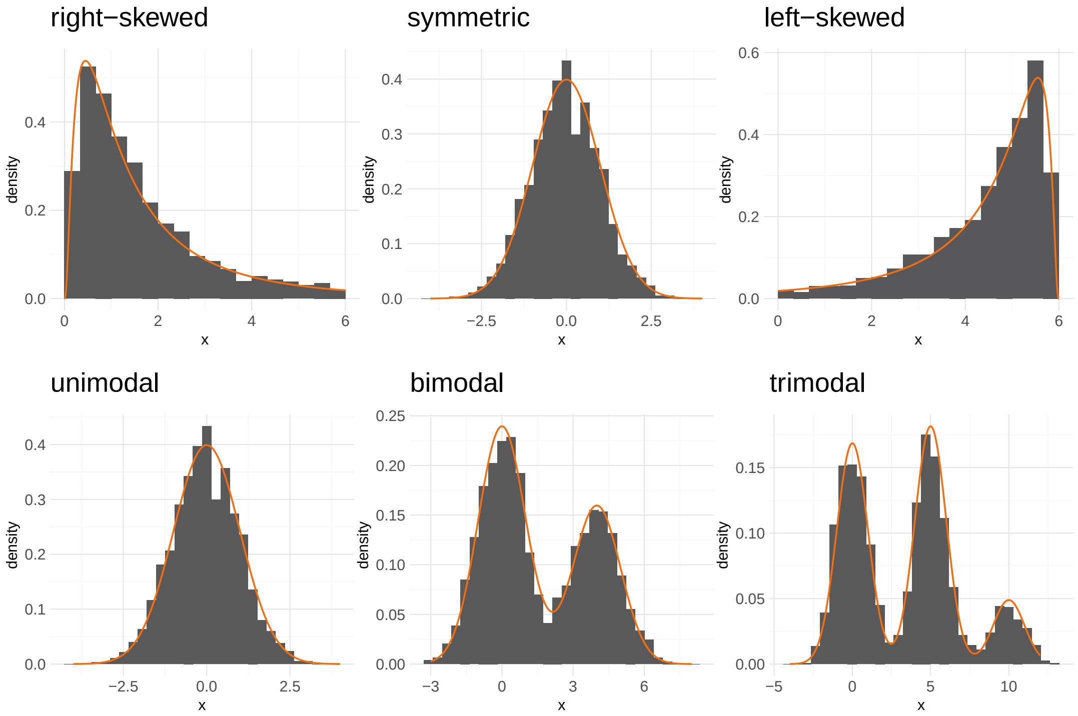


Figure 3.9: Common shapes of a distribution. Histograms with the the true density superimposed as orange line.

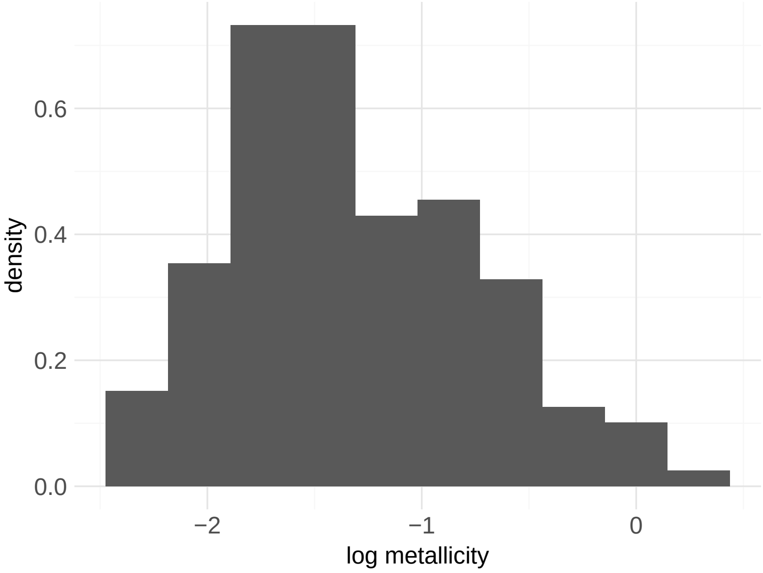


Figure 3.10: Histogram for the metallicity of globular clusters in the Mikly way (relative to the sun).

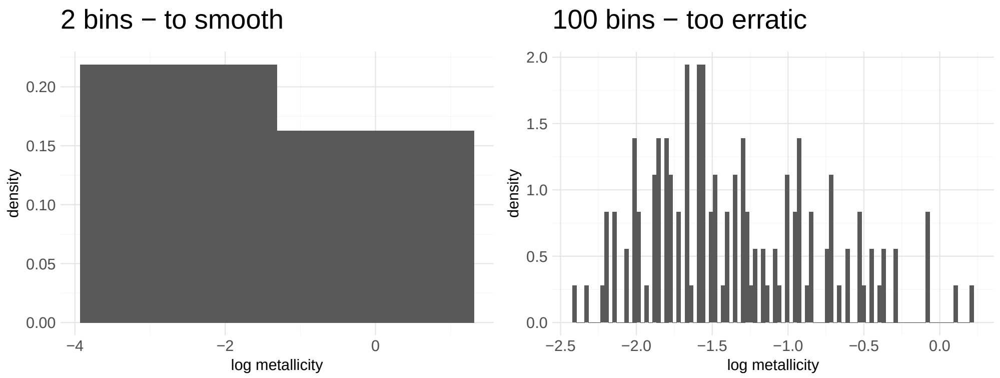


Figure 3.11: Histograms for the metallicity of globular clusters in the Mikly way (relative to the sun) with too few and too many bins.

**Definition 3.15** (Quantile)**.** *Let $F$ be a CDF. The corresponding $p$-quantile is defined as*

$$Q(p) = \min\{x \colon F(x) \geq p\}.$$

The definition is a bit weird, so some discussion is in place. First consider the case where $F$ is continuous. Then $Q(p) = F^{-1}(p)$ is just the inverse function of the CDF $F$. The weirdness only comes in for discrete distributions, where the CDF is not strictly increasing and, thus, no inverse exists. Conceptually, the definition above is equivalent to an inverse function though.

Quantiles answer the question: which value of $X$ is not exceed with a probability of $p$? For example, if $Q(0.01) = -5$, then the probability that $X$ is less than -5 is 1%. More generally, the $Q(p)$ divides the real line into two parts: the first part, $X \leq Q(p)$ has probability $p$, and the second part,$X > Q(p)$ has probability $1 - p$.

Given data $X_1, \dots, X_n$, we define the sample $p$-quantile as:

$$Q_n(p) = F_n^{-1}(p) = \min\{x \colon F_n(x) \geq p\}.$$

Let's write this in a more intuitive way. Denote $\lceil a \rceil$ as the smallest integer $k$ with $k \geq a$ ('rounding up'). Now $Q_n(p)$ is defined such that

- $\lceil np \rceil$ of the observations are less or equal to $Q_n(p)$,
- $\lceil n(1-p) \rceil$ of the observations are larger or equal to $Q_n(p)$,

So a proportion of at most $p$ of the data points is less than $Q_n(p)$, and a proportion of at most $1 - p$ is larger than $Q_n(p)$. To compute this number, do the following: First sort the data $X_1, \dots, X_n$ in ascending order. This gives an ordered data set $X_{(1)}, \dots, X_{(n)}$, where $X_{(k)}$ denotes the $k$th smallest observation. Then set[2]

$$Q_n(p) = X_{(\lceil np \rceil)}.$$

[2] Most software/books make an adjustment if $np$ is an integer.

Some quantiles have special names:

- $Q(1/2)$: median,
- $Q(1/4)$, $Q(3/4)$: lower quartile and upper quartile.

Similar to the mean, the median is another measure of location. It is defined such that half of the data/population is less than this number, and half of the data/population is larger than this number. So the median is really the "average guy". Similar to the variance, the *inter-quartile range* $IQR = Q(3/4) - Q(1/4)$ is a measure of variation. Such quantile-based measures are called *robust* because they're not sensitive to individual observations. Let's see this in an example:

**Example 3.16** (Robustness of median)**.** *Suppose we have data on the monthly net income (in* 1000 *EUR):*

$$X_1 = 3, \quad X_2 = 2.6, \quad X_3 = 3.6, \quad X_4 = 4.6, \quad X_5 = 1.7.$$

*We compute* $\bar{X}_n = 3.1$ *and* $Q_n(1/2) = 3$*. What happens to the mean and median if we change* $X_4 = 300$*? Well the median doesn't change at all, but the mean increases by an order of magnitude.*

The observation $X_4 = 300$ is called an outlier, because it is so different from all the other data points. The presence of a few outliers drastically changes the sample mean and variance, but the median and IQR are only little affected (if at all). In most countries, the majority of the population has a similar income, but there's the infamous 'top 1%' that earns muuuuch more than all the others. In that case, the vast majority of people would earn way less than the *average income*, so the sample average is not really representative of an average person. However, the *median income* always represents this average person: 50% earn more, 50% earn less.

The *boxplot* is a visual tool that i) gives a visual summary of the distribution, and ii) helps to identify potential outliers. An exemplary boxplot is shown in Figure 3.12. The line in the middle of the box indicates the median, a measure of location. The upper and lower boundary of the box are the upper and lower quartiles; the distance between them tell us something about the variability of the data. The lines leaving the box are called *whiskers*. (Their exact definition differs between implementations.) All data points that exceed the whiskers are drawn as dots. These points are potential outliers.

In the income example above, the outliers are a feature of our reality, so we should not ignore them (and not call them outliers). In other cases, outliers may come from faulty measurements or the guy entering the data having fat fingers. Such outliers we should filter out before proceeding with the data analysis. Detecting *potential* outliers is an important part of EDA. After having detected one, we must always ask ourselves:

(i) whether it is due to random variation or faulty measurement,

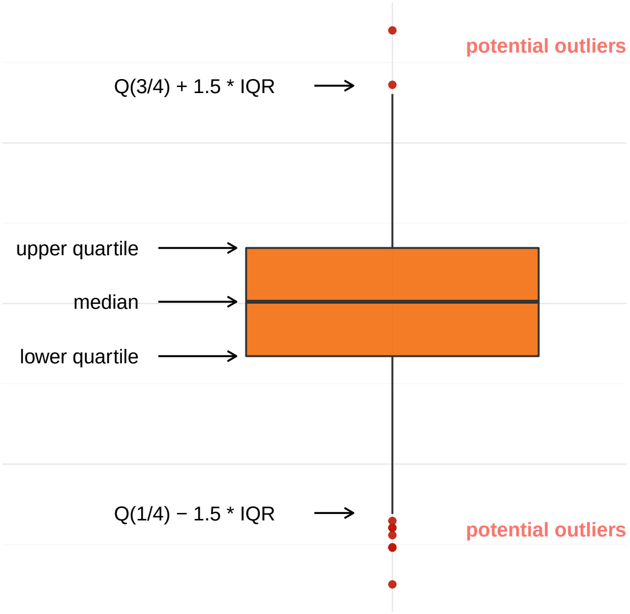


Figure 3.12: The components of the boxplot explained.

(ii) how it affects further analyses,

(iii) whether or not to remove it from the data.

Deciding whether to keep or remove a potential outlier is always a judgment call and should be based on domain knowledge.

## 3.6 A recipe for EDA

Now that we've learned about all these summaries, let's put them together. Whenever you start your data analysis, it's a good idea to follow the following recipe:

1. High-level summary: how many observations, which variables (units), missing values.
2. For each variable, compute and interpret summary statistics for location and scale.
3. For each variable, plot and interpret boxplot and/or histograms. Consider: skewness, modality, outliers
4. If more than one variable, also consider their dependence:
   - pair-wise scatterplots,
   - compute correlation,

- check for potential outliers.

Now let’s walk through this process with some real data.

## 3.7 Case study: Colors of quasars vs. Galaxies

We shall perform an EDA for data about the colors of quasars and galaxies. The data was obtained from

http://cas.sdss.org/dr16/en/tools/search/sql.aspx

by the following SQL query:

```sql
SELECT TOP 1000 p.u, p.g, p.r, s.z, s.class
FROM   photoobj AS p
       JOIN specobj AS s
       ON s.bestobjid = p.objid
WHERE  p.u BETWEEN 0 AND 19.6
       AND p.g BETWEEN 0 AND 20
       AND s.class <> ’UNKNOWN’
       AND s.class <> ’STAR’
       AND s.class <> ’SKY’
       AND s.class <> ’STAR_LATE’
```

Our goal is to get a feeling for what is going on in this data. What follows will be a very brief version of the process, mainly because I know too little about astronomy to tell you something interesting. Maybe you see some more interesting things?

### 3.7.1 High-level summary

- There are $n = 1\,000$ observations,
- There are 5 variables:
  - u: $u$-band apparent brightness (magnitude relative to sun),
  - g: $g$-band apparent brightness (magnitude relative to sun),
  - r: $r$-band apparent brightness (magnitude relative to sun),
  - z: redshift,
  - class: galaxy ($n_G = 863$) or quasar ($n_Q = 137$).
- There are no missing values.

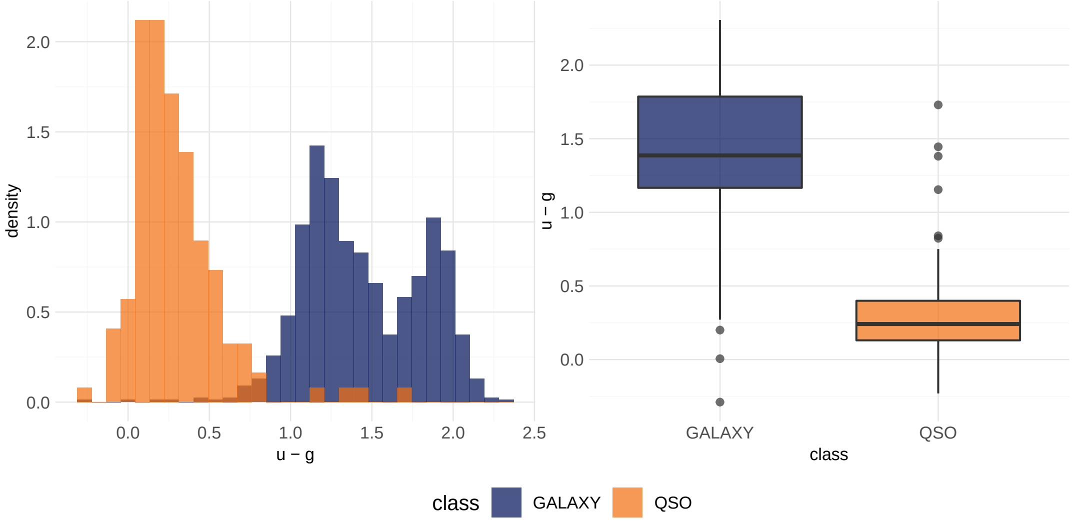
To assess the colors, we define the following new variables: ‘u - g’ = u - g (green-ness) and ‘g - r’ = g - r (red-ness). We can then forget about the original u, g, r variables.

### 3.7.2 EDA for individual variables

**Variable** u - g

| Galaxies | | Quasars | |
|---|---|---|---|
| mean | std. dev. | mean | std. dev. |
| 1.44 | 0.37 | 0.30 | 0.28 |

We see that galaxies in this data set tend do be greener than quasars and also the variability seems slightly larger for galaxies.

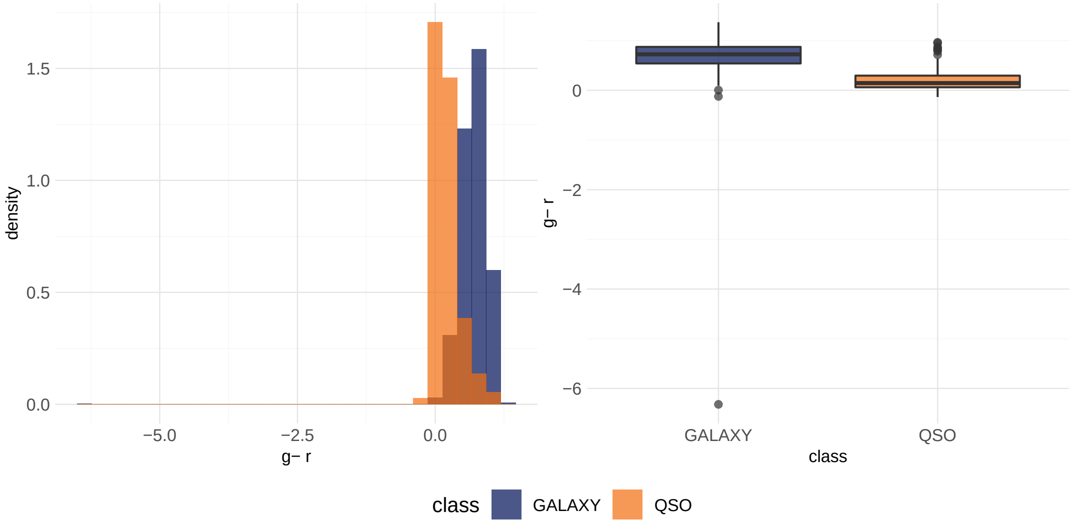


We can clearly identify the two populations in the u -g histogram and boxplot (galaxies and quasars). Above u - g= 1, we find almost no quasars, below this value we find almost no galaxies. The histogram for the quasars is roughly symmetric and unimodal. The histogram for the galaxies is bimodal, indicating that there may be two sub-populations of galaxies. The boxplots show a few potential outliers, but the points don't seem to crazy. It is certainly plausible that a Quasar has u - g $\approx 1.5$. Without further reasons, we should keep them in the data set.

**Variable** g - r

| Galaxies | | Quasars | |
|---|---|---|---|
| mean | std. dev. | mean | std. dev. |
| 0.70 | 0.32 | 0.21 | 0.22 |

We see that galaxies are redder than quasars on average and also more variable.

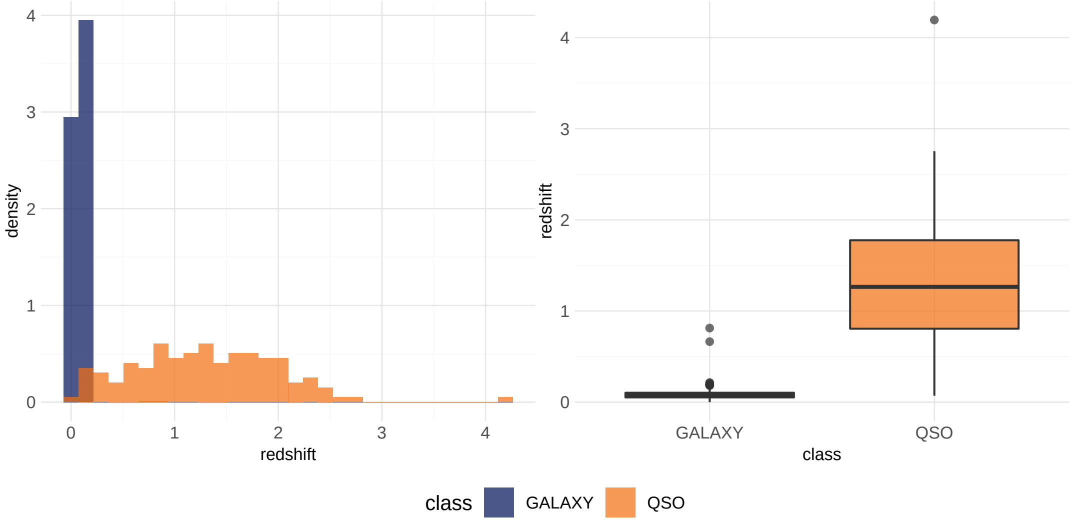


In the boxplot there is a clear outlier galaxy with `g - r` $= 6$. This looks very suspicious and it definitely affects our data analysis. Optimally, we would now look this galaxy up online and check whether it is really that red. And if it is, we still need to ask ourselves if we want to keep it or focus on more 'normal' galaxies. If we exclude it, we should also recompute all summaries and graphs above. But let's move on for now.

### Variable z

| Galaxies | | Quasars | |
|---|---|---|---|
| mean | std. dev. | mean | std. dev. |
| 0.08 | 0.05 | 1.30 | 0.69 |

Unsurpringly, Quasars tend to be further away from us than galaxies. The variablity in redshift is also much larger. That makes sense.



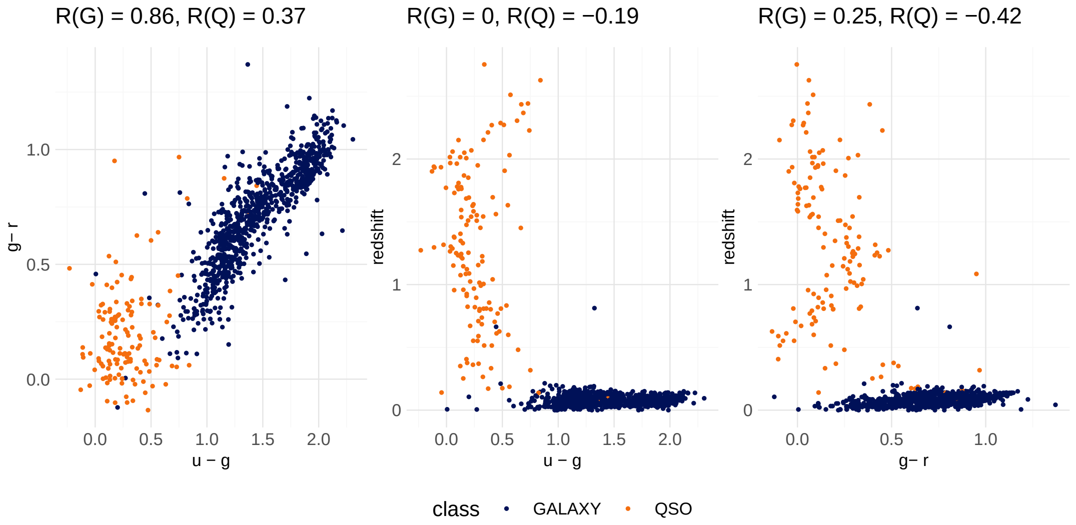
We see that more clearly in the histogram and boxplot. While all galaxies are fairly close ($z < 0.2$), the quasars are spread out wider with a center around 1.5 and one quasar with redshift of more than 4. This quasar also shows up as potential outlier in the boxplot, but $z \approx 4$ is certainly a plausible value for a quasar.

### 3.7.3 Removing outliers

The outlier for `g - r` is weird. Also the quasar with $z > 4$ seems very different from the rest of the objects under study. These outliers would distort the correlation heavily because it is not robust. Hence, I will remove both observations from the data set in what follows.

### 3.7.4 Joint behavior

The following graph contains scatterplots for all possible pairs of variables. At the top you can read the correlation ($R(C)$ =correlation in class $C$).



We see there is a strong positive dependence between colors of galaxies, but much less for quasars. The dependencies between redshift and color are weaker and negative for galaxies. So the further away a galaxy is the less green (or red) it tends to be. The pairwise scatterplots show some galaxies falling far away from the bulk. These should also be considered potential outliers.

### 3.7.5 Wrap up

This was a quick walk through the steps of an EDA. As a statistician, I can compute numbers and draw graphs. But that's only useful in combination with domain knowledge. As an astronomer, you should always try to interpret the result in the numbers and graphs in context. Always ask yourself if what you see is in line with your expectation. If something seems implausible, dig deeper. Zoom into a graph, compute summaries for interesting subsets of the data etc. At the end of this process you should i) have a feeling for what's going on in the data, and ii) trust that the data set you continue with is suitable for future modeling steps. We'll learn more about that in the next chapters.

# 4 Parametric statistical models

Let's once again recall the basic problem of statistics.

> *Having observed some data $X_1, \dots, X_n$, what can we say about the mechanism that generated them?*

In the last chapter we learned how to get a 'feeling' for this mechanism. We can now try and come up with plausible mechanisms that *could have* generated the data. Since we don't know the mechanism, what we come up with are just *models* of reality. A *statistical model* involves randomness and is hence characterized by a probability distribution (or density). A *parametric statistical model* is a family of distributions $\{F_\theta \colon \theta \in \Theta\}$ that is characterized by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$. Once we have a parametric model, the main question is which value of the parameter $\theta$ fits the data best. This will be the topic of the next chapter.

In the current chapter, we shall introduce some essential statistical models. This includes parametric families for discrete and continuous distributions as well as multi-dimensional data and prediction problems. At the end of this chapter, you should have heard about the most common models and know in which situations they may or may not apply. That's admittedly a bit boring, but it's the last thing we need in preparation for all the exciting things that follow in the next chapters.

## 4.1 Discrete distribution families

### 4.1.1 The Bernoulli distribution

**Definition 4.1.** *We say that $X$ follows a* ***Bernoulli distribution*** *with parameter $p \in (0,1)$ or $X \sim \mathrm{Bernoulli}(p)$ if the PMF is given as*

$$f(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \\ 0 & \text{otherwise,} \end{cases}$$

*or alternatively $f(x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$. It holds $\mathbb{E}[X] = p$ and $\mathbb{V}[X] = p(1-p)$*

We have already seen the Bernoulli distribution earlier in the course when speaking about coin flips. We can use the (arbitrary) encoding '0 = heads, 1 =

tails' to define a random variabl $X$ that represents the outcome of a single coin flip. We say that $X$ follows a *Bernoulli distribution*with parameter $p \in (0,1)$ or $X \sim \text{Bernoulli}(p)$. The parameter $p = \mathbb{P}(X = 1)$ is called *success probability*. Note that the interpretation of this parameter depends heavily on your coding of the categories (1 = heads vs. 1 = tails).

We rarely flip coins in reality, but the distribution is everywhere nevertheless. Its quite common to put study subjects into two categories:

- yes or no answers,
- dead or alive people,
- radio-quiet and radio-loud galaxies,
- red-sequence and blue-sequence galaxies,
- metal-rich or metal-poor globular clusters

All these categories can be recoded (arbitrarily) to a binary variable that only takes values 0 or 1. When checking the categorie of a random object, we're again faced with the Bernoulli distribution.

### 4.1.2 Binomial distribution

**Definition 4.2.** *Suppose* $Y_1, \dots, Y_k \overset{iid}{\sim}$ *Bernoulli*$(p)$*. Then we say that* $X = \sum_{i=1}^k Y_i$ *follows a* ***Binomial distribution*** *with parameters* $n$ *and* $p$*, or* $X \sim \text{Binomial}(n,p)$*. We have*

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, \dots, n \\ 0 & \text{otherwise,} \end{cases}$$

$\mathbb{E}[X] = np$ *and* $\mathbb{V}ar[X] = np(1-p)$*.*

If we throw a coin 10 times, how many heads do we get? That's again a toy problem of course, but we can replace the coin with other variables. Out of 100 people receiving cancer treatment, how many survive? Out of 50 random galaxies under study, how many are radio-quiet? In all these questions, we take a sum of Bernoulli variables, so the *Binomial distribution* arises naturally. It has two parameters: the *number of trials* $n$ and the *success probability* $p$.

**Example 4.3.** *Suppose that a proportion* $p$ *of all galaxies are radio-quiet. Pick* 50 *galaxies at random and let* $X$ *be the number of radio-quiet galaxies. Then* $X \sim \text{Binomial}(50, p)$*.*

The PMF is visualized for varying parameter choices in Fig. 4.1.[1] Note that all PMFs are zero for $x > n$ and that they have peak near $\mathbb{E}[X] = pn$.

[1] The dashed lines are only added as visual guides. The PMF is only defined where the dots are.

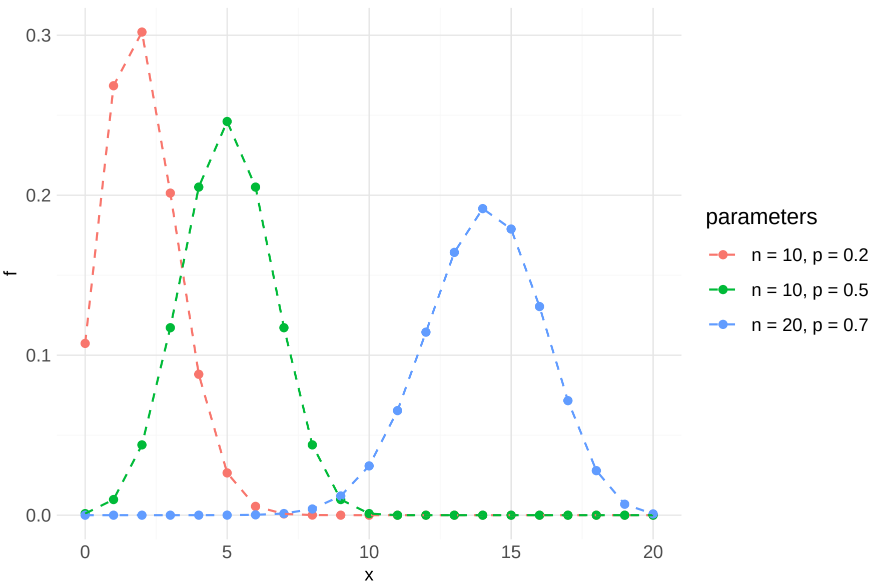


Figure 4.1: Probability mass function of the Binomial distribution for varying parameter choices.

### 4.1.3 Poisson distribution

**Definition 4.4.** *We say that $X$ follows a* ***Poisson distribution*** *with parameter $\lambda > 0$, written $X \sim \text{Poisson}(\lambda)$, if*

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots .$$

One may check $\mathbb{E}[X] = \lambda$, $\mathbb{V}ar[X] = \lambda$. Graphs of the PMF are shown in Fig. 4.2. Note that $f(x) > 0$ for all $x \in \mathbb{N}$.

The Poisson distribution can be derived formally as: the distribution of the *number of events* occurring in a *fixed period* (area/volume/...), if they occur at a *fixed rate* and independently of the time since the last event. Such situations arise often in The Poisson distribution often arises when modeling rare events:

- the number of mutations on a strand of DNA per unit length,
- telephone calls arriving in a system,
- number of radioactive decays in a given time interval.

The distribution has a few interesting properties:

- When $n$ is large and $p$ is small, the Binomial$(n, p)$ distribution is well approximated by the Poisson$(\lambda)$ distribution with $\lambda = np$.
- If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

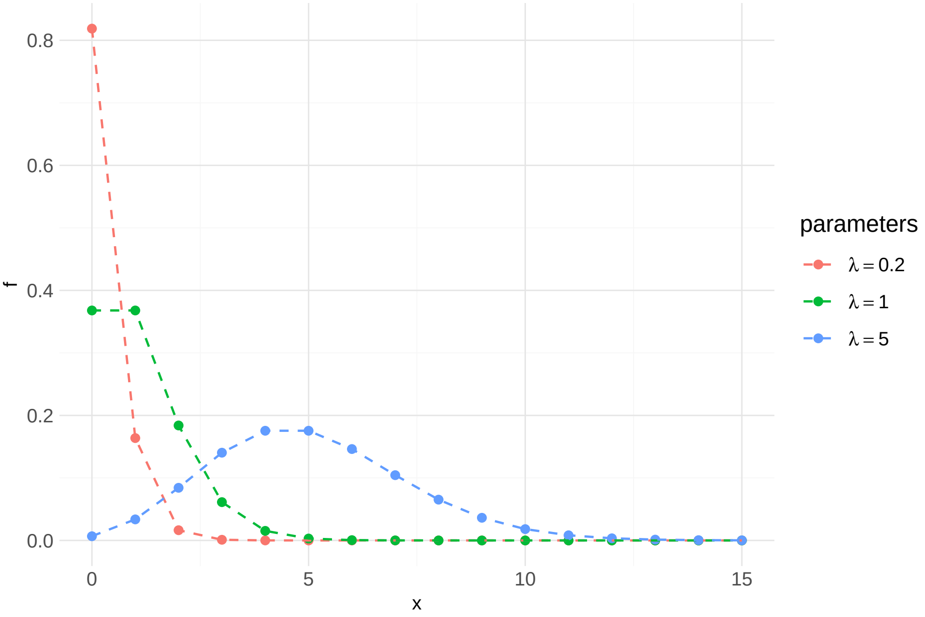


Figure 4.2: Probability mass function of the Poisson distribution for varying parameter choices.

**Example 4.5.** *A distant quasar emits $10^{64}$ photons per second in the X-ray band, but at a earth-orbiting X-ray telescope only ca. $10^{-3}$ photons arrive per second. In a typical observation period of $10^4$ seconds, only around $10^1$ of the $n \approx 10^{68}$ photons emitted during the observation period arrive, giving $p \approx 10^{-67}$. The number of photons arriving can be thought as* $\mathrm{Binomial}(n, p)$*-distributed, and the latter is well-approximated by the* $\mathrm{Poisson}(\lambda)$ *distribution with $\lambda = np \approx 10$.*

## 4.2 Continuous distribution families

### 4.2.1 Exponential distribution

**Definition 4.6.** *We say that $X$ has an **exponential distribution** with rate parameter[2] $\lambda > 0$, written $X \sim \mathrm{Exp}(\lambda)$, if*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

*It holds $\mathbb{E}[X] = 1/\lambda$, $\mathbb{V}ar[X] = 1/\lambda^2$. A graph is shown in Fig. 4.4*

The exponential distribution is widely applied to model times between random events. For example,

[2] Some authors use a different parametrization with 'scale' parameter $\alpha = 1/\lambda$.

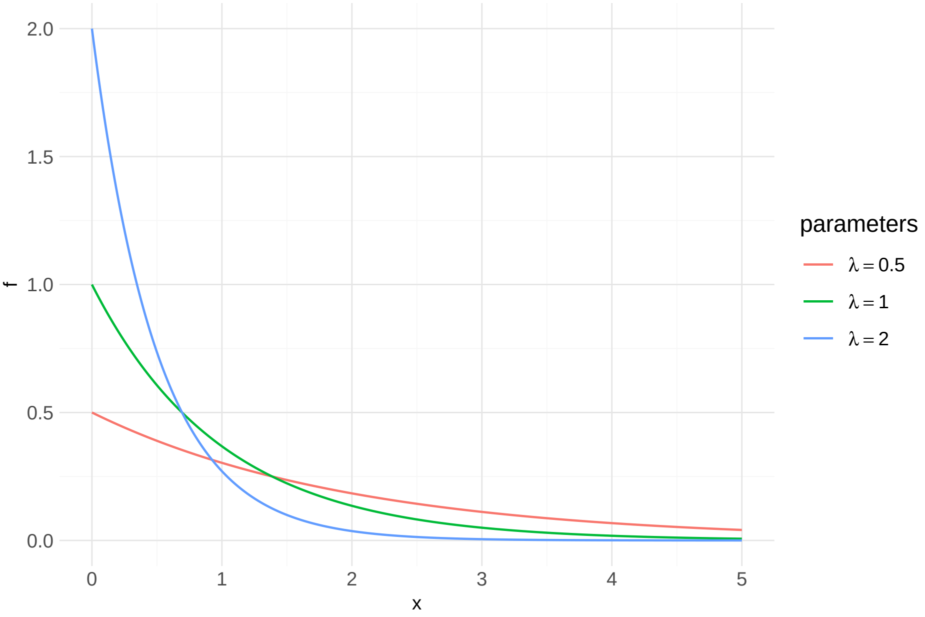


Figure 4.3: Probability density function of the Exponential distribution for varying parameter choices.

- the time it takes before your next telephone call,
- the time between clicks of a geiger counter,
- the lifespan of a bulb.

### 4.2.2 Gamma distribution

The Gamma distribution generalizes the exponential distribution. First we need to define the *Gamma function*:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy, \qquad \alpha > 0.$$

**Definition 4.7.** *X has a* ***Gamma distribution****with parameters shape $\alpha > 0$ and scale $\beta > 0$, written $X \sim \text{Gamma}(\alpha, \beta)$, if*

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

*It hods $\mathbb{E}[X] = \alpha\beta$, $\mathbb{V}ar[X] = \alpha\beta^2$.*

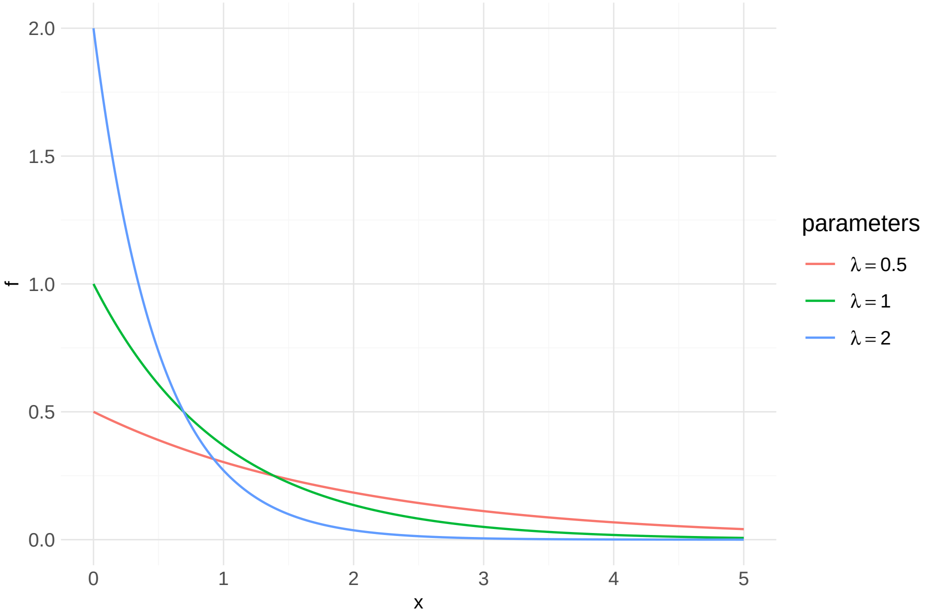


Figure 4.4: Probability density function of the Exponential distribution for varying parameter choices.

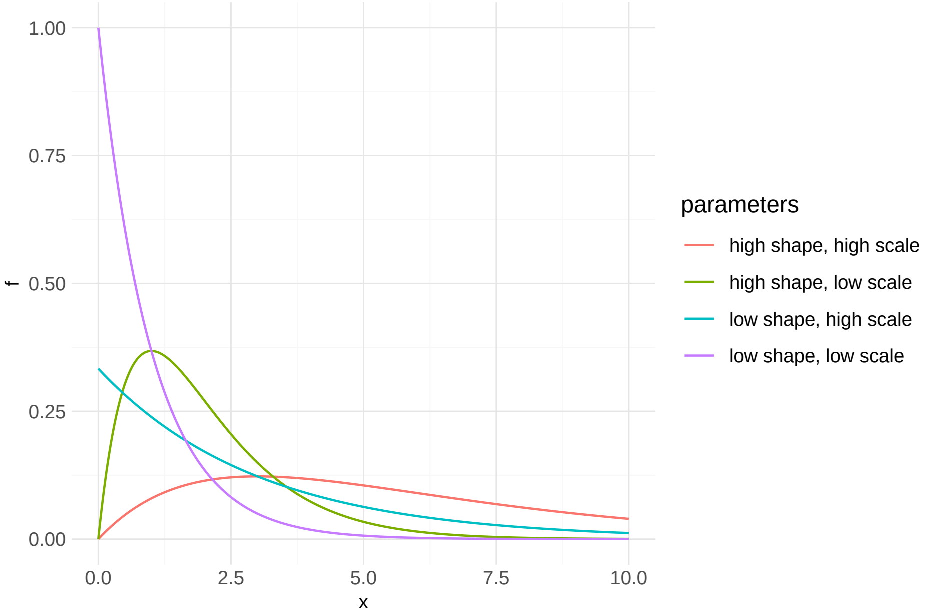


Figure 4.5: Probability density function of the Gamma distribution for varying parameter choices.

Note that the Gamma$(1, \beta)$ distribution is a Exp$(1/\beta)$ distribution. Fig. 4.5 shows the density for varying parameter choices. As you can see, the Gamma distribution is quite flexible. For small values of the shape parameter, the PDF is strictly decreasing. For larger values, the PDF shows a bump.

Because of its flexibility, the Gamma distribution is quite popular for modeling strictly positive, continuous variables. Typical examples are:

- aggregate insurance claims,
- the amount of rainfalls accumulated in a reservoir.

**Example 4.8.** *The gamma distribution has been studied extensively in extragalatic astronomy with respect to the distribution of luminosities, where it is known as the* Schechter luminosity function. *According to the Schechter function the number of stars or galaxies within a luminosity bin of fixed width at luminosity* $\ell$ *is proportional to* $\ell^\alpha \exp(-l/L^*)$.

### 4.2.3 Pareto distribution (power law)

**Definition 4.9.** *The **Power law** or **Pareto distribution** with shape* $\alpha > 0$ *and truncation point* $\xi > 0$*, written as* Pareto$(\alpha, \xi)$ *is defined through the PDF*

$$f(x) = \begin{cases} \alpha \frac{\xi^\alpha}{x^{\alpha+1}} & \textit{for } x \geq \xi, \\ 0 & \textit{otherwise.} \end{cases}$$

A graph of the Pareto density is shown in Fig. 4.6. The truncation parameter $\xi$ determines the smallest value that the random variable $X$ can take. In particular, $\mathbb{P}(X < \xi) = 0$. From there on, the density strictly decreases with what is called *polynomial decay* or a *polynomial tail*: $f(x) \propto x^{-(\alpha+1)}$. The smaller the value of alpha, the slower is the decay. Contrast this to *exponential tails* found in the exponential and Gamma distributions, where (approximately) $f(x) \propto \exp(-ax)$ which goes to zero much faster. The type of decay determines how probable very large values of the random variable $X$ are. When the density decays slowly, large values of $X$ occur at relatively high frequency. In fact we have,

$$\mathbb{E}[X] = \begin{cases} \infty, & \alpha \leq 1 \\ \frac{\alpha\xi}{\alpha-1}, & \alpha > 1. \end{cases}, \qquad \mathbb{V}ar[X] = \begin{cases} \infty, & \alpha \leq 2 \\ \frac{\alpha\xi^2}{(\alpha-1)^2(\alpha-2)}, & \alpha > 2. \end{cases}$$

So when $\alpha$ is very small, large values of $X$ are so probable that the expectation (or variance) is infinite. This is a bit of a mathematical oddity. The interpretation is that large values or so frequent that taking the average of such numbers doesn't yield a stable result, no matter how many numbers we average.

Typical application domains are similar to the Gamma distribution. But now we put more probability mass on very large values of the random variable. Examples are:

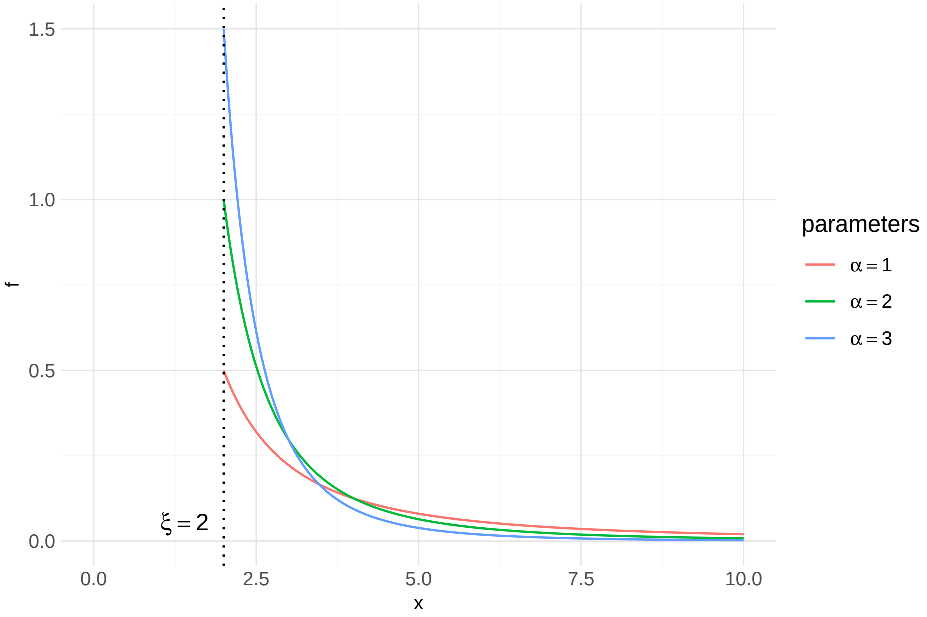


Figure 4.6: Probability density function of the Pareto distribution for varying parameter choices.

- the Angstrom exponent in aerosol optics,
- Pareto's law of income distribution,
- extreme value theory (stock market crashes, natural disasters),
- populations of cities (Gibrat's law),

**Example 4.10.** *Imagine taking a random sample of stars, which are just entering the main sequence. The masses of such stars are called initial masses. The probability density of their masses is called the* initial mass function *(IMF). Let us measure mass $m$ in multiples of 1 solar mass. Salpeter discovered that the* number of stars *with mass $m$ appears to decrease as a power law (at least for the larger stars).*

### 4.2.4 Normal distribution

The next family is more widely known: the *normal distribution*. Another common name is *Gaussian distribution*, because it was discovered by Carl Friedrich Gauss around the year 1800 as a by-product of his astronomical studies.

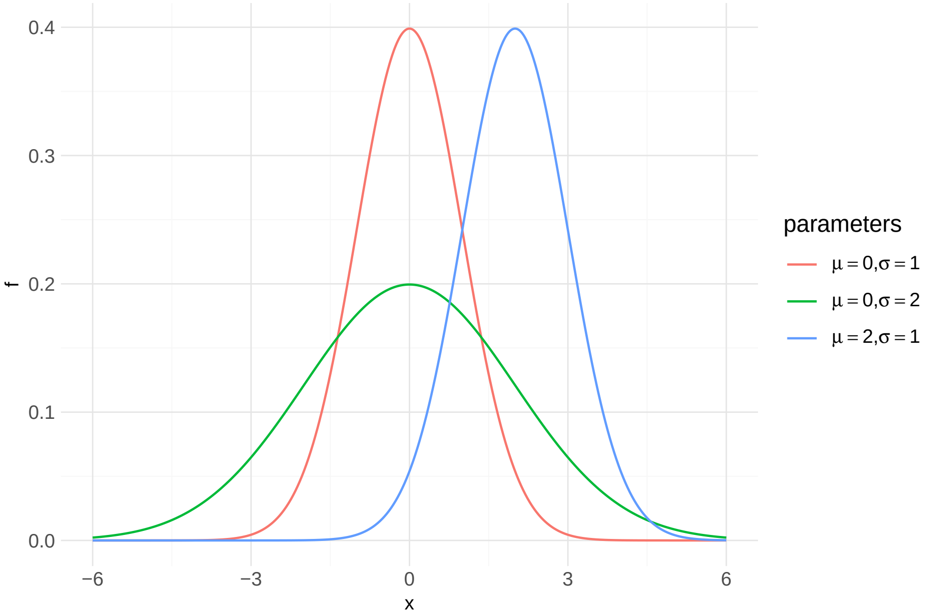


Figure 4.7: Probability density function of the Normal distribution for varying parameter choices.

**Definition 4.11.** *$X$ has a* ***normal distribution*** *with mean parameter $\mu \in \mathbb{R}$ and variance parameter $\sigma^2 > 0$, denoted $X \sim N(\mu, \sigma^2)$, if*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

As you might expect from the parameter names, it holds $\mathbb{E}[X] = \mu$, $\mathbb{V}ar[X] = \sigma^2$. If $\mu = 0$ and $\sigma = 1$, we shall say that $X$ has a *standard normal* distribution. The PDF and CDF of the standard normal random variable are conventionally denoted by $\phi(z)$ and $\Phi(z)$, respectively.

Fig. 4.7 shows graphs of normal density functions. The role of the parameters is quite obvious. The densities have bell shape, symmetric around a peak at $\mu$. So this parameter is used to shift the location of the distribution. The spread of the distribution is determined by the variance parameter $\sigma^2$, where larger values spread the probability mass out more.

The normal distribution is considered the *most important distribution* in statistics. Before the change to Euros, it was even featured prominently on the most common Deutsche Mark bill. (In Fig. 4.8 you can see during my PhD defence, explaining what a PDF is using the graph on the bill.) It is called ‘normal’ because so many things we observe appear to approximately follow a normal distribution. This is also the case in astronomy (example: the near-infrared $K$-band distribution of globular cluster magnitudes in the Milky Way Galaxy). There’s even a mathematical reason for that. We will later see that (most) averages of random variables are approximately normal. This is known as the *central limit*

Figure 4.8: The normal distribution was featured prominently on Deutsche Mark bills.

*theorem*, the second fundamental 'law' in statistics (the first being the law of large numbers). Many quantities are sums or averages of smaller contributions. That's true for both things that we observe and things that we compute (just look back to the previous chapter). We'll learn more about that later.

A particularly common application domain for the normal distribution are *measurement errors.*

**Example 4.12.** *Consider experiments of measuring the mass $m$ of the (anti-electron) neutrino. The $i$th experiment yields a measurement $M_i$. In many experiments, $M_i$ is computed as a difference two big and similar quantities, none perfectly known. Also some experiments report a negative value. A sensible first approach is to model $M_i \sim \mathcal{N}(m, \sigma^2)$, where $\sigma$ quantifies the precision.*

The normal distribution has many convenient and fascinating properties.

**Proposition 4.13.**

*(i)* *If* $X \sim \mathcal{N}(\mu, \sigma^2)$, *then* $Z = (X - \mu)/\sigma \sim N(0, 1)$.

*(ii)* *If* $Z \sim \mathcal{N}(0, 1)$, *then* $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

*(iii)* *If* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *and* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ *are independent, then*

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

The first property tells us that, by shifting and scaling a normal random variable, we can transform it to a standard normal one. This also works the other way around: by shifting and scaling a standard normal variable, we obtain a normal variable with arbitrary mean and variance (second property). Finally, if we add up to independent normal random variables, the result is again normal. A few more useful facts for computations:

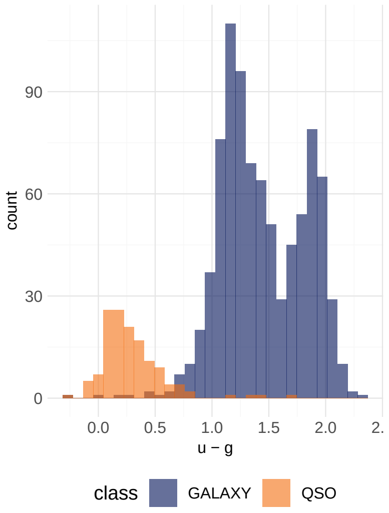
- If $X \sim N(\mu, \sigma^2)$, then it follows from the previous proposition
$$\begin{aligned}\mathbb{P}(a < X < b) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)\\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),\end{aligned}$$
so all the probabilities for $X$ can be computed from $\Phi$ (the CDF of a standard normal).
- Because of symmetry of the distribution, $\phi(x) = \phi(-x)$ and $\Phi(-x) = 1 - \Phi(x)$.

Oddly, $\Phi$ does not have a closed form, meaning that it's impossible to write it down (so don't even try to integrate the density function yourself). In practice, $\Phi$ is therefore always computed using a (highly accurate) numeric approximation. You'll never need to worry about that though, these approximation algorithms are implemented in any reasonable software — even on fancy hand calculators.

Because the normal distribution is so central, it has also been studied most thoroughly. There are many more properties we could list here, but let's not overdo it. As a good rule of thumb: whenever you want to know something about the normal distribution, Google will have the answer.

### 4.2.5 Normal mixtures

*Mixtures* are models that combine several distributions into a new one. Such distributions arise naturally whenever a data set contains multiple sub-populations. Mixtures of normal distributions are especially popular. Let's see this in an example.



On the right you see the histogram of the `u - g` color computed from the data set used in our EDA last chapter. The data consists of $\approx 87\%$ quasars and $\approx 13\%$ galaxies. The two sub-populations (quasars and galaxies) are easy to distinguish from this graph. Now suppose that the `u - g` color of galaxies is $\mathcal{N}(\mu_G, \sigma_G^2)$ of quasars is $\mathcal{N}(\mu_Q, \sigma_Q^2)$

Then the overall color distribution is a weighted mix of the two normal distributions. We write this as

$$0.87\mathcal{N}(\mu_G, \sigma_G^2) + 0.13\mathcal{N}(\mu_Q, \sigma_Q^2)$$

One might argue that the quasars could be split further into two sub-sub-populations. We could model that by a mixture of three normal distributions. Conceptually, that's no different than a mixture of two, but let's stick to the latter for simplicity.

Let's make these models more formal.

**Definition 4.14.** *For two groups with proportions $\alpha, 1-\alpha$, means $\mu_1, \mu_2$ ,and variances $\sigma_1, \sigma_2$,* ***Gaussian mixture distribution*** *function is*

$$F(x) = \alpha\Phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\alpha)\Phi\left(\frac{x-\mu_2}{\sigma_2}\right),$$

*and we write $X \sim \alpha\mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha)\mathcal{N}(\mu_2, \sigma_2^2)$.*

The corresponding density function is

$$f(x) = \alpha\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\alpha)\phi\left(\frac{x-\mu_2}{\sigma_2}\right)$$

The generalization to a mixture of $K$ groups with proportions $\alpha_1, \dots, \alpha_K$ is straightforward.

## 4.3 Multivariate normal distribution

A *multivariate* distribution is the joint distribution of a vector $\boldsymbol{X} \in \mathbb{R}^d$ of random variables. In the last chapter we only considered joint distributions of two variables, but you can probably figure out yourself how to adapt the definitions to $d$ variables. The normal distribution also has a generalization to this case.

**Definition 4.15.** *A random vector $\boldsymbol{X}$ is said to have* ***multivariate normal (Gaussian)*** *distribution with*

- *mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ ,*
- *covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$,*

*written $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its joint density is*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

You can check that the joint density of the multivariate normal simplifies to the density of the univariate normal when $d = 1$. The interpretation of the parameters is the same. The mean vector $\boldsymbol{\mu}$ shifts location, the covariance matrix determines the spread in every direction. Of course, the covariance matrix also contains information about the dependence between the components of $\boldsymbol{X}$.

**Theorem 4.16.** *Suppose* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then it holds:*

*(i)* $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ *(i.e.,* $\mathbb{E}[X_k] = \mu_k$ *for all* $k$*),*

*(ii)* $\mathbb{C}\mathrm{ov}[X_j, X_k] = \Sigma_{j,k}$ *for all* $j, k$*,*

*(iii)* $\boldsymbol{a}^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$ *for all* $\boldsymbol{a} \in \mathbb{R}^d$*,*

*(iv) If* $\mathbb{C}\mathrm{ov}[X_j, X_k] = 0$*, then* $X_j$ *and* $X_k$ *are independent.*

The first two properties are unsurprising. The third one generalizes the fact that the sum of two independent normals is again normal. It states that any linear combination of components of a multivariate normal vector is again normal. The statement from earlier is recovered for $d = 2$, $\boldsymbol{a} = (1, 1)$, and $\mathbb{C}\mathrm{ov}[X_1, X_2] = 0$. But why does $\mathbb{C}\mathrm{ov}[X_1, X_2] = 0$ mean that the variables are independent? For the multivariate normal distribution, variables are independent if and only if they are uncorrelated (fourth property). Note that this is a specific feature of the normal distribution. For most other distributions, zero correlation does *not* imply independence.

## 4.4 Regression models

A statistical model can consist of more than just a distribution. We shall briefly discuss such a model to give you a taste of what's coming in the second part of the course. A *regression model* is a model concerning several variables. In particular, we are interested how one of them, say $Y$, relates to some others, say $\boldsymbol{X}$.

As an example, let's ask how does brightness ($Y$) of a galaxy depend on its mass ($X$)? A simple model would be

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\beta_0, \beta_1$ are model parameters and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This model can be written equivalently as follows: the *conditional distribution* of $Y$ given $X = x$ is

$$\mathcal{N}(\mu(x), \sigma^2), \qquad \text{where } \mu(x) = \beta_0 + \beta_1 x.$$

More generally, if $\boldsymbol{X} \in \mathbb{R}^d$, a *linear regression model* is formulated as

$$Y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{X} + \epsilon, \qquad \mathbb{E}[\epsilon \mid X] = 0.$$

The model parameters $\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d$ are considered unknown. A further generalization are *generalized linear models*: Let $F_\theta$ be a distribution function with parameter $\theta \in \Theta$, then

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \;\sim\; F_\theta, \qquad \text{where } \theta = \theta(\boldsymbol{X}) = g(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{X}),$$

where $g\colon \mathbb{R} \to \Theta$ is a fixed, known function. Also here, $\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d$ are the unknown model parameters.

Generally speaking, regression models are models for conditional distributions. They are immensely important in many scientific fields. Common uses cases are:

- accounting for measurement errors,
- augmenting physical laws with randomness/uncertainty,
- prediction.

We will talk about such models in more detail in a few weeks.

# 5
# Parameter estimation

In the previous chapter we learned about common families of statistical distributions. These families are characterized by one or more parameters. In this chapter, we learn how to find parameter values that best fit observed data.

## 5.1 Motivation: distributions in times of Corona

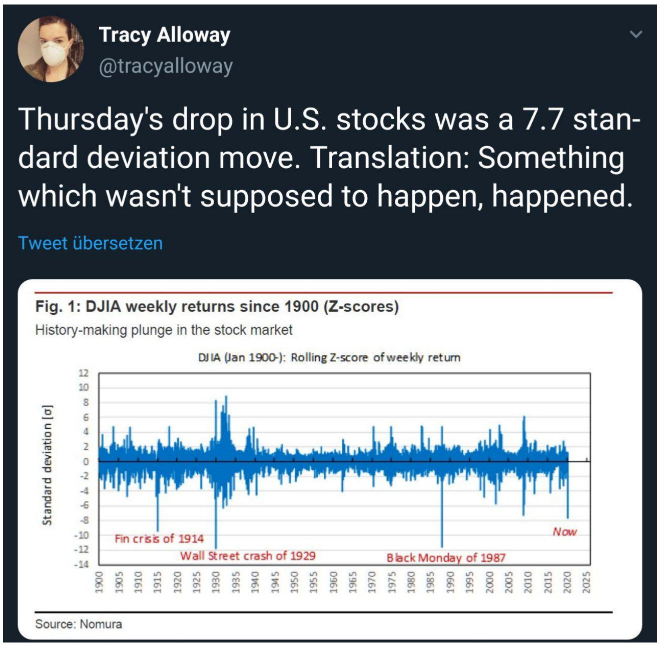
When Corona hit western Europe (March 9–13, 2020), the financial markets plummeted. The tweet in Figure 5.1 shows data for the Dow Jones Industrial Average, the most important stock index for the US industry. The observed quantities are weekly index returns (the percentage change of the index from one week to another). Judging from recent losses, the crash is just as bad as the 2008 financial crisis.

The tweet in Figure 5.1 alerts us that the Dow Jones made a '7.7 standard deviation move': the losses that week were 7.7 times as large as the sample standard deviation. Using the 'standard deviation scale' implicitly assumes a normal distribution for the returns. Under the normal distribution, the probability of such a move is $1 - \Phi(7.7) < 10^{-14}$. Put differently, we expect to see such an event every $10^{14}$ weeks, or $2 \cdot 10^{12}$ years. Even worse: in the last 120 years, we already encountered four similar events. Man, how unlucky are we!

A better explanation than 'bad luck' is 'bad model'. Extreme events like market crashes aren't normal — neither in the colloquial nor statistical sense. In fact, one can prove mathematically that extreme events follow a Pareto distribution.[1] To compute a crash probability under the Pareto$(\xi, \alpha)$ model, we need to specify the parameters. The parameter $\xi$ we can choose (only look at losses larger than $\xi$). But we also need to know the shape parameter $\alpha$. In this chapter, we learn how to find the parameter that fits the data best.

We shall compute a more realistic probability later. For now, take this story as a warning: statistical models need to be chosen wisely. That's why you should have an idea which model is suitable for which kind of problem.

[1] This is the key result of an area of statistics called *extreme value theory.* The theory says that events exceeding a large threshold approximately follow a *generalized Pareto distribution.* It is well established that the generalized model simplifies to the usual Pareto model for financial returns.



Tracy Alloway
@tracyalloway

Thursday's drop in U.S. stocks was a 7.7 standard deviation move. Translation: Something which wasn't supposed to happen, happened.

Tweet übersetzen

Fig. 1: DJIA weekly returns since 1900 (Z-scores)

History-making plunge in the stock market

Source: Nomura

Figure 5.1: Random tweet from March 14, 2020.

## 5.2 A general framework

### 5.2.1 Parametric statistical models

We can write the models from the last chapter in a more generic form. The key commonality is that we specify a (possibly conditional) PDF or PMF $f_\theta$ with parameter $\theta$. A *parametric statistical model* $\mathcal{F}$ is a collection of such functions:

$$\mathcal{F} = \left\{ f_\theta \colon \theta \in \Theta \right\}$$

**Example 5.1.** *The width of lines in electromagnetic spectra approximately follows a $\mathcal{N}(\mu, \sigma^2)$ distribution. Then $f_\theta$ is the density of a $\mathcal{N}(\mu, \sigma^2)$ random variable, where $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \infty)$.*

A statistical model is therefore a collection of many possible distributions, each corresponding to a different value of the parameter $\theta$. Finding the parameter value that best matches the data is called *parameter estimation* or *model fitting*.

### 5.2.2 Estimators and consistency

In Chapter 3, we already touched upon the concept of an *estimator*. Let's recall some important facts. In everything that follows, we assume that the data $X_1, \dots, X_n$ are *iid* random variables.

**Definition 5.2** (Estimator)**.** *Any quantity that can be expressed as $g(X_1, \dots, X_n)$ for some function $g$ is called an* ***estimator****.*

**Definition 5.3** (Consistency). *Let $\theta$ be an unknown quantity that we are interested in. An estimator $\widehat{\theta}_n$ is called* ***consistent*** *for $\theta^*$ if ‘$\widehat{\theta}_n$ converges to $\theta^*$ in probability:*

$$\widehat{\theta}_n \to_p \theta^*.$$

**Example 5.4.** *Let $\theta^* = \mathbb{E}[X]$. Then the sample average $\widehat{\theta}_n = \bar{X}_n$ is a consistent estimator.*

A statistical model $\mathcal{F}$ is called *correctly specified*, if the (unknown) true density (or PMF) $f^*$ is contained in $\mathcal{F}$, i.e., $f^* \in \mathcal{F}$. If the model is correctly specified, it is possible to construct estimators $\widehat{\theta}_n$ that are consistent for the parameter $\theta^*$. That is, we can learn the true distribution from the data if there are sufficiently many.[2] The subscript $n$ indicates that the estimator is different for any sample size $n$, but is often dropped for convenience.

### 5.2.3 Bias, variance, and MSE

Recall that estimators are functions of random variables. As a consequence, an estimator is itself a random variable. It thus makes sense to speak about the expectation and variance of an estimator.

The expectation $\mathbb{E}[\widehat{\theta}]$ is related to a concept we call *bias*.

**Definition 5.5** (Bias). *The* ***bias*** *of an estimator $\widehat{\theta}$ is defined as*

$$\text{bias}[\widehat{\theta}] = \mathbb{E}[\widehat{\theta}] - \theta^*.$$

Optimally, we would like to have $\mathbb{E}[\widehat{\theta}] = \theta^*$: on average, the estimator $\widehat{\theta}$ is equal to the true parameter. In this case, $\text{bias}[\widehat{\theta}] = 0$ and we call the estimator *unbiased*.

**Example 5.6.** *Let $\theta^* = \mathbb{E}[X]$ and $\widehat{\theta}_n = \bar{X}_n$. Then*

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_i\right] = \frac{n}{n}\mathbb{E}[X_1] = \mathbb{E}[X] = \theta^*.$$

*Hence, the sample average $\bar{X}_n$ is an unbiased estimator for the true mean $\mathbb{E}[X]$.*

Although unbiasedness is desirable, it is not necessary for consistency. For example, one can show that the sample variance is biased:

$$\mathbb{E}[S_n^2] = \frac{n-1}{n}\mathbb{V}[X] \neq \mathbb{V}[X].$$

[2] This seems like dark magic to some people. To reflect that, statisticians put a magician's hat on the parameter $\theta$ and denote the estimator by $\widehat{\theta}_n$. (That's my interpretation at least.)

In fact, there are problems for which it is impossible to construct an estimator that is both consistent and unbiased (but that topic is too advaced for now). Normally, we're satisfied with asymptotically unbiased estimators. That is, estimators $\widehat{\theta}$ for which $\mathbb{E}[\widehat{\theta}] \to \theta^*$ as $n \to \infty$. Indeed, for the sample variance it holds $\lim_{n\to\infty} \mathbb{E}[S_n^2] \to \mathbb{V}[X]$.[3]

Unbiasedness is not sufficient for consistency either. The estimator $\widehat{\theta}$ is a random variable, but the limiting value in Definition 5.3 is not. In particular, the limiting value has zero variance. Hence, we would also like the variance of the estimator to vanish asymptotically, i.e.,

$$\mathbb{V}[\widehat{\theta}] \to 0, \qquad \text{as } n \to \infty.$$

Any asymptotically unbiased estimator with vanishing variance is consistent.

**Theorem 5.7.** *Let $Y_1, Y_2, \ldots$ be a sequence of random variables. If $\mathbb{E}[Y_n] \to y$ for some $y \in \mathbb{R}$ and $\mathbb{V}[Y_n] \to 0$, then $Y_n \to_p y$.*

**Example 5.8.** *We already know that the sample average is unbiased: $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X]$ for all $n$. Furthermore, we can compute $\mathbb{V}[\bar{X}_n] = \frac{1}{n}\mathbb{V}[X] \to 0$. Hence, $\bar{X}_n \to_p \mathbb{E}[X]$. (This is one way to prove the law of large numbers.)*

Instead of considering bias and variance separately, we can also look at a single measure for the quality of an estimator.

**Definition 5.9** (Mean squared error, MSE)**.** *The **mean squared error** of an estimator $\widehat{\theta}$ is defined as*

$$\text{MSE} = \mathbb{E}[(\widehat{\theta} - \theta^*)^2].$$

The squared error $(\widehat{\theta} - \theta^*)^2$ is a measure for the accuracy of $\widehat{\theta}$. The MSE tells us how accurate we are *on average.*[4] It turns out that the MSE is just a composition of bias and variance.

**Theorem 5.10.** *It holds*

$$E[(\widehat{\theta} - \theta^*)^2] = \text{bias}[\widehat{\theta}]^2 + \mathbb{V}[\widehat{\theta}].$$

*Proof.* Note that

$$\widehat{\theta} - \theta^* = \widehat{\theta} - \underbrace{\mathbb{E}[\widehat{\theta}] + \mathbb{E}[\widehat{\theta}]}_{=0} - \theta^* = (\widehat{\theta} - \mathbb{E}[\widehat{\theta}]) + (\mathbb{E}[\widehat{\theta}] - \theta^*).$$

[3] The expectation is not a random variable, so this convergence is in the usual, non-probabilistic sense.

[4] To preserve units, we may take the root of MSE.

Then using the binomial theorem,

$$\begin{aligned} E[(\widehat{\theta}-\theta^*)^2] &= \mathbb{E}\big[\big((\widehat{\theta}-\mathbb{E}[\widehat{\theta}])+(\mathbb{E}[\widehat{\theta}]-\theta)\big)^2\big] \\ &= \mathbb{E}[(\widehat{\theta}-\mathbb{E}[\widehat{\theta}])^2] + 2\,\underbrace{\mathbb{E}[(\widehat{\theta}-\mathbb{E}[\widehat{\theta}])]}_{=0}\big(\mathbb{E}[\widehat{\theta}]-\theta^*\big) + (\mathbb{E}[\widehat{\theta}]-\theta^*)^2 \\ &= \underbrace{\mathbb{E}[(\widehat{\theta}-\mathbb{E}[\widehat{\theta}])^2]}_{=\mathbb{V}[\widehat{\theta}]} + \underbrace{(\mathbb{E}[\widehat{\theta}]-\theta^*)^2}_{\text{bias}[\widehat{\theta}]^2}. \end{aligned}$$

□

## 5.3 The method of moments

We shall now turn to specific examples of estimators for the parameter of a statistical model. We start with a simple and intuitive method, the *method of moments (MOM)*. The $k$th *moment* of a random variable $X$ is simply $\mathbb{E}[X^k]$. $\mathbb{E}[X]$ is the first moment; $\mathbb{E}[X^2] = \mathbb{V}[X] + \mathbb{E}[X]^2$ is the second moment, and so on.

In the previous chapter, we have seen explicit expressions for the mean and variance of random variables adhering to models $f_\theta$ (exponential, Poisson, etc.). In addition, we know that we can estimate the mean and variance of a random variable consistently by the sample mean $\bar{X}_n$ and sample variance $S_n^2$. We can therefore define the estimated parameter $\widehat{\theta}$ such that theoretical and estimated mean and variance match.

**Example 5.11.** *Assume $X \sim \text{Bernoulli}(p)$. We know that $\mathbb{E}[X] = p$. The MOM estimator is defined as $\widehat{p} = \bar{X}_n$. Consistency of this estimator follows from consistency of $\bar{X}_n$.*

**Example 5.12.** *Assume $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e., $\theta = (\mu, \sigma^2)$. We know that $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$. We define the MOM estimator $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2)$ by*

$$\widehat{\mu} = \bar{X}_n, \qquad \widehat{\sigma}^2 = S_n^2.$$

*Consistency of this estimator follows from consistency of $\bar{X}_n$ and $S_n^2$.*

These examples were a bit boring: the parameters are equal to the mean and variance. It's often more involved, though.

**Example 5.13.** *Assume $X \sim \text{Gamma}(\alpha, \beta)$, i.e., $\theta = (\alpha, \beta)$. We know that $\mathbb{E}[X] = \alpha\beta$ and $\mathbb{V}[X] = \alpha\beta^2$. We define the MOM estimator $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$ by solving the system of equations*

$$\widehat{\alpha}\widehat{\beta} = \bar{X}_n, \qquad \widehat{\alpha}\widehat{\beta}^2 = S_n^2.$$

*You can verify that the solution is*

$$\widehat{\alpha} = \frac{\bar{X}_n^2}{S_n^2}, \qquad \widehat{\beta} = \frac{S_n^2}{\bar{X}_n}.$$

*Consistency of $\widehat{\theta}$ again follows from consistency of $\bar{X}_n$ and $S_n^2$.*

In the examples above, $\theta$ was at most two-dimensional. Hence, we used two moments to fit the parameters to data. In general, if theta is $p$-dimensional, we need to match $p$ theoretical and sample moments to another. The $k$th sample moment is just $\frac{1}{n}\sum_{i=1}^n X_i^k$, which convergences to $\mathbb{E}[X^k]$ by the law of large numbers. This ensures consistency of the MOM estimator also in the more general case.

## 5.4 Maximum likelihood estimation

The method of moments was especially popular in times where data analysis was performed with nothing but pencil and paper. It presupposes that the theoretical moments of a distribution are known and have a simple form. We shall now discuss a more general method: *maximum likelihood estimation.*

### 5.4.1 Motivation

Suppose we have decided on a statistical model $\mathcal{F} = \{f_\theta \colon \theta \in \Theta\}$ and want to construct an estimator for the parameter $\theta$. Consider the following question:

> *Given observed data $X_1, \dots, X_n$, which value for the true parameter $\theta^*$ is most likely?*

It seems natural to use the answer to this question as an estimator $\widehat{\theta}$. There is a subtle conceptual issue, however. The question asks for a probabilistic assessment of a fixed (not random!) parameter $\theta^*$. But when there's no randomness, probabilities become trivial. The true parameter always takes the same (unknown) value, every other value has probability zero.

To fix this, we would need to think of the parameter $\theta$ as a random variable. This requires a different conceptual framework, called *Bayesian* paradigm. We will touch on this later in the course. To resolve the paradox in the current framework, we must ask a different question. Let's reverse the one above:

> *Given a parameter value $\theta$, how likely is it that we observe the data $X_1, \dots, X_n$?*

The data are random variables, so it's adequate to assess them probabilistically. We then define an estimator $\widehat{\theta}$ as the value $\theta$ under which the observed data are most likely: we maximize the likelihood of the data given the parameter.

### 5.4.2 The likelihood function

We first need to define what we mean by *likelihood* of the data. Suppose a random variable $X$ has PDF/PMF $f$. The value $f(x)$ measures how likely it is that the random variable $X$ takes the value $x$. It will be unnecessary to distinguish between continuous and discrete variables in what follows, so let's just say 'density' when we mean either PDF or PMF.

When we say 'likelihood of the data', we therefore mean 'joint density of all observations $X_1, \dots, X_n$'. Now assume that the data are *iid* with $X_i \sim f_{\theta^*}$. So what's the joint density of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$? Take a moment to think about this.

Because the components of $\boldsymbol{X}$ are independent, the joint density is just the product of marginal densities (see Theorem 2.47). Hence, we define the likelihood function as

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

$L$ is a function of the model parameter $\theta$ and tells us how likely it is to observe the data $X_1, \dots, X_n$ if the true model were $f_\theta$.

### 5.4.3 The maximum likelihood estimator

The *maximum-likelihood estimator (MLE)* $\widehat{\theta}$ is defined as the value $\theta^*$ that maximizes the function $L$. In mathematical notation:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(X_i). \tag{5.1}$$

The method is extremely general. Everything we need to compute the estimator is knowledge of the density function. Another formulation of the MLE will be more convenient. Define the *log-likelihood* $\ell(\theta) = \ln L(\theta)$ and the MLE as

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta^*) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f_\theta(X_i) \tag{5.2}$$

Convince yourself that the two defintions (5.1) and (5.2) are in fact equivalent (the logarithm is a strictly increasing function).

So what's the advantage of taking the logarithm? Recall that, to find the maximum of a function, we equate the first derivative of the function to zero. The derivative of a product of $n$ terms is unwieldy (think applying the product rule dozens of times). The derivative of a sum is just the sum of derivatives. This simplifies computations a lot. If the maximization problem is solved numerically, a sum also tends to be more stable than a product, but that's a topic for another course.

**Remark 5.1.** *Suppose we are not actually interest in the parameter $\theta^*$, but some transformation $\tau^* = g(\theta^*)$ of it. If $\widehat{\theta}$ is the MLE for $\theta^*$, then $g(\widehat{\theta})$ is the MLE for $\tau^*$. This property is called* ***equivariance*** *of the MLE.*

### 5.4.4 Computing the MLE

In many cases, the MLE can be computed theoretically. The strategy is always the same:

Step 1. Compute the log-likelihood function $\ell(\theta)$.

Step 2. Compute the first derivative with respect to all components of $\theta$ ($\theta$ may be multidimensional).

Step 3. Equate the derivatives to zero. For $k$-dimensional $\theta$, we get the system of equations

$$\frac{\partial \ell(\theta)}{\partial \theta_1} = 0, \quad \dots, \quad \frac{\partial \ell(\theta)}{\partial \theta_k} = 0. \tag{5.3}$$

Step 4. Define the MLE $\widehat{\theta}$ as the value of $\theta$ that solves (5.3).

**Example 5.14.** *Suppose $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Bernoulli}(p)$. Then $\theta = p$, and $f_\theta(x) = p^x(1-p)^{1-x}$. We have*

$$\begin{aligned}\ell(p) &= \sum_{i=1}^n \ln f_\theta(X_i) \\ &= \sum_{i=1}^n \ln\Big(p^{X_i}(1-p)^{1-X_i}\Big) \\ &= \sum_{i=1}^n X_i \ln p + \sum_{i=1}^n (1-X_i)\ln(1-p).\end{aligned}$$

*Taking the derivative with respect to the parameter $p$ yields*

$$\frac{d\ell(p)}{dp} = \frac{1}{p}\sum_{i=1}^n X_i - \frac{1}{1-p}\sum_{i=1}^n (1-X_i) \overset{!}{=} 0.$$

*To solve the above equation, multiply both sides with $p(1-p)$, which gives*

$$\begin{aligned}
&(1-p)\sum_{i=1}^n X_i - p\sum_{i=1}^n (1-X_i) = 0\\
\Leftrightarrow\quad &\sum_{i=1}^n X_i - p\sum_{i=1}^n X_i - p\sum_{i=1}^n 1 + p\sum_{i=1}^n X_i = 0\\
\Leftrightarrow\quad &\sum_{i=1}^n X_i - p\sum_{i=1}^n 1 = 0\\
\Leftrightarrow\quad &\sum_{i=1}^n X_i - pn = 0\\
\Leftrightarrow\quad &p = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n.
\end{aligned}$$

*Hence the MLE and MOM estimator coincide: $\widehat{\theta} = \widehat{p} = \bar{X}_n$.*

**Example 5.15.** *Consider the regression model*

$$Y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{X} + \epsilon,$$

*with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\boldsymbol{X}$. Recall that this is equivalent to saying $Y \mid \boldsymbol{X} = \boldsymbol{x} = \mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$. Let's assume for simplicity that the variance $\sigma^2$ is known. The statistical model for $Y$ is then $\mathcal{F} = \{f_\theta \colon \theta \in \Theta\}$, where $f_\theta$ is the density of a $\mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{X}, \sigma^2)$ random variable and $\theta = (\beta_0, \boldsymbol{\beta})$. Note that*

$$2\ell(\beta_0, \boldsymbol{\beta}) = 2\sum_{i=1}^n \ln f_\theta(Y_i) = -n\ln(2\pi\sigma^2) - \sum_{i=1}^n (Y_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{X}_i)^2/\sigma^2.$$

*Maximizing this expression is equivalent to minimizing*

$$\sum_{i=1}^n (Y_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{X}_i)^2.$$

*That's why the MLE under a Gaussian likelihood is also referred to as* (ordinary) least-squares estimator *or OLS for short. The OLS estimator can be computed theoretically, but let's reserve that for another time.*

When the statistical model is too complicated, it may be hard to derive the MLE theoretically. If that's the case (or you just feel lazy), the MLE can be computed using numerical optimization algorithms (e.g., `scipy.optimize`). Theoretical expressions are much faster to compute, however, so they are still useful in practice (and for exam problems).

Let's get back to our Corona crash example from the beginning. We can also compute the MLE for the Pareto distribution.

**Example 5.16.** *Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Pareto}(\xi, \alpha)$ and $\xi$ is known. Then $\theta = \alpha$ and*

$$f_\alpha(x) = \frac{\alpha \xi^\alpha}{x^{\alpha+1}}, \qquad \textit{for } x > \xi.$$

*The log-likelihood is*

$$\ell(\alpha) = n \ln(\alpha) + n\alpha \ln(\xi) - (\alpha + 1) \sum_{i=1}^n \ln(X_i).$$

*Taking the derivative we get*

$$\frac{\partial \ell(\alpha)}{\partial \alpha} = \frac{n}{\alpha} + n \ln(\xi) - \sum_{i=1}^n \ln(X_i) \overset{!}{=} 0.$$

*Solving for $\alpha$ gives the MLE $\widehat{\alpha} = n / \sum_{i=1}^n \ln(X_i/\xi)$.*

Now we can fit the parameter and compute a probability for a crash as extreme as last week. I downloaded data for Dow Jones returns for the last 35 years from yahoo finance[5]. Let's set $\xi = 0.05$: every week with a loss larger than 5% is considered extreme. By this definition, 39 of the weeks ($\approx 2.1\%$ of the data) were larger than than $\xi$ and the MLE gives $\widehat{\alpha} \approx 2.8$. In the tweet, the weekly loss was a whopping 17%. The probability of an event at least as extreme is

$$0.021\mathbb{P}(X > 0.17) = 0.021\big(1 - F_{\xi, \widehat{\alpha}}(0.17)) \approx 0.0007.$$

Thus, we expect a crash like this every $1/0.0007 \approx 1400$ weeks or every $1/(52 \times 0.0007) \approx 27$ years. While this is still unlikely, the event is orders of magnitude more probable as you would expect under a normal distribution. It also aligns well with what we observed over the last 120 years. Seems like we're not that unlucky after all.

### 5.4.5 Consistency

The MLE enjoys several nice theoretical properties. In some sense, it is even the best possible estimator (think: no other consistent estimator can have a smaller MSE), but that's beyond the current scope. For now, we shall content ourselves with the fact that the MLE is consistent.

**Theorem 5.17.** *Suppose $X_1, \dots, X_n \overset{iid}{\sim} f_{\theta^*}$ for some $f_{\theta^*} \in \mathcal{F}$. Under some regularity conditions[6], then the MLE $\widehat{\theta}$ exists[7]and is consistent.*

[5] https://finance.yahoo.com/

[7] The 'regularity conditions' are mostly unproblematic. Their main purpose is to exclude pathological cases; for example, densities $f_\theta$ that aren't continuous in $\theta$.

[7] The 'exists' refers to the fact that the likelihood actually has a maximum.

**Remark 5.2.** *We are not going to prove the above theorem. But in case you're interested, here's the idea: Instead of maximizing $\ell(\theta)$, we could just as well maximize $\ell(\theta)/n$ (scaling doesn't change the maximal point). By the law of large numbers, it holds*

$$\ell(\theta)/n = \frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(X_i) \to_p \mathbb{E}_{\theta^*}[\ln f_\theta(X_i)],$$

*where the expectation on the right is computed under the true model with parameter $\theta^*$ and density $f_{\theta^*}$. That is,*

$$\mathbb{E}_{\theta^*}[\ln f_\theta(X_i)] = \int \ln f_\theta(x) f_{\theta^*}(x)dx,$$

*which is called the* cross-entropy *between two densities $f_\theta$ and $f_{\theta^*}$. Hence, $\widehat{\theta}$ is a value that maximizes cross-entropy asymptotically (as $n \to \infty$). Finally, one can show that the cross-entropy is maximized by the true parameter value $\theta^*$.*

## 5.5 Checking for misspecification and model fit

There is one condition in Theorem 5.17 that you should worry about. The theorem assumes that the model is correctly specified. If it is not, the estimated model will not converge to the true one. As we have seen from the Corona crash example, this can have severe consequences. If the misspecification is less extreme, the MLE may still be useful, however.

Philosophically speaking, it's unreasonable to assume that the true distribution really belongs to a specific model class $\mathcal{F}$. There is an apt quote from one of the greatest figures in 20th century statistics:

> *All models are wrong, but some are useful.* — *George E. Box*

In that sense, it's useless to worry about the model being incorrect. But we should certainly think about whether a statistical model is useful. If what we observe doesn't align with the properties of the model, it's probably not a useful one.

So how do we check? There's one simple tool, called quantile-quantile plot or just *QQ-plot*. Suppose $\widehat{\theta}$ is the estimated parameter and $F_{\widehat{\theta}}$ the associated distribution function. The QQ-plot is simply a graph with the theoretical quantiles $F_{\widehat{\theta}}^{-1}(p)$ on the $x$-axis and the empirical quantiles $F_n^{-1}(p)$ on the $y$-axis. If the model fit is good, all points should lie on the main diagonal $x = y$.

A QQ-plot for the Dow Jones data is shown in Fig. 5.2. The QQ-plot for the normal distribution is indicated by black dots (parameters were estimated by MOM). The points deviate a lot from the main diagonal. For large losses, the sample quantiles are much larger than the theoretical quantiles. Hence, the normal distribuition is a poor model for large losses. The orange triangles are

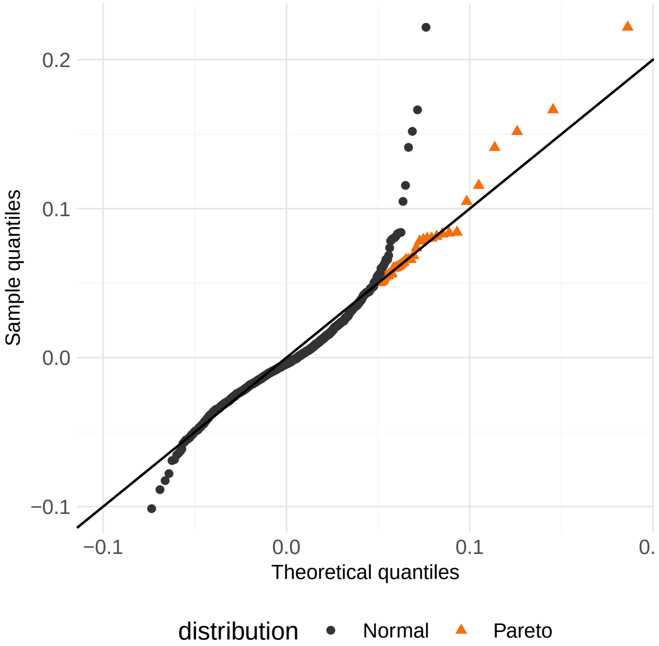


Figure 5.2: QQ-plots for the normal and Pareto distributions in the Corona crash. Data are weekly losses on the Dow Jones Index. Since we cut of the Pareto at $\xi = 0.05$, only values above this threshold are shown.

the QQ-pairs for the Pareto distribution. They are generally quite close to the diagonal, so the Pareto model seems to provide a good fit.

## 5.6 Chi-square fitting

Let's be clear: use MLE or MOM whenever possible. But there's one more method we need to speak about. Early in the last century, astronomers came up with a method called $\chi^2$ fitting. Although antiquated and sub-optimal in many ways, it is still applied all over the field. Because it's used so widely, you should have at least seen it.

The procedure in a nutshell:

1. Compute a histogram of the data. Let $\xi_1, \dots, \xi_K$, be the centers of the bins and $\widehat{h}(\xi_k)$ the estimated density.

2. Because in general $\widehat{h}(\xi_k) \neq f(\xi_k)$, the histogram makes a random estimation error $\epsilon_k = \widehat{h}(\xi_k) - f(\xi_k)$ for which we assume $\epsilon_k \sim \mathcal{N}(0, \sigma_k^2)$. The variances $\sigma_k^2$ may be different for each bin $k$ and can be estimated.

3. Now find the parameter $\theta$ that minimizes the $\chi^2$ criterion

$$\chi^2(\theta) = \sum_{k=1}^{K} \frac{\left(\widehat{h}(\xi_k) - f_\theta(\xi_k)\right)^2}{\widehat{\sigma}_k^2}.$$

There are numerous issues with the above procedure: binning causes bias, it's unclear how to choose number and location of bins, error variances $\sigma_k^2$ need to be estimated, ... There are equally many modifications of the above: taking logarithms of $X_i$ first, taking logarithms of $\widehat{h}$ and $f_\theta$ when computing the criterion, how to compute $\sigma_k$, etc. — many things that try to fix issues caused by binning.

In the old days, data were often recorded or shared in the form of binned counts. Then the only option is to go through these chores. Gladly, this is rare in modern times and we can just use MLE or MOM.

# 6 Uncertainty quantification

By now, we have seen many estimators $\widehat{\theta}$ of some parameter $\theta^*$:

- The sample average $\bar{X}_n$, variance $S_n^2$, correlation $R_n$ as estimators of the parameters $\mathbb{E}[X]$, $\mathbb{V}[X]$, and $\rho(X, Y)$.
- The empirical distribution function $F_n(x)$ and quantile $F_n^{-1}(p)$ as estimators of a CDF $F(x)$ and quantile $F^{-1}(p)$.
- The histogram $\widehat{h}(x)$ as an estimator for the PDF/PMF $f(x)$.
- The MOM and maximum-likelihood method as estimators for the parameter of a parametric statistical model.

All but MOM and MLE are also called *nonparametric* estimators, because we do not need to specify a parametric model for them to work.

So if we have computed an estimator $\widehat{\theta}$, can we say that $\theta^* = \widehat{\theta}$? Of course, not! The estimator $\widehat{\theta}$ is a random variable, but $\theta^*$ is not. Every time we compute an estimator $\widehat{\theta}$ we will make an estimation error $\widehat{\theta} - \theta^* \neq 0$. We cannot know the exact error without knowing $\theta^*$. If the estimator is consistent, we know that it converges for infinitely many observations. But on finite samples, there there is some *uncertainty* how close we are to the truth.

In this chapter, we learn how to quantify this uncertainty probabilistically.

## 6.1 The central limit theorem

The estimation error $\widehat{\theta} - \theta^*$ is a random variable, so it has a distribution. The main question is therefore what this distribution is. In special cases, the distribution can be derived exactly. But more commonly, we need to rely on asymptotic approximations.

Consistency (related to the law of large numbers) tells us that the distribution converges to a point mass in the limit. But that's not helpful to quantify uncertainty. There is another important limit theorem, the *central limit theorem (CLT)*. We first need a definition for convergence of distributions.

**Definition 6.1** (Convergence in distribution)**.** *Let $Y_n$ be a sequence of random variables and $Y$ be another random variable. Denote the CDF of $Y_n$ by $F_n$ and the one of $Y$ by $F$. Then we say that $Y_n$* ***converges in distribution*** *to $Y$ or*

$$Y_n \to_d Y,$$

*if for all $y \in \mathbb{R}$ where $F$ is continuous,*

$$F_n(y) \to F(y), \qquad \text{as } n \to \infty.$$

The restriction to continuity points is necessary to allow for non-continuous distributions. There is a relationship between convergence in probability and distribution. If $Y_n \to_p Y$, then also $Y_n \to_d Y$. The reverse in not true in general. But if the limiting random variable $Y$ is constant, i.e., $\mathbb{P}(Y = c) = 1$ for some $c \in \mathbb{R}$, then $Y_n \to_d c$ also implies $Y_n \to_p c$.

For uncertainty quantification, we are really interested in situations where the limit is genuinely random. The CLT provides just that.

**Theorem 6.2** (Central limit theorem, CLT)**.** *Let $Y_1, \dots, Y_n$ be* iid *with mean $\mathbb{E}[Y_1] = \mu$ and variance $\mathbb{V}[Y_1] = \sigma^2$ and define $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then*

$$\frac{\bar{Y}_n - \mathbb{E}[\bar{Y}_n]}{\sqrt{\mathbb{V}[\bar{Y}_n]}} = \frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma} \to_d \mathcal{N}(0,1),$$

*and we say that the sequence $\bar{Y}_n$ is* ***asymptotically normal****.*

**Remark 6.1.** *The statement of the theorem uses the common short notation $(Y_n - \mu)/\sigma \to_d \mathcal{N}(0,1)$. The long form is "there is a random variable $Y \sim \mathcal{N}(0,1)$ such that $(Y_n - \mu)/\sigma \to_d Y$." An alternative way to write it is*

$$\sqrt{n}(\bar{Y}_n - \mu) \to_d \mathcal{N}(0, \sigma^2).$$

Remember when we said that (most) averages behave like a Gaussian random variable? The CLT is the mathematically precise formulation of this fact. The interpretation is that, for large enough $n$, the sample average $\bar{Y}_n$ behaves approximately[1] like a $\mathcal{N}(\mu, \sigma^2/n)$ random variable. This also explains why the Gaussian distribution is found everywhere in nature. It is the natural model when many independent factors contribute to an outcome.

As $n \to \infty$, the variance $\mathbb{V}[\bar{Y}_n] = \sigma^2/n$ vanishes. Hence, in a probabilistic sense, the difference $\bar{Y}_n - \mu$ gets closer to 0 (that's the law of large numbers). The scaling with $\sqrt{n}$ allows us to obtain a non-trivial limit. You can think of it this way: multiplying a random variable by $\sqrt{n}$ blows up its variance. The rate $\sqrt{n}$ strikes just the right balance: $\mathbb{V}[\sqrt{n}\bar{Y}_n] = (\sqrt{n})^2\mathbb{V}[\bar{Y}_n] = \sigma^2 \in (0, \infty)$.

[1] "Approximately behaves like" refers to probability statements: probability statements concerning $\bar{Y}_n$ are approximated by probability statements concerning $\mathcal{N}(\mu, \sigma^2/n)$.

The central limit theorem is quite remarkable. The only assumptions are that the sequence is *iid* with finite variance. It is called *central* because it plays such a central role in probability and statistics. The name was first used by George Pólya[2] in 1920 (in German, "Zentraler Grenzwertsatz"), but the idea is older and many other famous mathematicians contributed, including Laplace, Cauchy, Bessel, Poisson (all part-time astronomers!).

As a side note, let me mention that there are several generalizations of the CLT. The multivariate CLT states that an average of random vectors behaves like a multivariate normal random variable. Furthermore, the variables do not have to be *iid*. For example, we can allow their distribution to change with $n$ or for (weak) dependence between observations.

There is a joke about statisticians taking averages all day and, in a sense, this is true. Many estimators we have seen so far can be expressed as averages (or functions of averages). We shall see that even when they don't, they can often be approximated by a suitable average. The CLT tells us that *all* these estimators behave like a Gaussian when properly scaled. How nice is that?

## 6.2 Asymptotic normality of estimators

So how is this useful for uncertainty quantification? If $\widehat{\theta} - \theta^* \approx \mathcal{N}(0, \sigma^2/n)$, we can compute an (approximate) probability that $\widehat{\theta}$ is within some distance of $\theta^*$. In particular, for any $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}(|\widehat{\theta} - \theta| < \epsilon) &= \mathbb{P}\left(\left|\frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma}\right| < \frac{\sqrt{n}\epsilon}{\sigma}\right) \\ &= \mathbb{P}\left(-\frac{\sqrt{n}\epsilon}{\sigma} < \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma} < \frac{\sqrt{n}\epsilon}{\sigma}\right) \\ \text{(CLT)} \qquad &\approx \Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right) - \Phi\left(-\frac{\sqrt{n}\epsilon}{\sigma}\right). \end{aligned}$$

If the variance $\sigma^2$ is known, we can actually compute this number. Often it is unknown, but can be estimated.

As $n \to \infty$, the probability above approaches 1: the more data we have, the more certain we are that $\widehat{\theta}$ is close to $\theta^*$. Note that the standard deviation of $\widehat{\theta}$ is approximately $\mathrm{se}[\widehat{\theta}] = \sigma/\sqrt{n}$. This term is called *standard error* and often used as a measure of uncertainty. As $n \to \infty$, the standard error goes to zero, which reflects our increase in certainty.

The CLT applies directly to the sample average $\widehat{\theta} = \bar{X}_n$. This is an estimator for the parameter $\theta^* = \mathbb{E}[X]$. Let's revisit some of the other examples from the beginning. As always, we assume that the data are *iid* random variables $X_1, \dots, X_n \overset{iid}{\sim} F$.

[2] You might have been tortured by his 'urn' in high school.

**Example 6.3.** *The empirical distribution function is defined as*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x), \quad x \in \mathbb{R}.$$

*Convince yourself that*

$$\mathbb{E}[\widehat{F}_n(x)] = F(x), \qquad \mathbb{V}[\widehat{F}_n(x)] = \frac{F(x)\big(1 - F(x)\big)}{n}.$$

*(Hint: what's the distribution of $\mathbb{1}(X_i \leq x)$?) By the CLT,*

$$\frac{\sqrt{n}(\widehat{F}_n(x) - F(x))}{\sqrt{F(x)\big(1 - F(x)\big)}} \to_d \mathcal{N}(0, 1).$$

*The variance $\sigma^2 = F(x)\big(1 - F(x)\big)$ is not known, because it involves the unknown distribution $F$. However, we can estimate it by $\widehat{\sigma}^2 = \widehat{F}_n(x)\big(1 - \widehat{F}_n(x)\big)$.*

**Example 6.4.** *Suppose for simplicity that $F$ is continuous. The histogram for $x \in (x_{k-1}, x_k]$ is defined as*

$$\widehat{h}_n(x) = \frac{1}{n(x_k - x_{k-1})}\sum_{i=1}^n \mathbb{1}(x_{k-1} < X_i \leq x_k).$$

*Setting $p_k = F(x_k) - F(x_{k-1})$, we get*

$$\mathbb{E}[\widehat{h}_n(x)] = \frac{p_k}{x_k - x_{k-1}}, \qquad \mathbb{V}[\widehat{h}_n(x)] = \frac{p_k(1 - p_k)}{n(x_k - x_{k-1})^2}$$

*and therefore*

$$\widehat{h}_n(x) \approx \mathcal{N}\left(\frac{p_k}{x_k - x_{k-1}}, \frac{p_k(1 - p_k)}{n(x_k - x_{k-1})^2}\right).$$

*Note that $\mathbb{E}[\widehat{h}_n(x)] \neq f(x)$, so the histogram is biased. (One can check that it is asymptotically unbiased, however.) If we want to compute probabilities like $\mathbb{P}(|\widehat{h}_n(x) - f(x)| < \epsilon)$, we would need to estimated not just the variance of $\widehat{h}_n(x)$, but also its bias. That's beyond the scope of this course, but be aware that biased estimators complicate things.*

## 6.3 Asymptotic normality of the MLE

Because the MLE is so important, it deserves special treatment. We will sometimes write $f_\theta(x)$ as $f(x; \theta)$ to indicate more clearly that $f$ is a function of $\theta$. Our main result in this section is the asymptotic normality of the MLE.

**Theorem 6.5.** *Suppose $X_1, \dots, X_n \overset{iid}{\sim} f_\theta$ for some $f_{\theta^*} \in \mathcal{F}$. Under some regularity conditions, the MLE $\widehat{\theta}$ satisfies*

$$\sqrt{n}(\widehat{\theta} - \theta^*) \to_d \mathcal{N}\big(0, I(\theta^*)^{-1}\big),$$

*where*

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)^2\right].$$

The function $I$ determining the variance is called the *Fisher information.* It is interpreted as the amount of information one observation $X_i$ carries about the unknown parameter $\theta^*$. If $I(\theta)$ is large ($X_i$ provides a lot of information), the variance of the MLE will be small. That makes sense: if the data carry more information, we can be more certain about the estimate.

There is a lot of beautiful theory around the Fisher information that you don't need to worry about. For example, one can prove that for any estimator $\widehat{\theta}$ of $\theta^*$, $\mathbb{V}[\widehat{\theta}] \geq 1/\big(nI(\theta^*)\big)$. Hence, no estimator can have smaller variance than the MLE. Put differently: The MLE is the most efficient way to extract information from the data. One useful fact is the alternative representation[3]

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ln f_\theta(X)}{(\partial\theta)^2}\right].$$

**Example 6.6.** *Recall from Example 5.16 that for $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Pareto}(\xi^*, \alpha^*)$ and known $\xi^*$,*

$$f_\alpha(x) = \frac{\alpha\xi^\alpha}{x^{\alpha+1}}, \qquad \text{for } x > \xi.$$

*The log-density is*

$$f(x;\alpha) = \ln(\alpha) + \alpha\ln(\xi) - (\alpha+1)\ln(x).$$

*Thus,*

$$\frac{\partial f(x;\alpha)}{\partial\alpha} = \frac{1}{\alpha} + \ln(\xi) - \ln(x), \qquad \frac{\partial^2 f(x;\alpha)}{(\partial\alpha)^2} = -\frac{1}{\alpha^2}.$$

*Hence, $I(\alpha) = 1/\alpha^2$ and the MLE satisfies $\widehat{\alpha} - \alpha^* \approx \mathcal{N}\big(0, (\alpha^*)^2/n\big)$.*

**Remark 6.2.** *Theorem 6.5 also generalizes to multi-dimensional parameters $\theta$. In that case, the limit is a multivariate normal distribution and the Fisher information is a matrix.*

Deriving Theorem 6.5 is more involved than the previous cases. Except in special cases, the MLE cannot be expressed as an average. We shall see that we

[3] There's a proof on wikipedia if you don't believe it.

can approximate it by an average, however. This technique is quite powerful and commonplace in mathematical statistics.

*Proof of Theorem 6.5.* Let's simplify notation and write $\partial_\theta = \partial/(\partial\theta)$. To find the maximum of the log-likelihood $\ell(\theta)$, we equate its first derivative to zero. Thus, the MLE $\widehat{\theta}$ solves

$$\sum_{i=1}^n \partial_\theta \ln f(X_i; \widehat{\theta}) = 0.$$

A first order Taylor approximation around $\theta^*$ gives

$$0 = \sum_{i=1}^n \partial_\theta \ln f(X_i; \widehat{\theta}) \approx \sum_{i=1}^n \partial_\theta \ln f(X_i; \theta^*) + \sum_{i=1}^n \partial_\theta^2 \ln f(X_i; \theta^*)(\widehat{\theta} - \theta^*).$$

Solving the above for $\widehat{\theta} - \theta^*$ yields

$$\widehat{\theta} - \theta^* \approx -\frac{\sum_{i=1}^n \partial_\theta \ln f(X_i; \theta^*)}{\sum_{i=1}^n \partial_\theta^2 \ln f(X_i; \theta^*)} = \frac{\frac{1}{n}\sum_{i=1}^n \partial_\theta \ln f(X_i; \theta^*)}{-\frac{1}{n}\sum_{i=1}^n \partial_\theta^2 \ln f(X_i; \theta^*)}.$$

By the law of large numbers, the denominator converges to $I(\theta^*)$ in probability. Thus,

$$\widehat{\theta} - \theta^* \approx \frac{\frac{1}{n}\sum_{i=1}^n \partial_\theta \ln f(X_i; \theta^*)}{I(\theta^*)}.$$

Now apply the central limit theorem to the the right hand side. (This was the interesting part of the proof, you can skip the following details if you want.)

Because, by the chain rule

$$\partial_\theta \ln f(X_i; \theta^*) = \frac{\partial_\theta f(X_i; \theta^*)}{f(X_i; \theta^*)},$$

it holds

$$\begin{aligned}
\mathbb{E}\Big[\partial_\theta \ln f(X_i; \theta^*)\Big] &= \int \frac{\partial_\theta f(x; \theta^*)}{f(x; \theta^*)} f(x; \theta^*) dx \\
&= \int \partial_\theta f(x; \theta^*) dx \\
&= \partial_\theta \int f(x; \theta^*) dx \\
&= \partial_\theta 1 \\
&= 0.
\end{aligned}$$

Further,

$$\begin{aligned}\mathbb{V}\left[\frac{\partial_\theta \ln f(X_i;\theta^*)}{I(\theta^*)}\right] &= \frac{1}{I(\theta^*)^2}\mathbb{V}\big[\partial_\theta \ln f(X_i;\theta^*)\big] \\ &= \frac{1}{I(\theta^*)^2}\left(\mathbb{E}\big[\big(\partial_\theta \ln f(X_i;\theta^*)\big)^2\big] - \mathbb{E}\big[\partial_\theta \ln f(X_i;\theta^*)\big]^2\right) \\ &= \frac{1}{I(\theta^*)^2}\mathbb{E}\big[\big(\partial_\theta \ln f(X_i;\theta^*)\big)^2\big] \\ &= \frac{1}{I(\theta^*)}.\end{aligned}$$

Then the result follows from the CLT. □

## 6.4 The delta method

Sometimes we are interested in a transformation $g(\theta^*)$ of a parameter $\theta^*$ rather than the parameter itself. If we know that $\widehat{\theta}$ is asymptotically normal, what does that mean for $g(\widehat{\theta})$? For example, one can show that the sample variance $S_n^2$ is asymptotically normal. To preserve units, we would like to look at the sample standard deviation $g(S_n^2) = \sqrt{S_n^2}$.

The answer is simple. If $g$ is continuously differentiable, then $g(\widehat{\theta}) - g(\theta^*)$ is also asymptotically normal. This follows from the Taylor approximation

$$g(\widehat{\theta}) - g(\theta^*) \approx g'(\theta^*)(\widehat{\theta} - \theta^*).$$

The derivative $g'(\theta^*)$ tells us how to adjust the asymptotic variance.

**Theorem 6.7** (Delta method). *Suppose $\sqrt{n}(\widehat{\theta} - \theta^*) \to_d \mathcal{N}(0,\sigma^2)$ and that $g$ is continuously differentiable. Then,*

$$\sqrt{n}\big(g(\widehat{\theta}) - g(\theta^*)\big) \to_d \mathcal{N}\big(0, g'(\theta^*)^2\sigma^2\big).$$

**Example 6.8.** *Let $\sigma^2 = \mathbb{V}[X]$. For the sample variance $S_n^2$, one can show $\sqrt{n}(S_n^2 - \sigma^2) \to_d \mathcal{N}(0, \mu_4 - \sigma^4)$, where $\mu_4 = \mathbb{E}[(X-\mu)^4]$. Now consider the sample standard deviation $S_n = g(S_n^2) = \sqrt{S_n^2}$. It hold's $g'(x) = 1/(2\sqrt{x})$ and therefore*

$$\sqrt{n}(S_n - \sigma) \to_d \mathcal{N}\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right).$$

**Example 6.9.** *The delta rule is often useful when computing probabilities from an estimated model. Recall the Corona crash example following Example 5.16. We computed the MLE $\widehat{\alpha}$ and then a probability $p(\widehat{\alpha}) = 0.021 \times \big(1 - F_{\xi,\widehat{\alpha}}(0.17)\big)$.*[4]

[4] Let's treat 0.021 as a fixed number for simplicity. Strictly speaking it's also a random variable.

*You can check that $F_{\xi,\alpha}(x) = 1 - (\xi/x)^\alpha$ for $x > \xi$. Hence,*

$$p'(\alpha) = -0.021 \times (\xi/0.17)^\alpha \ln(\xi/0.17),$$

*and*

$$\sqrt{n}\big(p(\widehat{\alpha}) - p(\alpha^*)\big) \to_d \mathcal{N}\big(0, p'(\alpha^*)^2(\alpha^*)^2\big).$$

*For easier interpretation, we also computed an inverse probability $t(\widehat{\alpha}) = 1/(52 \times p(\widehat{\alpha}))$, which we interpret as the expected number of years between such events. It holds,*

$$t'(\alpha) = -\frac{p'(\widehat{\alpha})}{52 \times p(\widehat{\alpha})^2} = \frac{0.021 \times (\xi/0.17)^\alpha \ln(\xi/0.17)}{52 \times p(\widehat{\alpha})^2},$$

*and*

$$\sqrt{n}\big(t(\widehat{\alpha}) - t(\alpha^*)\big) \to_d \mathcal{N}\big(0, t'(\alpha^*)^2(\alpha^*)^2\big).$$

## 6.5 Confidence intervals

*Confidence intervals* are the most common way to communicate uncertainty. We want to construct an interval $(\widehat{\theta}_l, \widehat{\theta}_u)$ around around the estimated value $\widehat{\theta}$ in a way that we can be confident that it covers the true parameter $\theta^*$. Our confidence is quantified by a probability $\gamma$, the *confidence level.* A $\gamma$-confidence interval is an interval that includes the true parameter with probability at least $\gamma$.

**Definition 6.10** (Confidence interval)**.** *An interval $(\widehat{\theta}_l, \widehat{\theta}_u)$ is called a **$\gamma$-confidence interval**, if*

$$\mathbb{P}\big(\theta^* \in (\widehat{\theta}_l, \widehat{\theta}_u)\big) \geq \gamma.$$

There is a subtlety: The parameter $\theta^*$ is a fixed number, it is the interval that is random. So the probability in Definition 6.10 is a statement about the interval, not about the true parameter. The graph[5] in Fig. 6.1 might help to understand this better. We repeat the same experiment 14 times:

(i) simulate data $X_1, \dots, X_n$,

(ii) compute the MLE $\widehat{\theta}$,

(iii) construct a confidence interval $(\widehat{\theta}_l, \widehat{\theta}_u)$.

[5] Taken from `https://seeing-theory.brown.edu`, a beautiful introduction to statistics with interactive graphics. Check it!

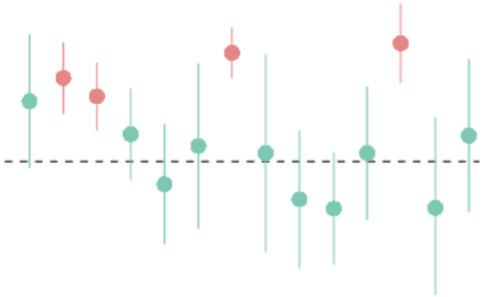


Figure 6.1: Illustration of confidence intervals. The dashed line is the true parameter $\theta^*$, intervals are constructed repeatedly from simulated data.

The dashed line indicates the fixed location of the true parameter $\theta^*$. The dots are the estimates $\widehat{\theta}$, the bars indicate the intervals $(\widehat{\theta}_l, \widehat{\theta}_u)$. The estimates and intervals are random, so they are different for every of the 14 runs. Some of the intervals cover the true value $\theta^*$, some don't. For $\gamma$-confidence intervals, we expect that the long-run proportion[6] of intervals covering $\theta^*$ is at least $\gamma$.

So how do we construct such intervals? Suppose that an estimator $\widehat{\theta}$ is asymptotically normal, that is $\widehat{\theta} \approx \mathcal{N}(\theta^*, \mathrm{se}[\widehat{\theta}]^2)$. The standard error $\mathrm{se}[\widehat{\theta}]$ may not be known, but estimated by some $\widehat{\mathrm{se}}[\widehat{\theta}]$ (see, e.g., Example 6.3). Recall that $\Phi$ is the cdf of the standard normal function. Set $\gamma = 1 - \alpha$ ($\alpha$ is called *significance level* in a related context, but we'll get to that). Define $z_{\alpha/2}$ as the corresponding $(1 - \alpha/2)$-quantile

$$z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$$

and note that, by symmetry of $\Phi$, $-z_{\alpha/2} = \Phi^{-1}(\alpha/2)$.

**Theorem 6.11.** *If $\widehat{\theta} \approx \mathcal{N}(\theta^*, \mathrm{se}[\widehat{\theta}]^2)$ and $\widehat{\mathrm{se}}[\widehat{\theta}] \to_p \mathrm{se}[\widehat{\theta}]$, the interval*

$$(\widehat{\theta}_l, \widehat{\theta}_u) = \left(\widehat{\theta} - z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\theta}], \widehat{\theta} + z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\theta}]\right)$$

*is an approximate $(1 - \alpha)$-confidence interval.*

[6] 'Long-run' means that we repeat the experiment a large number of times

*Proof.* It holds

$$\begin{aligned}
\mathbb{P}\big(\theta^* \in (\widehat{\theta}_l, \widehat{\theta}_u)\big) &= \mathbb{P}\big(\widehat{\theta} - z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\theta}] < \theta^* < \widehat{\theta} + z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\theta}]\big) \\
&= \mathbb{P}\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta^*}{\widehat{\mathrm{se}}[\widehat{\theta}]} < z_{\alpha/2}\right) \\
(\widehat{\mathrm{se}}[\widehat{\theta}] \to_p \mathrm{se}[\widehat{\theta}]) \qquad &\approx \mathbb{P}\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta^*}{\mathrm{se}[\widehat{\theta}]} < z_{\alpha/2}\right) \\
(\text{asymptotic normality}) \qquad &\approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\
(\text{definition of } z_{\alpha/2}) \qquad &= (1 - \alpha/2) - \alpha/2 \\
&= 1 - \alpha.
\end{aligned}$$

□

The choice of $\gamma$ (or, equivalently, $\alpha$) is up to the researcher. The most common choices are $\gamma = 90\%$ and $\gamma = 95\%$. The corresponding quantiles are

$$z_{5\%} \approx 1.64, \quad \text{and} \quad z_{2.5\%} \approx 1.96.$$

**Example 6.12.** *Let $\theta^* = \mathbb{E}[\theta^*]$ and $\widehat{\theta} = \bar{X}_n$. The CLT states $\widehat{\theta} \approx \mathcal{N}(\theta^*, \mathbb{V}[X]/n)$, so $\mathrm{se}[\widehat{\theta}] = \sqrt{\mathbb{V}[X]/n}$, which we can approximate by $\widehat{\mathrm{se}}[\widehat{\theta}] = S_n/\sqrt{n}$. Hence,*

$$\left(\bar{X}_n - \frac{z_{\alpha/2}S_n}{\sqrt{n}}, \bar{X}_n + \frac{z_{\alpha/2}S_n}{\sqrt{n}}\right)$$

*is a $(1-\alpha)$-confidence interval for the sample average.*

**Example 6.13.** *Let's reconsider our Corona crash example. We computed the MLE for the Pareto shape as 2.87 following Example 5.16. In Example 6.9, we have shown that*

$$\sqrt{n}\big(t(\widehat{\alpha}) - t(\alpha^*)\big) \to_d \mathcal{N}\big(0, t'(\alpha^*)^2(\alpha^*)^2\big),$$

*where $t(\alpha^*)$ is the expected number of years between events. Recall that the MLE was $\widehat{\alpha} = 2.87$, $\xi = 0.05$, and $n = 39$. Substituting these values in the expressions derived in Example 6.9 yields*

$$\widehat{\mathrm{se}}[\widehat{\theta}] = \frac{|t'(\alpha^*)|\alpha^*}{\sqrt{n}} \approx 13.8.$$

*Therefore, a 90%-confidence interval for $t(\widehat{\alpha})$ is*

$$(27 - 1.64 \cdot 13.8, 27 + 1.64 \cdot 13.8) \approx (4.23, 49.63).$$

*With 90% confidence, we expect Corona-like crashes to happen every 4 to 50 years.*

**Remark 6.3.** *Note that the conditions of Theorem 6.11 do not apply to the histogram because of its bias. The intervals can still be used to guide intuition,* *but the confidence level will not be correct.*

## 6.6 The bootstrap

The above procedure requires three things:

(i) asymptotic normality of an estimator $\widehat{\theta}$,

(ii) an expression for the standard error $\mathrm{se}[\widehat{\theta}]$,

(iii) a consistent estimator $\widehat{\mathrm{se}}[\widehat{\theta}]$ for the standard error.

The first is rarely an issue. The second and third require hard work. For complex statistical models or estimators, the standard error may not be known, difficult to derive, or difficult to estimate.

Luckily, Bradley Efron came up with an ingenious idea in 1987. The *bootstrap* is one of the most celebrated and widely used techniques for uncertainty quantification. Recall that to quantify uncertainty, we need to approximate the distribution of the random variable $\widehat{\theta} - \theta$.[7] Alas, we only observe a single realization of this variable: the estimate computed from the observed data $X_1, \dots, X_n$.

Suppose for a moment that we can simulate from the true distribution $F$. Consider the following **bootstrap algorithm**:

Step 1. Simulate $B \in \mathbb{N}$ independent data sets $X_{1,b}, \dots, X_{n,b}$ from $F$, for $b = 1, \dots, B$.

Step 2. For each $b$, compute the estimator $\widehat{\theta}_b = g(X_{1,b}, \dots, X_{n,b})$.

Step 3. Define $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$ as the $\alpha/2$ and $(1 - \alpha/2)$ sample quantiles of the 'observations' $\widehat{\theta}_1, \dots, \widehat{\theta}_B$.

Step 4. Define the confidence interval

$$(\widehat{\theta}_l, \widehat{\theta}_u) = (\widehat{q}_{\alpha/2}, \widehat{q}_{1-\alpha/2}).$$

For large $B$ and $n$, $(\widehat{\theta}_l, \widehat{\theta}_u)$ is an approximate $(1 - \alpha)$-confidence interval if $\widehat{\theta}$ is unbiased. Why so? Recall that we are interested in the distribution of the random variable $\widehat{\theta}$. The algorithm allows us to 'observe' $B$ independent realizations from this distributions. If this feels like we are "pulling ourselves up by our bootstraps"[8], you now also know where name comes from.

However, we made the assumption that we can simulate from $F$. But if we would know $F$, we wouldn't need to estimate anything. In practice, we replace $F$ by a consistent estimate $\widehat{F}$. There are two ways to do that:

[7] This distribution is also called *sampling distribution* of $\widehat{\theta}$.

[8] https://en.wiktionary.org/wiki/pull_oneself_up_by_one%27s_bootstraps

- **Parametric bootstrap**: We postulate a parametric model $\mathcal{F} = \{f_\beta \colon \beta \in \mathcal{B}\}$ and compute an estimator $\widehat{\beta}$ for its parameter. Then we simulate from the model $F_{\widehat{\beta}}$ in step 1 of the algorithm.

- **Nonparametric bootstrap**: We approximate $F$ by the empirical distribution function $F_n$ of the data $X_1, \dots, X_n$. Simulating $X_{1,b}, \dots, X_{n,b}$ from $F_n$ is equivalent to drawing $n$ times from the set $\mathcal{X} = \{X_1, \dots, X_n\}$ with replacement. That is, for each $i = 1, \dots, n$, draw a random variable $J$ from the uniform distribution on $\{1, \dots, n\}$ and set $X_{i,b} = X_J$.

The parametric bootstrap fails if the model $\mathcal{F}$ is misspecified. The nonparametric bootstrap essentially always works. The number $B$ should be chosen large enough for the quantile estimates $\widehat{q}_{\alpha/2}, \widehat{q}_{1-\alpha/2}$ to be stable. $B = 100$ is the absolute minimum, better use 200 or 500.

So why should we care about asymptotic normality at all? The bootstrap is computationally demanding. For example, numerically maximizing the likelihood of a complex model with many parameters can take forever, especially on larger data sets. If computing the estimator $\widehat{\theta}$ once is expensive, computing it a large number of times is infeasible.

# 7 Testing

The history of science is full of hypotheses: Pavlov's dogs, Mendel's peas, Newton's Radiant Prisms, and so forth. You surely heard about a few famous hypotheses in astronomy: Copernicus' hypothesis that planets circulate around the sun, Hubble's hypothesis of the expanding universe, the hypothesis that our universe contains dark energy. An intrinsic property of an hypothesis is that, at the time it is stated, we don't know whether it is true or not. What we can do is a reality check: does the data contradict the hypothesis? The statistical term for this check is *hypothesis testing*.

This chapter is a bit different than the others. The statistical framework for hypothesis testing consists of many different concepts and definitions. It's easy to get lost in the details and miss the bigger picture. To avoid that, we shall quickly walk through all concepts with a concrete example at hand (Section 7.1). We then discuss a number of issues with hypothesis testing in Section 7.2. These two sections contain the main takeaways from this chapter. The following sections introduce the concepts more formally and give additional details.

Something else is unusual: we will see only one concrete examples of a statistical test. There are thousands, each of them different in its own way. I don't find it useful to teach you specific tests that you will immediately forget after this course is over. What's important is that you understand the core principles and the problems associated with them.

## 7.1 Hypothesis testing: a quick walk through

### Null and alternative hypotheses

Marius mentioned in his session that star-forming galaxies tend to emit bluer light than passive ones. He said there are physical reasons to believe this, but without having seen any data, we should treat it as an hypothesis. I have no clue about physics and want to disprove him.

To use statistics (or do science), we need to formalize the hypothesis. Let the random variable $X^{(A)}$ be the blue light ($g - r$ apparent magnitude) emitted by a star-forming galaxy and $X^{(P)}$ the blue light emitted by a passive galaxy. Our hypothesis is that the difference $\Delta^* = \mathbb{E}[X^{(A)}] - \mathbb{E}[X^{(P)}]$ is negative: on average, star-forming galaxies emit bluer light.

The hypothesis we want to test is called the *null hypothesis* or $H_0$. Every state of our world where $H_0$ does not hold, is collected in the *alternative hypothesis*,

$H_1$. In our example,

$$H_0\colon\ \Delta^* < 0, \qquad H_1\colon\ \Delta^* \geq 0.$$

Now we want to check whether the data contradicts the hypothesis. Note that I want to reject $H_0$ to prove that I'm right. This is how statistical tests are usually set up, but more on that later.

### Test statistics

Suppose that we know which galaxies in the data are star-forming and which are not. We have data $X_1^{(A)}, \dots, X_n^{(A)}$ from active galaxies and data $X_1^{(P)}, \dots, X_m^{(P)}$ from passive galaxies. The true value of $\Delta^*$ is unknown to us, but can be estimated by

$$\widehat{\Delta} = \bar{X}_n^{(A)} - \bar{X}_m^{(P)} = \frac{1}{n}\sum_{i=1}^n X_i^{(A)} - \frac{1}{m}\sum_{i=1}^m X_i^{(P)}.$$

One can use the CLT to show that

$$\widehat{T}_n = \frac{\widehat{\Delta} - \Delta^*}{\sqrt{\widehat{\sigma}_A^2/n + \widehat{\sigma}_P^2/m}} \to_d \mathcal{N}(0,1),$$

where $\widehat{\sigma}_A^2$ and $\widehat{\sigma}_P^2$ are the sample variances of active and passive galaxies. The random variable $\widehat{T}_n$ is called *test statistic*, because we will use it to decide whether to reject the null hypothesis. Recall that $H_0$ states that $\Delta^*$ is negative. The larger (more positive) $\widehat{\Delta}$ (or $\widehat{T}_n$), the more evidence the data provide against $H_0$.

### P-values

Note that $\widehat{T}_n$ is a random variable (because $\widehat{\Delta}$ is) from which we see only one realization — the one computed from the data we observed. Let's denote this number by $t$ to make the distinction between the random variable and the realization more clear.

To decide whether or not to reject $H_0$, we compute the *p-value*: the probability of seeing a value of $\widehat{T}_n$ at least as large as $t$, *if* $H_0$ *would be true*:

$$p = \mathbb{P}_{H_0}(\widehat{T}_n \geq t),$$

where the subscript $H_0$ indicates that this probability is computed assuming that $H_0$ is true. If the probability is small, we have found evidence against $H_0$. Warning: If $p$ is large, we can only conclude that we found *no evidence against* $H_0$, not that we found evidence *for* it.

$H_0$ spans many possible values of $\Delta^*$ and its good practice to assume the worst case, $\Delta^* = 0$.[1] In that case,

$$\widehat{T}_n = \frac{\widehat{\Delta}}{\sqrt{\widehat{\sigma}_A^2/n + \widehat{\sigma}_P^2/m}} \to_d \mathcal{N}(0, 1).$$

and

$$p = \mathbb{P}_{(\Delta^*=0)}(\widehat{T}_n > t) = 1 - \Phi(t).$$

The test above is a *Wald test*. That is, a test constructed from an asymptotically normal estimator. Since you already know many estimators that are asymptotically normal, you should be able to construct Wald tests for other types of hypothesis as well.

### Significance

Ultimately, we want to make a decision: do we reject $H_0$ or not? We do this by comparing the *p-value* against a *significance level* $\alpha$. Recall that, a small value of $p$ constitutes evidence against $H_0$. Hence, we use the rule

- if $p < \alpha$: reject $H_0$,
- if $p \geq \alpha$: don't reject $H_0$.[2]

If $p < \alpha$, we also say that the result is *statistically significant* at level $\alpha$. Similar to the confidence level $\gamma$, choosing the significance level $\alpha$ is up to the researcher. The most common value is 5%, but this depends on the field and type of research.

But what does it actually mean? The value of $\alpha$ controls the probability of a *false positive*: rejecting $H_0$ although it is true. We call this a 'positive', because most tests use $H_0$ as the hypothesis of 'no effect'. If we want to establish an effect, we actually *want to reject* $H_0$. $\alpha = 5\%$ means that, if $H_0$ is true, we expect it to be rejected in 5% of the cases — just due to chance. This is unacceptable in many physical experiments, where much smaller levels for $\alpha$ are used.

However, the smaller $\alpha$, the harder it is to reject $H_0$ (or to 'detect an effect'). The probability of detecting a real effect (rejecting $H_0$ when it is false) is also called *power* of the test. If a test has little power, we will rarely reject $H_0$, no matter if it is true or not. That's why a large $p$-value should not be interpreted as evidence for $H_0$. Generally, the power of a test increases if we have more data to base our decision on.

### Multiple testing

Let's assume we found $t = -3$, such that $p \approx 0.999$. Unfortunately, I couldn't find evidence against Marius' hypothesis that star-forming galaxies emit bluer light.

[1] Here, worst case means that it is harder to find evidence against $\Delta^* < 0$ than against $\Delta^* < c$ for any other $c \leq 0$.

[2] Again, we never "accept", we only "not reject".

That's a bit embarrassing. Maybe I can at least prove that I'm not a complete idiot and the difference is small. Let's test a new hypothesis $H_0 : \Delta^* < -0.2$. We find $p = 0.04$ and conclude that I'm not a complete idiot at significance level $\alpha = 5\%$.

That would be even more embarrassing, because it would mean that I also have no clue about statistics. By testing two hypotheses, we increase the probability of a false positive (reject $H_0$ although it is true). It means that, even if we compare $p$ against 5%, the level of the two tests combined is larger than 5%. This is a *multiple testing problem* and for sake of good science, we need to correct for it.

There are two popular ways to do that:

- The *Bonferroni correction* compares $p$ against $\alpha/m$, where $m$ is the number of tests. This correction guarantees that the false positive rate is at most $\alpha$. It is generally conservative and safeguards against the worst case.

- The *Benjamini-Hochberg (BH) method* is a bit more complicated and does not control the false positive rate (the proportion of false rejections *among all tests*). Instead, it controls the *false discovery rate*: the proportion of false rejections *among all rejections*. This is generally less conservative.

## 7.2 Issues with hypothesis testing

There are several issues with hypothesis testing[3] and we should address them early.

### 7.2.1 Overuse

There is a tendency to overuse statistical tests. Very often, estimation and confidence intervals are better tools. It's a good idea to ask yourself three questions:

Q1. Do I have a well-defined and well-motivated hypothesis to test for?

Q2. Do I really want to make a yes-or-no decision?

Q3. Do I really care about error probabilities?

If one of the answers is 'no', then testing is probably not the right tool. Let me give a few common examples where tests are misused:

- *Exploratory data analysis*: There is absolutely no reason to do formal tests while exploring a data set. In EDA, we would normally answer Q1 and Q3 above with 'no'. Visualization, estimation, and confidence intervals tell you everything you need. Yes-or-no decisions like "should I remove this outlier?" should be based on the scientific context rather than a statistical test.

[3] The American Statistical Association even issued a statement on this: https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN

- *Assessing estimation uncertainty*: There is a reason why we didn't speak about tests in the previous chapter. If you want to assess or communicate how variable estimates are, there's no need to make a yes-or-no decision (Q2). Use confidence intervals!
- *Model selection*: Quite often, we have multiple plausible models $\mathcal{F}_1, \dots, \mathcal{F}_M$ and want to decide which is best. Let $F$ be the unknown true distribution of the data. A widespread (mal-)practice is to test the hypothesis $H_0 \colon F \in \mathcal{F}_m$ for all $m = 1, \dots, M$ and pick the one with the largest $p$-value. That's not what testing is for (see Q1-Q3). If possible, use visualizations like histograms and QQ-plots to see if the models fit the data or where they deviate. We shall see formal methods for model selection later in the course.

### 7.2.2 Statistical significance vs. scientific relevance

Often, we know upfront that the hypothesis is wrong. For example, if $H_0$ is 'there is no difference between populations A and B' or '$F \in \mathcal{F}$', it's hard to believe that this is exactly right. But a statistically significant violation of $H_0$ may be scientifically irrelevant. If we have enough data, we can detect even the smallest deviations from a hypothesis. But would it be relevant if blue light emissions in active and passive galaxies differ by $10^{-100}$ magnitudes? Effect sizes matter — and whether they are relevant depends on the scientific context. The $p$-value alone has little meaning.

### 7.2.3 Misinterpretation of p-values

The $p$-value is *not* the probability that $H_0$ is true (or $H_1$ is false). This is a widespread mistake, commonly found in articles on public media.

The $p$-value is a probability, but a weird one: if $H_0$ was true, how unlikely is it that the test statistic is as large (or small) as the one we observed. The important part is 'if $H_0$ was true'. If $H_0$ is not true (and that's what we want to show), then the probability bears no real-world meaning. Because the actual interpretation is so unwieldy, it is not a good a measure to communicate scientific results to broader audiences.

### 7.2.4 The replication crisis

You may have heard of the *replication crisis* in psychology and medicine. Large consortia of researchers set out to reproduce results of high profile scientific 'discoveries'. The shocking outcome was that — even with huge sample sizes and identical study protocols — many findings could not be reproduced. Conservative estimates today say that around a third of 'discoveries' in these fields are in fact false positives. Many of them were considered well-established and formed entire lines of research, with hundreds of publications over several decades. Similar observations were made in other disciplines.

But with $\alpha = 5\%$ or lower, how can it be that the false positive rate is so large? There are several likely reasons for this, some good and some bad. Among the good ones is that multiple testing issues are ignored across a huge proportion of the scientific literature — mainly due to a lack of awareness and insufficient statistics education. If this reason counts as 'good', you get an idea of what comes next.

Academic journals are less likely to publish research with no significant results. Because scientists know this, many don't even try to publish insignificant ones. We don't really know how often someone failed to reject a hypothesis, we only see the significant results. This is known as the *file drawer effect*.

It gets worse. If people are aware of it or not, uncorrected multiple testing actually makes it easier to claim 'scientific discoveries'. If we use $\alpha = 5\%$, we can test 100 things where there is no effect, but will make significant 'discoveries' in 5 of them — just by chance. Scientists acquire fame and secure their job through 'discoveries', so there is an incentive to make as many as possible. As a consequence, the incentives suggest to *not* correct for multiple testing.

Much worse. Remember when I was unhappy with the outcome of my test, so I tested another hypothesis instead? This is known as *HARKing* or *hypothesizing after results are known* and it's problematic. If the same data is used to form a hypothesis and to test it, all inferences (like error probabilities) are corrupted. Unfortunately, this is practice is widespread. Intentions don't need to be bad. For example, data can be expensive or even impossible to collect twice (for example, when testing hypotheses about a certain time period). What one can do nevertheless, is to clearly communicate how (and when) a hypothesis was formed and whether this has implications for inference.

In fact, surveys suggest that many researchers torture their data until they make a 'discovery'. This can mean to come up with and test new hypotheses until $p < 0.05$ for one of the tests. It can also mean to change the data to push the $p$-value beyond the significance boundary by, e.g., excluding or including outliers, control variables, or sub-groups of the data. These practices are known as *p-hacking* and are poison to scientific progress.

In the past five years or so, these issues started to attract attention and things are changing for the better. Luckily, astronomy and physics are fields where such practices have been less problematic. But they are not immune to these issues either. Take the above as a cautionary tale. Small violations of the rules accumulate and corrupt the scientific endeavor. So it's better to be aware and avoid corrupting your own field.

## 7.3 Null and alternative hypotheses

Now it's time to formalize our walk-through above. A statistical hypothesis is a statement about an unknown parameter $\theta^* \in \Theta$. We separate the parameter space $\Theta$ into two parts, $\Theta = \Theta_0 \cup \Theta_1$. A statistical hypothesis has the form

$$H_0\colon\ \theta^* \in \Theta_0, \qquad H_1\colon\ \theta^* \in \Theta_1.$$

There are two common types of hypotheses involving a specific value $\theta_0$:[4]

- **one-sided hypotheses**:
  - $H_0\colon \theta^* < \theta_0$ and $H_1\colon \theta^* \geq \theta_0$,
  - $H_0\colon \theta^* \leq \theta_0$ and $H_1\colon \theta^* > \theta_0$,
  - $H_0\colon \theta^* > \theta_0$ and $H_1\colon \theta^* \leq \theta_0$,
  - $H_0\colon \theta^* \geq \theta_0$ and $H_1\colon \theta^* < \theta_0$,
- **two-sided hypotheses:** $H_0\colon \theta^* = \theta_0$ and $H_1\colon \theta^* \neq \theta_0$.

Marius' hypothesis above was one-sided. Two-sided hypotheses are generally more common. Let's see a few examples before we continue.

**Example 7.1.** *Consider the hypothesis that, on average, stars in the Milky Way and Andromeda galaxies have the same mass. If $\mu_{MW}$ is the expected mass of a star in the Milky Way and $\mu_A$ the expected mass of an Andromeda star. Then $\theta^* = \mu_{MW} - \mu_A$ and*

$$H_0\colon\ \mu_{MW} - \mu_A = 0, \qquad H_1\colon\ \mu_{MW} - \mu_A \neq 0.$$

**Example 7.2.** *Consider the hypothesis that the metallicity of a quasar is independent of its age. If they are independent, the theory predicts that they must be uncorrelated. Denote by $\rho$ the correlation between metallicity and age. Then $\theta^* = \rho$ and*

$$H_0\colon\ \rho = 0, \qquad H_1\colon\ \rho \neq 0.$$

**Example 7.3.** *Consider the hypothesis that the luminosity of stars in the Milky way (in magnitudes) follows a normal distribution. Denote by $\Phi_{\mu,\sigma^2}$ the corresponding CDF and let $\mathcal{F} = \{\Phi_{\mu,\sigma^2}\colon (\mu, \sigma^2) \in \mathbb{R} \times (0,\infty)\}$ be the statistical model. The parameter of interest is the true CDF $F$. That is, $\theta^* = F$[5] and*

$$H_0\colon\ F \in \mathcal{F}, \qquad H_1\colon\ F \notin \mathcal{F}.$$

*This is called* goodness-of-fit (GoF) testing *and sometimes used as plausibility check. (I'm not a big fan of this.)*

[4] The 'side' refers to the alternative hypothesis.

[5] Note that here the parameter $\theta^*$ is not just a number, but an entire function.

## 7.4 Test statistics

A statistical test is an evidence-based decision: given the data, should we reject the null hypothesis? A *test statistic* is a number that helps us to make this decision. It is similar to an estimator in the sense that it is a function of the data: $\widehat{T}_n = t(X_1, \dots, X_n)$ for some function $t$.

**Definition 7.4.** *A* ***statistical test*** *is a decision rule based on a test statistic $\widehat{T}_n$ and set $\mathcal{R}$ (called rejection region):*

- *if $\widehat{T}_n \in \mathcal{R}$, reject the null-hypothesis,*
- *if $\widehat{T}_n \notin \mathcal{R}$, do not reject the null-hypothesis.*

Most commonly, the rejection region takes the form $\{\widehat{T}_n > c\}$ (for one-sided tests) or $\{|\widehat{T}_n| > c\}$ (for two-sided tests), where $c \in \mathbb{R}$ is a *critical value*. It is more common to reformulate the decision rule in terms of $p$-values, to which we'll get in a minute.

Note that we never 'accept' the null-hypothesis. If we don't reject it, this can have several reasons. The main one is that the test statistic is not informative enough. That does not mean that we found evidence for $H_0$, only that we couldn't find any against it.

## 7.5 Test errors

Wasserman gives a nice analogy in his book:

> *Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain $H_0$ unless there is strong evidence to reject $H_0$.*

There are two types of errors we can make: convicting someone innocent and letting the perpetrator go unpunished. The same is true for statistical tests:

- **type I error**: rejecting $H_0$ although it is true,
- **type II error**: retaining $H_0$ although it is false.

I've said it before, let me say it again: statisticians are terrible at naming things.[6] A more intuitive terminology comes from medicine. The outcome of a medical test is termed *positive* if it indicates disease (as in 'HIV-positive') and *negative* if not. The type I error corresponds to a *false positive*: diagnosing a disease when the patient is healthy. The type II error corresponds to a *false negative*: not detecting the disease although the patient is ill. See the table below for a summary.

[6] Confession time: I need to check Wikipedia every time 'type I/II' errors are mentioned.

| | retain $H_0$ | reject $H_0$ |
|---|---|---|
| $H_0$ true | ☺ | type I/false positive |
| $H_0$ false | type II/false negative | ☺ |

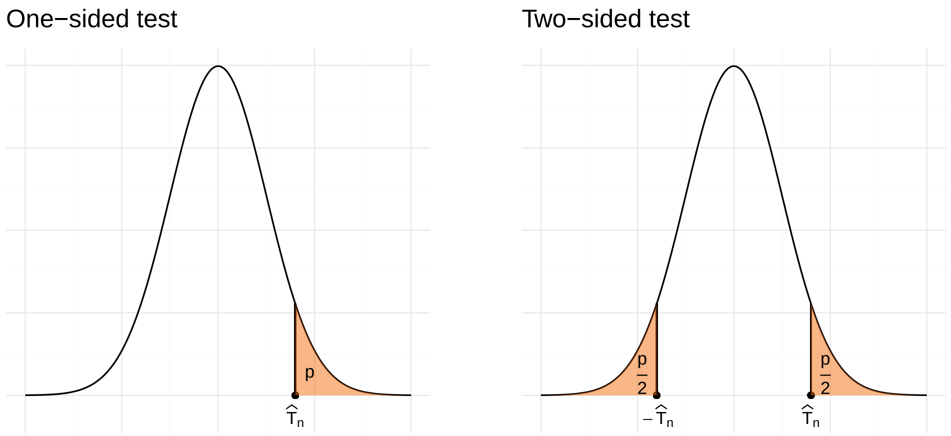


Figure 7.1: Illustration of $p$-values: The curve is the density of the test statistic under the null hypothesis. The $p$-value is the area under the curve beyond the observed test statistic $\widehat{T}_n$.

## 7.6 Significance and p-values

Just like estimators, test statistics are random variables. Hence, we can make probabilistic statements about the outcome of a statistical test. This motivates the convention for setting up the rejection region: choose the critical value $c$ such that the probability of a false positive does not exceed some target level $\alpha$. This level is called *significance level* or *size* of the test.

A false positive is a rejection of $H_0$ although it is true. So assuming that $H_0$ is true, the probability of a false positive is[7]

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta(\widehat{T}_n > c) = \mathbb{P}_{\theta_0}(\widehat{T}_n > c) \qquad \text{(one-sided)} \quad \text{or}$$
$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta(|\widehat{T}_n| > c) = \mathbb{P}_{\theta_0}(|\widehat{T}_n| > c) \quad \text{(two-sided)}.$$

This is also called the *false positive rate.* The subscript $\theta$ in $\mathbb{P}_\theta$ indicates that the probability is computed under the assumption $\theta^* = \theta$: "If $\theta^*$ was equal to $\theta$, what is the probability of rejecting $H_0$?" In the two-sided case, $\Theta_0 = \{\theta_0\}$; in the one-sided case, $\theta_0$ is the worst-case parameter of $\Theta_0$.

Let's denote $F_{\widehat{T}_n}(t) = \mathbb{P}_{\theta_0}(\widehat{T}_n \leq t)$ as the CDF of $\widehat{T}_n$ under $H_0$. Then

$$\mathbb{P}_{\theta_0}(\widehat{T}_n > t) = 1 - F_{\widehat{T}_n}(t), \qquad \mathbb{P}_{\theta_0}(|\widehat{T}_n| > t) = 1 - F_{\widehat{T}_n}(t) + F_{\widehat{T}_n}(-t).$$

The $p$-value is defined as $p = 1 - F_{\widehat{T}_n}(\widehat{T}_n)$ and $p = 1 - F_{\widehat{T}_n}(|\widehat{T}_n|) + F_{\widehat{T}_n}(-|\widehat{T}_n|)$,

[7] The max is actually a sup, but that ship has sailed.

respectively. This is illustrated in Fig. 7.1. By construction, the decision rule

- if $p < \alpha$, reject $H_0$,
- if $p \geq \alpha$, retain $H_0$,

gives a false positive rate of at most $\alpha$. If $p < \alpha$ we speak of a *statistically significant* result at level $\alpha$.

The most widely used level is $\alpha = 5\%$: if $H_0$ is true, we want to reject in less than 5% of the cases. Physics is a notable exception, where much smaller values are common. For example, for the Higgs boson, a statistical test of

$$H_0\colon \text{a Higgs boson does not exist at 126 GeV mass}$$

was required to be significant at level[8] $\alpha = \Phi(-5) \approx 3 \times 10^{-6}$: If there was no Higgs boson, the probability of thinking there is one should be at most $3 \times 10^{-6}$.

## 7.7 Power

The probabilities used above can be generalized as follows.

**Definition 7.5** (Power)**.** *The **power function** of a statistical test is defined as*

$$\beta(\theta) = \mathbb{P}_\theta(\widehat{T}_n \in \mathcal{R}).$$

The value $\beta(\theta)$ is the probability of rejecting the null hypothesis if $\theta^*$ was equal to $\theta$.

When the null hypothesis is false ($\theta^* \in \Theta_1$), we want a to reject it with high probability. That is, we want the power $\beta(\theta)$ to be as large as possible for all $\theta \in \Theta_1$.

**Example 7.6.** *Recall that in the example from Section 7.1,*

$$\widehat{T}_n = \frac{\widehat{\Delta} - \Delta^*}{\sqrt{\widehat{\sigma}_A^2/n + \widehat{\sigma}_P^2/m}} \to_d \mathcal{N}(0, 1),$$

*so*

$$\beta(\Delta) = \mathbb{P}_\Delta(\widehat{T}_n > c) \approx 1 - \Phi\left(\frac{c - \Delta}{\sqrt{\sigma_A^2/n + \sigma_P^2/m}}\right).$$

*A few observations:*

- *As the true difference $\Delta$ grows larger (more positive), the power increases. That is, we are more likely to reject the hypothesis that $\Delta^* < 0$. This makes sense: the more positive the true $\Delta$ is, the easier it is to detect.*

[8] The probability of a 'five standard deviation event' or $5\sigma$-event.

- *The larger the critical value $c$, the stronger the deviation for $H_0$ has to be for us to reject and, consequently, the less powerful is the test.*
- *If $c - \Delta < 0$ and the sample sizes $n, m$ increase, the test becomes more powerful. For if we have more data, it becomes easier to detect deviations from the null.*

The observations made in this example hold more generally. Large deviations from the null are easier to detect, and more data helps.

When the power of a test is low, the probability of rejecting $H_0$ is small, no matter if it is true or not. That's why $\widehat{T}_n \notin \mathcal{R}$ should not be interpreted as evidence for $H_0$.

## 7.8 Multiple testing

Suppose we are testing multiple hypotheses $H_0^{(k)}$, $k = 1, \dots, m$. The *family-wise error rate (FWER)* is defined as the probability of having at least one false positive. Define $A_k$ as the event that test $k$ results in a false positive. Suppose that all hypotheses are true and that each test has level $\alpha$, i.e., $\mathbb{P}(A_k) = \alpha$. Clearly,

$$\text{FWER} = \mathbb{P}\left(\bigcup_{k=1}^{m} A_k\right) \geq \mathbb{P}(A_1) = \alpha.$$

But equality only holds when $A_1 = \dots = A_m$. In the worst case, the events $A_1, \dots, A_m$ are all disjoint. Therefore,

$$\text{FWER} = \mathbb{P}\left(\bigcup_{k=1}^{m} A_k\right) \leq \sum_{k=1}^{m} \mathbb{P}(A_k) = m\alpha.$$

So to ensure FWER $\leq \alpha^*$, we need to take $\alpha = \alpha^*/m$. This is called **Bonferroni correction**.

The Bonferroni method is quite conservative: first, it protects us against the worst-case; second, it protects us against making even a single rejection. If the number of tests $m$ is very large, the Bonferroni method can be prohibitive. For example, such a situation appears in studies of the *cosmic microwave background (CMB)*. To measure the CMB, a satellite is looking in the sky. As the CMB is only the background, it is masked by other objects in the foreground (like thermal dust and galaxies). To filter out the foreground objects, a statistical test is performed with $H_0$: "there is no foreground object in the way". This is done hundreds or thousands of times: once for every tiny part of the sky. In this case, the Bonferroni method would post such a high hurdle that almost no foreground could be detected.

In such situations, it is more reasonable to control a less restrictive measure. The *false discovery rate (FDR)* is defined as the expected proportion of false rejections among all the rejections. A key difference is that we don't assume that all hypotheses are true. Let $m$ be the number of tests, $R$ be the total number of rejections, and $R_0$ be the number of false positives. Then $\mathrm{FDR} = \mathbb{E}[R_0/R]$. The **Benjamini-Hochberg (BH) procedure** allows to control the FDR. To ensure that $\mathrm{FDR} \leq \alpha$:

1. Let $P_{(1)}, \dots, P_{(m)}$ be the $p$-values of the tests in ascending order.
2. Find the largest $k$ such that $P_{(k)} \leq k\alpha/m$. Denote it by $k^*$.
3. Reject $H_0^{(i)}$, for all hypotheses where $P_i \leq P_{(k^*)}$; retain $H_0^{(i)}$ otherwise.

## 7.9 Some classes of tests

To conclude this chapter, we shall briefly discuss some common classes of tests. There's no need to memorize any of this; it's enough to have heard the names. When you really need to do a test in your research (or read about someone else's) and have understood the concepts above, you'll easily be able to find what you need on the internet.[9]

- *t-test*: A test for the mean, i.e., $H_0 : \mathbb{E}[X] = \mu_0$ or $H_0 : \mathbb{E}[X] = \mathbb{E}[Y]$. If the data are exactly Gaussian (so never), the test statistic has a 'Student $t$-distribution'.
- *Wald tests*: Tests derived from asymptotic normality of an estimator. The example in Section 7.1 was such a test.
- *One- and two-sample tests*: A one-sample test is about a property of a single population, e.g., $H_0 : \mathbb{E}[X] = \mu_0$. A two-sample test is about the similarly of two populations, e.g., $H_0 : \mathbb{E}[X] = \mathbb{E}[Y]$.
- $\chi^2$*−tests*: all tests where the test statistic follows a $\chi^2$ distribution: $X \sim \chi^2(k)$ if $X = \sum_{j=1}^k Y_j^2$ with $Y_k \overset{iid}{\sim} \mathcal{N}(0,1)$. Among them are some tests for independence, goodness-of-fit, and many more.
- *Likelihood-ratio tests*: Tests for nested models, like $H_0 : X \sim \mathcal{N}(0,1)$ vs $H_1 \colon X \sim \mathcal{N}(\mu, 1)$ with $\mu \neq 0$. The test statistic is the difference of log-likelihoods and typically follows a $\chi^2$ distribution.
- *Rank-based tests*: Construct test statistics from ranking the data. Among them are tests for equality of distributions and independence.
- *Permutation tests*: Similar as rank-based, but based on random permutations of observations instead.

[9] See, for example, `https://en.wikipedia.org/wiki/Category:Statistical_tests`.

# 8 Regression models

Broadly speaking, *Regression models* are statistical models for conditional distributions. The goal is usually to explain some target quantity ($Y$) with the help of others ($\boldsymbol{X}$). The models can be used to formalize scientific theories and make predictions. Outside of statistics the term *regression* has become out of fashion. But most methods trading under the names *machine learning* and *artificial intelligence* today are fundamentally regression models. We touched on regression models briefly in Chapter 4 and Example 5.15 and will expand on them a bit more in this chapter.

## 8.1 Terminology

A regression model involves two types of variables:

- a *response* variable $Y \in \mathbb{R}$ (also called *independent variable* in social sciences or *label* in machine learning).
- a vector of *covariates* $\boldsymbol{X} \in \mathbb{R}^p$ (also called *dependent variables*, *predictors*, or *features*).

Generally speaking, a regression model is a model for some aspect of the conditional distribution $F_{Y|\boldsymbol{X}}$. Mostly, interest is in the conditional expectation $\mathbb{E}[Y \mid \boldsymbol{X}]$: given the information provided by $\boldsymbol{X}$, what is our best guess for the value of $Y$? This is also called *mean regression*. Less frequently, the conditional distribution $F_{Y|\boldsymbol{X}}$ itself (*distribution regression*) or conditional quantiles $F_{Y|\boldsymbol{X}}^{-1}$ (*quantile regression*) are considered.

Also here, we distinguish between parametric and nonparametric models. As always, parametric models are characterized by a finite-dimensional parameter. This is quite a strong assumption, but facilitates building and interpreting models. Nonparametric models make almost no assumptions, which makes them harder to estimate and interpret. We shall therefore focus on parametric regression models and only briefly discuss nonparametric ones towards the end.

## 8.2 The linear regression model

### 8.2.1 Model formulation

The linear model is both the simplest and most common type of regression model. It takes the form

$$Y = \boldsymbol{\beta}^\top \boldsymbol{X} + \epsilon, \tag{8.1}$$

where

- $\boldsymbol{\beta} \in \mathbb{R}^p$ are model parameters, also called *regression coefficients*;
- $\epsilon$ is a *noise* or *error* term and assumed to satisfy $\mathbb{E}[\epsilon \mid \boldsymbol{X}] = 0$.

An equivalent formulation of model (8.1) is $\mathbb{E}[Y \mid \boldsymbol{X}] = \boldsymbol{\beta}^\top \boldsymbol{X}$: we assume that the conditional expectation $\mathbb{E}[Y \mid \boldsymbol{X}]$ is linear in $\boldsymbol{X}$.

**Remark 8.1.** *The model assumes a linear relationship between the response and the predictors. Note that we could take, for example, $X_3 = X_2^2$, so that non-linear relationships can be represented as well. We will speak more about this later.*

The first element of $\boldsymbol{X}$ is usually set to $X_1 = 1$ and the corresponding coefficient $\beta_1$ called *intercept*. All other elements of $\boldsymbol{X}$ are proper random variables. In this case, the model can be written equivalent as

$$Y = \beta_1 + \sum_{k=2}^{p} \beta_k X_k + \epsilon.$$

In Example 5.15, we separated the intercept from the remaining predictors, but the current formulation will be more convenient.

**Example 8.1.** *Let $V$ and $B$ be the visual and blue band magnitudes of a star. A linear regression model for the color-magnitude diagram is*

$$V = \beta_1 + \beta_2 \times (B\text{-}V) + \epsilon,$$

*where $B$-$V$ is the color index. Fig. 8.1 shows an example of a linear regression model for the color-magnitude diagram of selected stars from the Hipparcos catalog. Each point represents a star, the straight line is the function $\beta_1 + \beta_2 \times (B\text{-}V)$. We see that, on average, the data exhibit an (almost) linear relationship on: the bluer the star, the brighter it tends to be. Of course, not every star falls on the line $\beta_1 + \beta_2 \times (B\text{-}V)$. The vertical distance to the line is the error term $\epsilon$. For some stars it is positive, for some negative; for some larger, for some smaller.*

The regression coefficients in Fig. 8.1 were not chosen arbitrarily, but estimated from the data. That's our next topic.

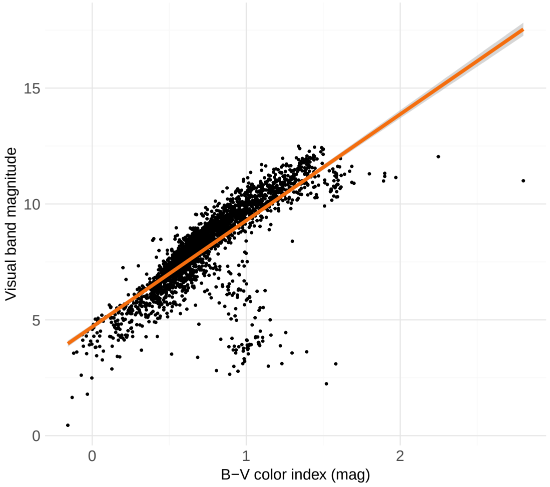


Figure 8.1: Linear regression model for the color-magnitude diagram of 2655 stars from the Hipparcos catalog.

### 8.2.2 Parameter estimation

In Example 5.15, we have derived a way to estimate the regression coefficients. By assuming $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$, the MLE is equivalent to the minimizer of the *least-squares* criterion:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (Y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i)^2.$$

It turns out that this criterion also works if $\epsilon$ is not Gaussian (but then $\widehat{\boldsymbol{\beta}}$ is no longer the MLE). The intuition is that $(Y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i)^2$ is a measure for prediction error. The smaller it is (on average), the better the model is at explaining $Y_i$ from $\boldsymbol{X}_i$. The true parameter $\boldsymbol{\beta}^*$ is the one that explains $Y_i$ best in the sense that the expected error $\mathbb{E}[(Y - \boldsymbol{\beta}^\top \boldsymbol{X})^2]$ is minimal.

To find an explicit expression for $\widehat{\boldsymbol{\beta}}$, we equate the derivative of the criterion to zero:

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)\boldsymbol{X}_i^\top = 0 \\ \Leftrightarrow \quad & \frac{1}{n}\sum_{i=1}^n Y_i \boldsymbol{X}_i^\top = \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \\ \Leftrightarrow \quad & \widehat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n Y_i \boldsymbol{X}_i. \end{aligned} \tag{8.2}$$

Note that $\boldsymbol{X}_i \boldsymbol{X}_i^\top$ is a $p \times p$ matrix, so that the solution involves matrix inversion. The estimator $\widehat{\boldsymbol{\beta}}$ above is also called the *ordinary least squares (OLS)* estimator.

**Theorem 8.2.** *Define* $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(Y - \boldsymbol{\beta}^\top \boldsymbol{X})^2]$. *Then the OLS* (8.2) *is consistent for* $\boldsymbol{\beta}^*$: $\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^*$.

*Proof.* Recall that $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(Y - \boldsymbol{\beta}^\top \boldsymbol{X})^2]$. Using the same arguments as in (8.2), we can show that

$$\boldsymbol{\beta}^* = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]^{-1}\mathbb{E}[Y\boldsymbol{X}].$$

Then the claim follows from the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top \to_p \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top], \qquad \frac{1}{n}\sum_{i=1}^n Y_i\boldsymbol{X}_i \to_p \mathbb{E}[Y \mid \boldsymbol{X}]. \qquad \square$$

**Remark 8.2.** *Note that Theorem 8.2 does not assume that the model* (8.1) *is correctly specified. The OLS converges to the best linear predictor* $\boldsymbol{\beta}^*$ *(the one minimizing* $\mathbb{E}[(Y - \boldsymbol{\beta}^\top \boldsymbol{X})^2]$*) in any case. However, if* (8.1) *does not hold,* $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ *does not converge to* $\mathbb{E}[Y \mid \boldsymbol{X}]$.

Let use look at the OLS estimator in a bit more detail in the simple case where $\boldsymbol{X} = (1, X_2)$. One can show that the formula simplifies to

$$\widehat{\beta}_1 = \bar{Y}_n - \widehat{\beta}_2 \bar{X}_{2,n}, \qquad \widehat{\beta}_2 = R_n \frac{S_{n,Y}}{S_{n,X_2}},$$

where $R_n$ is the sample correlation of $(Y, X_2)$ and $S_{n,Y}$, $S_{n,X_2}$ are the sample standard deviations of $Y$ and $X_2$, respectively. First note that the correlation is unit-free, while the sample standard deviations have the same units as the variables they're computed from.

Now let's interpret the coefficients above:

- The intercept $\widehat{\beta}_1$ is (literally) the average value of $Y_i$ after the average effect of $X_{i,2}$ has been removed. It has the same units as $Y_i$. It's interpretation is sometimes meaningful and sometimes not. Essentially it is the expected value of $Y_i$ if $X_{i,2} = 0$.

- The coefficient $\widehat{\beta}_2$ is proportional to the correlation between $Y_i$ and $X_{i,2}$, but adjusted according to the scales of the variables. The unit of $\widehat{\beta}_2$ is the ratio of units of $Y_i$ and $X_{i,2}$. It holds $\widehat{\beta}_2 = 0$ if and only if $R_n = 0$.[1] If $\widehat{\beta}_2 \neq 0$, the interpretation is as follows: if $X_{i,2}$ is increased by one unit, then $Y_i$ is expected to increase by $\widehat{\beta}_2$ units.

**Example 8.3.** *Consider again the color-magnitude diagram in Fig. 8.1. The straight line in the graph is in fact the OLS estimate which gives*

$$\widehat{\beta}_1 = 4.70, \qquad \widehat{\beta}_2 = 4.59.$$

[1] This should remind you of the fact that the correlation measures linear dependence.

*The coefficient $\widehat{\beta}_1$ is in the same units as $V$, i.e., magnitudes. It tells us that a star with (B-V) $= 0$ is expected to have a $V$-band magnitude of 4.7. The coefficient $\widehat{\beta}_2$ is unit-free, because $V$ and B-V have the same units. It tells us that, for an increase of 1 mag in B-V, we expect to see an increase of 4.59 mag in $V$.*

### 8.2.3 Confidence intervals and significant covariates

If you look closely to Fig. 8.1, you will see that there is a gray shaded area around the regression line. This area is a 95%-confidence region for the line. It is very narrow, because there are so many data ($n$) compared to the number of covariates ($p$). To construct these intervals, you can use the bootstrap or a closed-form expression derived from asymptotic normality.

**Theorem 8.4.** *Define $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(Y - \boldsymbol{\beta}^\top \boldsymbol{X})^2]$. Then the OLS $\widehat{\boldsymbol{\beta}}$ satisfies for all $j = 1, \dots, p$,*

$$\sqrt{n}\widehat{\Sigma}^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to_d \mathcal{N}(0, I_{p\times p}),$$

*where $I_{p\times p}$ is the $p \times p$ identity matrix and*

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)^2 \times \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1}.$$

The results follows from the multivariate CLT, but we won't bother with it. Note that $\widehat{\Sigma}$ is a $p \times p$ matrix, not a single number. The standard error for $\widehat{\beta}_j$ is computed as the $j$th diagonal element of $\widehat{\Sigma}^{1/2}/\sqrt{n}$.

Theorem 8.4 can be used for a Wald test for the effect of individual covariates:

$$H_0\colon \beta_j^* = 0, \qquad H_1\colon \beta_j^* \neq 0.$$

If we reject $H_0$, we say that covariate $j$ has a significant effect. Beware: if we test multiple covariates, we need to correct for multiple testing!

### 8.2.4 Fitted values and residuals

While $Y_i$ and $\boldsymbol{X}_i$ are observed, the error term $\epsilon_i$ in the equation

$$Y_i = \boldsymbol{\beta}^\top \boldsymbol{X}_i + \epsilon_i$$

is not. Given an estimate $\widehat{\boldsymbol{\beta}}$, our model predictions for $Y_i$ would be

$$\widehat{Y}_i = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i,$$

which are called the *fitted values*. The *regression residuals* are defined as

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i.$$

They residuals approximate the error term $\epsilon_i$ and are quite useful to check for model fit and misspecification.

The *residual sum of squares (RSS)* is defined as

$$\mathrm{RSS} = \sum_{i=1}^n \widehat{\epsilon}^2 = \sum_{i=1}^n (Y_i - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)^2,$$

and measures the quality of the fit. We already saw it pop up in the asymptotic variance in Theorem 8.4. If the RSS is small, it means that our model predictions are close to the observed values. However, the RSS depends crucially on the variance of $\epsilon$. If this variance is large, the RSS will be large, too. (Why?) A standardized version, called *R-squared*, is

$$R^2 = 1 - \frac{\sum_{i=1}^n \widehat{\epsilon}^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} = 1 - \frac{S^2_{n,\widehat{\epsilon}}}{S^2_{n,Y}}.$$

and measures the proportion of $Y$'s variability that is explained by the model. $R^2 = 0$ means that the model has no explanatory power; $R^2 = 1$ means that all observations can be explained perfectly ($\widehat{\epsilon}_i = 0$ for all $i$).

An issue with $R^2$ is that we can always improve by adding more covariates. This is fixed by the *adjusted* $R^2$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

which penalizes large $p$.

### 8.2.5 Implementation in Python

You will probably never need to implement any of the methodology yourself; any reasonable statistics software does this for you. In Python, you may use the `statsmodels` library. Let me shamelessly copy & paste an excerpt from the documentation:

```
  # Load modules and data
In [1]: import numpy as np
In [2]: import statsmodels.api as sm
In [3]: spector_data = sm.datasets.spector.load(as_pandas=False)
In [4]: spector_data.exog = sm.add_constant(spector_data.exog, prepend=False)

# Fit and summarize OLS model
In [5]: mod = sm.OLS(spector_data.endog, spector_data.exog)
In [6]: res = mod.fit()
In [7]: print(res.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     6.646
Date:                Fri, 21 Feb 2020   Prob (F-statistic):            0.00157
Time:                        13:59:19   Log-Likelihood:                -12.978
No. Observations:                  32   AIC:                             33.96
Df Residuals:                      28   BIC:                             39.82
```

```
20 Df Model:                           3
21 Covariance Type:            nonrobust
22 ==============================================================================
23                  coef    std err          t      P>|t|      [0.025      0.975]
24 ------------------------------------------------------------------------------
25 x1             0.4639      0.162      2.864      0.008       0.132       0.796
26 x2             0.0105      0.019      0.539      0.594      -0.029       0.050
27 x3             0.3786      0.139      2.720      0.011       0.093       0.664
28 const         -1.4980      0.524     -2.859      0.008      -2.571      -0.425
29 ==============================================================================
30 Omnibus:                        0.176   Durbin-Watson:                   2.346
31 Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
32 Skew:                           0.141   Prob(JB):                        0.920
33 Kurtosis:                       2.786   Cond. No.                         176.
34 ==============================================================================
```

The first three instructions import the libraries and some data set. The fourth instruction adds an intercept to the covariates (as the last element, because `prepend=false`). The fifth instruction specifies which variables are response and which are the covariates ($Y =$ `spector_data.endog`, $\boldsymbol{X} = ($`spector_data.exog`, $1)$). The sixth instruction, computes the OLS and the seventh instruction prints a summary of the fitted model.

There's more information in the output than you will normally need and more than what's covered here. So let me just point you to the important bits.

- The (adjusted) $R^2$ is given in lines 13–14 on the right.
- Below you'll find the log-likelihood (assuming Gaussian errors) and the model selection criteria AIC and BIC, which we'll cover later.
- In the table below (lines 23–28) you see everything related to parameter estimates. `x1`, `x2`, `x3` are the names of the (random) covariates, `const` refers to the intercept that we added to the model.
- The first column (`coef`) contains the estimated parameters $\widehat{\beta}_k$ followed by the standard error.
- The fourth column (`P>|t|`) is the $p$-value for $H_0\colon \beta_k^* = 0$ — not corrected for multiple testing!
- The last two columns are the lower and upper bounds of a 95% confidence interval. If you want another confidence level, you can compute your own from the standard errors in the second column.

### 8.2.6 Heteroscedasticity

In the model formulation (8.1), we made no assumption about the variance of $\epsilon$. To derive the OLS criterion from the normal distribution, we assumed that this variance is constant. *Heteroscedasticity* (another terrible name) refers to situations where the variance depends on $\boldsymbol{X}$, i.e., $\mathbb{V}[\epsilon \mid \boldsymbol{X}]$ is not constant.

Consider for example the model in Fig. 8.1. In some B-V-regions the residuals $\widehat{\epsilon}_i$ tend to be larger (in absolute terms). This is a common phenomenon, especially when $\widehat{\epsilon}_i$ is a measurement error. For example, measurements of redshift are generally less accurate for distant objects, which translates to a larger error variance.

While the OLS is still consistent under heteroscedasticity, it is not very efficient. Observations with large $\mathbb{V}[\epsilon_i \mid \boldsymbol{X}_i]$ are overweighted. This can be fixed if we know the error variance $\sigma_i^2 = \mathbb{V}[\epsilon_i \mid \boldsymbol{X}_i]$ for each observation or have an estimator $\widehat{\sigma}_i^2$. Then, we should use the *weighted least squares* criterion

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i)^2}{\widehat{\sigma}_i^2}.$$

## 8.3 Generalized linear models

### 8.3.1 The idea

The linear model Eq. (8.1) made no assumptions at all about the distribution of $Y$ or $\epsilon$. It works best when $\epsilon$ is at least approximately normal. If $Y$ has a different distribution, predictions from the model can be nonsensical, though. Take for example $Y \sim \text{Bernoulli}(p)$ and the model $Y = X_2 + \epsilon$. For large values of $X_2$, the model would predict values of $Y$ that are much larger than 1, which does not make sense. Similarly, if the distribution is such that $Y > 0$, the model might nevertheless predict negative values.

When we have a good guess for the (conditional) distribution of $Y$, a generalized version of the linear model is more appropriate. If $\epsilon \sim \mathcal{N}(0, \sigma^2)$, yet another formulation of (8.1) is that the conditional distribution $F_{Y|\boldsymbol{X}}$ is normal with parameter $\mu = \boldsymbol{\beta}^\top \boldsymbol{X}$ and variance $\sigma^2$. In short,

$$Y \mid \boldsymbol{X} \;\sim\; \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{X}, \sigma^2).$$

From a more abstract point of view, we assume that $F_{Y|\boldsymbol{X}}$ belongs to a parametric family where the parameter $\mu$ is a linear function of $\boldsymbol{X}$.

So why not just replace the normal distribution by another parametric family? That's the idea behind the class of *generalized linear models (GLMs)*. There is a minor caveat though. Take for example $Y \mid X \sim \text{Bernoulli}(p)$ with $p = \boldsymbol{\beta}^\top \boldsymbol{X}$. We still have the issue that $\boldsymbol{\beta}^\top \boldsymbol{X}$ may fall outside of $(0, 1)$. This is easily fixed by inserting a *link function* $g\colon \mathbb{R} \to \mathcal{P}$, where $\mathcal{P}$ is the set of allowed parameter values. In the classical linear regression model, we have the *identity link*, $g(x) = x$.

### 8.3.2 Popular GLMs

This idea can be used with all the parametric families we've seen so far. Let's consider a few examples:

- *Logistic regression:* $Y \mid X \sim \text{Bernoulli}(p)$ with $p = g(\boldsymbol{\beta}^\top \boldsymbol{X})$ and the logistic link

$$g(x) = \frac{e^x}{1 + e^x}.$$

Instead of regression, we often call this a *classification model*, because it models the probability of class-membership, e.g., radio-loud vs radio-quite, star-forming or not, etc.

- *Binomial regression:* $Y \mid X \sim \text{Binomial}(p, n)$ with fixed $n$, $p = g(\boldsymbol{\beta}^\top \boldsymbol{X})$ and the logistic link.
- *Poisson regression:* $Y \mid X \sim \text{Poisson}(\lambda)$ with $\lambda = g(\boldsymbol{\beta}^\top \boldsymbol{X})$ and $g(x) = e^x$. This is called *log-link*, which refers to the inverse $g^{-1}(x) = \ln(x)$.
- *Exponential regression:* $Y \mid X \sim \text{Exp}(\lambda)$ with $\lambda = g(\boldsymbol{\beta}^\top \boldsymbol{X})$ and log-link.
- *Gamma regression:* We assume $Y \mid X \sim \text{Gamma}(\alpha, \nu)$ with $\nu$ not depending on $\boldsymbol{X}$ and $\alpha = g(\boldsymbol{\beta}^\top \boldsymbol{X})/\nu$. The default link function in most software is the *inverse link* $g(x) = 1/(\beta^\top \boldsymbol{X})$, but that doesn't really enforce the right parameter values. I recommend to use the log-link here too.

This is only a small sample from an extremely rich class of models. For example, one can play with other link functions or let both parameters of the Gamma family vary with $\boldsymbol{X}$.

### 8.3.3 Parameter estimation

The MLE can be used to estimate the parameters. Let

$$\mathcal{F} = \left\{ f_{g(\boldsymbol{\beta}^\top \boldsymbol{X}),\eta} \colon \boldsymbol{\beta} \in \mathbb{R}^p, \eta \in \mathcal{E} \right\}$$

be the statistical model for the conditional distribution of $Y$ given $\boldsymbol{X}$. Here, $\eta$ refers to all parameters in the model that are not a function of $\boldsymbol{X}$. Then the log-likelihood is

$$\ell_n(\boldsymbol{\beta}, \theta) = \sum_{i=1}^n \ln f_{g(\boldsymbol{\beta}^\top \boldsymbol{X}_i),\eta}(Y_i),$$

and all results from Section 5.4 apply.

### 8.3.4 Implementation in Python

The GLMs mentioned above are just easily implemented as the linear model. Again from the `statsmodels` documentation:

```
  # Load modules and data
In [1]: import statsmodels.api as sm
In [2]: data = sm.datasets.scotland.load(as_pandas=False)
In [3]: data.exog = sm.add_constant(data.exog)
```

```
 5
 6 # Instantiate a gamma family model with the default link function.
 7 In [4]: gamma_model = sm.GLM(data.endog, data.exog, family=sm.families.Gamma())
 8 In [5]: gamma_results = gamma_model.fit()
 9 In [6]: print(gamma_results.summary())
10                  Generalized Linear Model Regression Results
11 ==============================================================================
12 Dep. Variable:                      y   No. Observations:                   32
13 Model:                            GLM   Df Residuals:                       24
14 Model Family:                   Gamma   Df Model:                            7
15 Link Function:          inverse_power   Scale:                       0.0035843
16 Method:                          IRLS   Log-Likelihood:                -83.017
17 Date:                Fri, 21 Feb 2020   Deviance:                     0.087389
18 Time:                        13:59:13   Pearson chi2:                   0.0860
19 No. Iterations:                     6
20 Covariance Type:            nonrobust
21 ==============================================================================
22                  coef    std err          z      P>|z|      [0.025      0.975]
23 ------------------------------------------------------------------------------
24 const         -0.0178      0.011     -1.548      0.122      -0.040       0.005
25 x1          4.962e-05   1.62e-05      3.060      0.002    1.78e-05    8.14e-05
26 x2             0.0020      0.001      3.824      0.000       0.001       0.003
27 x3         -7.181e-05   2.71e-05     -2.648      0.008      -0.000   -1.87e-05
28 x4             0.0001   4.06e-05      2.757      0.006    3.23e-05       0.000
29 x5         -1.468e-07   1.24e-07     -1.187      0.235   -3.89e-07    9.56e-08
30 x6            -0.0005      0.000     -2.159      0.031      -0.001   -4.78e-05
31 x7         -2.427e-06   7.46e-07     -3.253      0.001   -3.89e-06   -9.65e-07
32 ==============================================================================
```

There are only minor differences to the example we've seen above. The code above sets up a Gamma regression model (fourth instruction) instead of a linear model. The model summary contains a bit less information, but the important parts are still there. The estimate $\widehat{\nu}$ of the fixed parameter is given as `Scale` in line 15, right column.

## 8.4 Non-linear models

Consider again the color-magnitude diagram in Figure 8.1. The linear model looks like an OK approximation, but we also see some systematic deviations. For very low and large values of the B-V-index, most points fall below the regression line. For medium values of the index, most points fall above the regression line. So it seems that the relationship is not exactly linear.

One might argue that nothing's ever truly linear. It then depends on the scientific goals whether it's worthwhile to capture nonlinearities explicitly. In the following we learn how this can be done within the framework of (generalized) linear models.

In non-linear models, we assume that the true relationship is characterized by a function $h(\boldsymbol{X})$. For example, our model may take the form

$$Y = h(\boldsymbol{X}) + \epsilon, \qquad \text{or} \qquad Y \mid \boldsymbol{X} \sim F_{g(h(\boldsymbol{X})),\eta}.$$

We wish to estimate the function without making restrictive assumptions (like linearity). For simplicity, we shall mostly restrict ourselves to the simpler model on the left. Everything transfers naturally to the generalized non-linear model on the right.

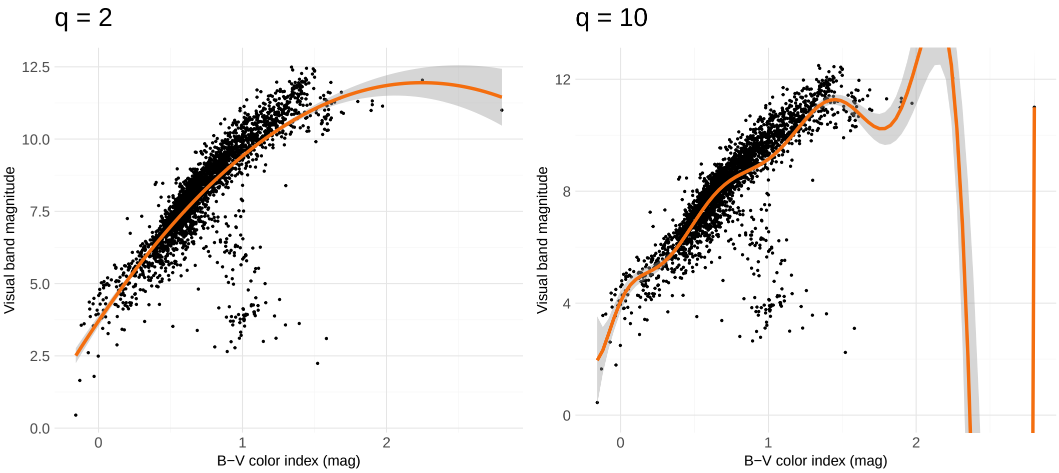


Figure 8.2: Non-linear regression model for the color-magnitude diagram of 2655 stars from the Hipparcos catalog using polynomial expansions of order 2 (left) and 10 (right).

### 8.4.1 Polynomial expansions

Earlier in this chapter, I already mentioned a simple way to model non-linear relationships. For sake of simplicity, we shall assume that there is only a single covariate $X$. Now define $\boldsymbol{Z} = (1, X, X^2, \dots, X^q)$ and use this as the new vector of covariates. For example, the linear model (8.1) becomes

$$Y = \boldsymbol{\beta}^\top \boldsymbol{Z} + \epsilon = \beta_1 + \beta_2 X + \cdots + \beta_{q+1} X^q + \epsilon. \tag{8.3}$$

Instead of assuming a linear relationship between $Y$ and $X$, we now assume that the relationship is a polynomial of (up to) order $q$. The intuition behind this is Taylor's theorem. If the function $h$ is sufficiently smooth, it can be well approximated by a polynomial of some order. This also works for expansions in more than one variable.

Similarly, we may replace an initial covariates $X$ in a GLM with an extended vector $\boldsymbol{Z}$. Since we merely replaced the covariate vector by another, the OLS/MLE methods and related theory still work when assuming a polynomial relationship.

Fig. 8.2 shows the OLS fit of the extended linear model with a quadratic ($q = 2$) and tenth-order ($q = 10$) polynomial (right). Clearly, our models now predict a non-linear relationship between the V-band magnitude and color index. We see less systematic deviations, but we also see that the confidence intervals become wider as we increase $q$. This is a general rule: the more complex we make the model, the more uncertainty we have in parameter estimates.

We also see that the $q = 10$ curve behaves rather erratically, especially in areas where there are little observations. The reason is that we have *too much* flexibility in the model, so that we can trace the points at the right boundary almost perfectly. Of course, we should not trust the model with $q = 10$ unless there are good physical reasons to believe in such erratic behavior. There's always

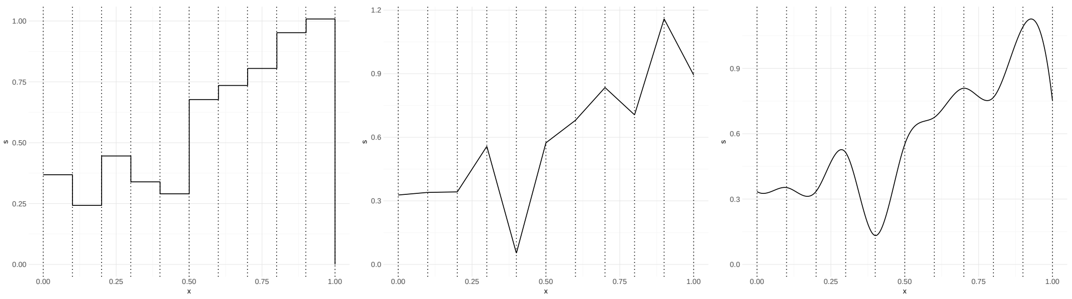


Figure 8.3: Examples of spline functions of degrees 0 (left), 1 (middle), and 3 (right).

a sweet spot for $q$, but unfortunately it's hard to know where it is in advance. Later, we'll discuss methods on how to find this sweet spot.

## 8.4.2 Spline expansions

Polynomial expansions are simple, but have known issues. One is that the function $h$ must be very smooth for Taylor's theorem to apply. A second is that polynomial expansions converge rather slowly to the function they're supposed to approximate. A smarter way to approximate $h$ are *spline expansions*.

### Spline functions

Splines are functions that are composed piece-wise from polynomials.

**Definition 8.5.** *Let $[a, b]$ be some interval. A* ***spline of degree q*** *is a function $s$ defined on a grid of* ***knots*** *$a = \xi_0 < \xi_1 < \ldots < \xi_m = b$, such that*

$$s(x) = P_k(x), \qquad \text{for } x \in [\xi_{k-1}, \xi_k),$$

*where $P_k$, $k = 1, \ldots, m$, are polynomials of order $q$. Additionally, we require that the $(q-1)$th derivative is continuous.*

Less formally, a spline is a function that 'stitches' together $m$ different polynomials defined on sub-intervals.

Fig. 8.3 shows spline functions of degree $q = 0$ (left), $q = 1$ (middle), and $q = 3$ (right). In all graphs, the interval [0, 1] is split into 10 sub-intervals, indicated by vertical dashed lines. On each of these intervals, the function is simply a polynomial of degree $q$. On the left ($q = 0$), the function is constant on every interval; in the middle ($q = 1$), it is linear on every interval; on the right ($q = 3$) it is cubic on every interval. Cubic splines are used most commonly and should be the default choice.

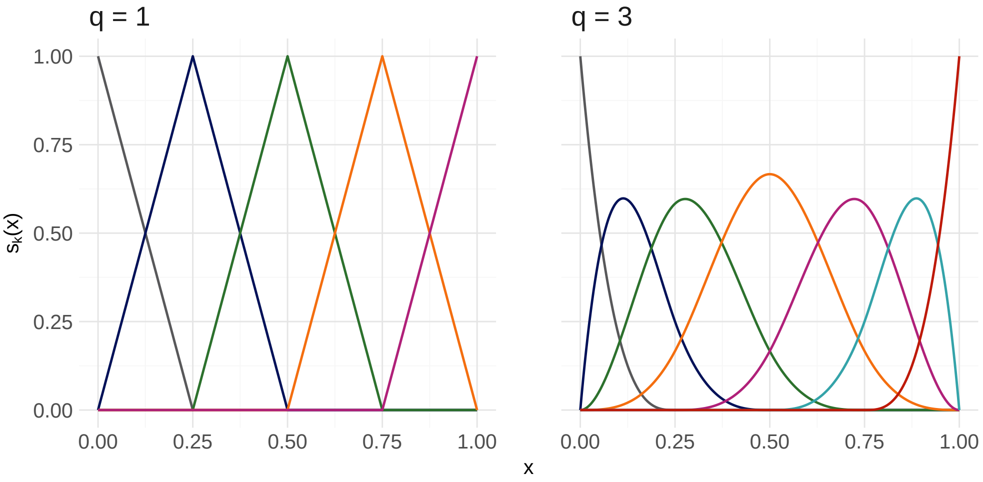


Figure 8.4: Spline basis functions on 5 knots for degrees 0 (left) and 3 (right).

### Spline basis

So let's get back to the statistical problem, estimating the unknown function $h$ in (8.3). If we assume that $h$ is a spline on $m$ knots, we now need to estimate 10 polynomials of degree $q$, giving $m \times q$ coefficients. That doesn't sound very convenient. But hold on: we also have this little side constraint of continuous derivatives. Quite remarkably, this seemingly innocent assumptions reduces our degrees of freedom a lot. Even more remarkably, any spline can be represented conveniently as a linear combination of just $m + q + 1$ known *basis functions* $B_j$:

$$s(x) = \sum_{j=0}^{q+m} \beta_j B_j(x),$$

where

$$\begin{aligned} B_j(x) &= x^j, & j &= 0, \dots, q, \\ B_{q+1+j}(x) &= \max\{0, (x - \xi_j)\}^q, & j &= 0, \dots, m-1, \end{aligned}$$

and $\beta_j$ are unknown coefficients (to be estimated). Other forms of the basis exist, but the number of functions stays the same. You can create such a basis in Python using `statsmodels.gam.smooth_basis.BSplines`. Such basis functions $B_j$ for $q = 1, 3$ and $m = 4$ are shown in Fig. 8.4, where each color corresponds to a different $j$. A spline function is simply a linear combination of these functions.

For a single covariate $X \in [a, b]$, we can therefore assume

$$h(X) \approx s(X) = \sum_{j=0}^{q+m} \beta_j B_j(X),$$

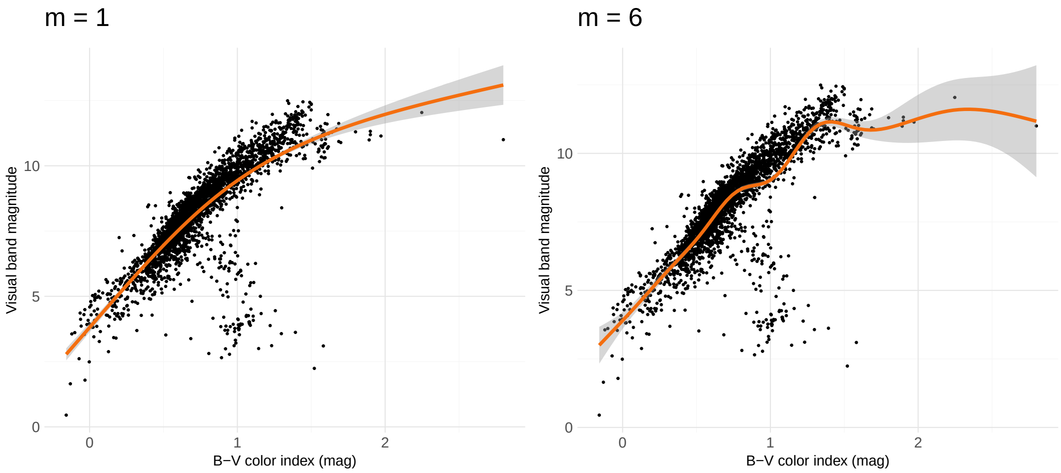


Figure 8.5: Spline regression model for the color-magnitude diagram of 2655 stars from the Hipparcos catalog using cubic splines ($q = 3$).

or

$$Y \approx \sum_{j=0}^{q+m} \beta_j B_{j,q}(X) + \epsilon$$

which is just a linear model

$$Y \approx \boldsymbol{\beta}^\top \boldsymbol{Z} + \epsilon$$

on the transformed covariates

$$\boldsymbol{Z} = (B_0(X), \dots, B_{q+m}(X)) \in \mathbb{R}^{q+m+1}.$$

Since the basis functions $B_j$ are known, the coefficients $\boldsymbol{\beta}$ can be estimated easily with OLS or MLE. An example is shown in Fig. 8.5, where we fit cubic splines with $m = 1, 6$ to the Hipparcos data. We observe that splines tend to be more stable than the polynomials in the prevous section. The spline with $m = 6$ has 10 parameters, but produces a very reasonable model (in contrast to the polynomial of 10th order.)

### The bias-variance trade-off

Although the usual OLS/MLE results apply, we're not really interested in how well the coefficients are estimated, but rather how well the function $h$ is estimated. Define

$$\widehat{h}(x) = \sum_{j=0}^{q+m} \widehat{\beta}_j B_j(x)$$

as our estimated function.

One can prove the following:

**Theorem 8.6.** *Suppose that* $|\xi_k - \xi_{k-1}| = 1/m$ *for all* $k = 1, \dots, m$. *Under some regularity conditions*

$$\mathbb{E}[\widehat{h}(x)] = h(x) + O(m^{-(q+1)}), \qquad \mathbb{V}[\widehat{h}(x)] = O\left(\frac{m}{n}\right), \quad \textit{for all } x \in [a, b].$$

**Remark 8.3.** *The* $O$*-symbol is to be read as 'is of the same order as'. Formally, for two sequences* $a_n, b_n$, $a_n = O(b_n)$ *means* $\lim_{n\to\infty} |a_n/b_n| < \infty$. *More intuitively, it says that if* $b_n \to 0$, *then also* $a_n \to 0$ *at least as fast.*

Let's take this apart.

- We assume that the distance between any two subsequent knots $\xi_{k-1}$ and $\xi_k$ is the same. That is, the spline is defined on intervals of the same length. (This is only to simplify result, it's by no means necessary in practice.)
- The 'regularity conditions' are very mild. The main assumption is that the true function $h$ is a few times continuously differentiable. This just excludes cases where $h$ goes completely wild.
- The bias behaves like $m^{-(q+1)}$, which is decreasing in $m$. If $m$ is fixed, then the estimator is usually biased. But if we take $m \to \infty$, the bias vanishes. Moreover, the bias vanishes faster for larger $q$. We conclude that we prefer large $m$ to make the bias small.
- The variance is of order $m/n$. If $m$ is fixed, the variance goes to zero as $n \to \infty$. However, the variance is increasing in $m$, so we prefer small $m$ to keep the variance small.

The last two bullets describe a ubiquitous phenomenon in function estimation (as opposed to parameter estimation). We have a *tuning parameter* (here $m$) that controls a trade-off between bias and variance. If we decrease the variance, we increase the bias; if we decrease the bias, we increase the variance.

Gladly, large sample sizes $n$ reduce the variance. We can therefore afford larger values of $m$ when there's a lot of data. Again there's a sweet spot that balances bias variance in an optimal way. One can show that the mean squared error is optimal if $m$ increases at the order $n^{1/(2q+3)}$, but that's not helpful in practice. We'll see how to solve this shortly.

**Remark 8.4.** *One can mathematically prove that the bias-variance trade-off is unavoidable when estimating regression or density functions. For example, the same phenomenon appears for the histogram, where more bins decrease bias, but increase variance.*

**Remark 8.5.** *Just so you know: there is another popular way to control the bias-variance tradeoff for splines. Here, we take a large number of knots $m$, but put a penalty on the magnitude of coefficients $\beta_j$. This is called a* penalized spline *and the strength of the penalty is controlled by a parameter $\alpha \geq 0$. For $\alpha = 0$ there is no penalty, and for $\alpha = \infty$, all coefficients are $\beta_j = 0$.*

### Generalized Additive Models (GAMs)

So far, we only considered the case of a single covariate $X$. Let's briefly discuss two methods to handle multi-dimensional $\boldsymbol{X} \in \mathbb{R}^p$. The first, is to assume that

$$h(\boldsymbol{X}) = \sum_{k=1}^{p} h_k(X_k), \tag{8.4}$$

where each $h_k$ is a spline function. This is called an *additive* model: we assume that the the multi-dimensional function $h$ can be decomposed into a sum of one-dimensional functions $h_k$. The *generalized additive model* (GAM) takes a parametric model $\{F_{\theta,\eta} \colon \theta \in \Theta, \eta \in \mathcal{E}\}$ and postulates

$$Y \mid \boldsymbol{X} \sim F_{g(h(\boldsymbol{X})),\eta},$$

where $h$ is as in (8.4) and $g$ is an appropriate link. Such models are still quite easy to estimate and interpret, because the functions $h_k$ can be treated separately. Don't worry about the details of estimation, these models have great implementations (e.g., `statsmodels.gam`).

### Tensor product splines

The additivity assumption (8.4) can also be relaxed. A $p$-dimensional version of the spline is the *tensor product spline*:

$$s(\boldsymbol{x}) = \sum_{j_1=0}^{m+q} \cdots \sum_{j_p=0}^{m+q} \beta_{j_1,\dots,j_p} B_{j_1,q}(x_1) \cdots B_{j_p,q}(x_p),$$

where $B_{j_p,q}$ are the basis functions from before. Tensor product splines can approximate all continuous, $p$-dimensional functions with arbitrary accuracy. However, the model has a $(m+q+1)^p$ parameters to estimate and is much harder to interpret. As a rule of thumb, tensor product splines are only useful when $p \leq 3$.

### More on non-linear models

Non-linear and nonparametric regression is a huge and very active field of research. We've only seen a small glimpse of it, so I want to mention a few popular alternatives:

- Kernel methods estimate non-linear functions by (roughly) taking clever weighted averages.
- Support vector methods use different types of expansions of the covariates with appropriate penalties on the parameters.
- Neural networks essentially nest multiple GLMs into each other.

The principles we learned in this course also apply to these models, but require more advanced mathematics. If you want to learn more about splines or GAMs, there is an excellent book by Simon Wood "Generalized Additive Models: An Introduction with R".

## 8.5 Model selection

By now you've learned about many different models: linear models, GLMs, non-linear models. For each of these models, you can additionally decide whether to include or exclude some of the covariates $(X_1, \dots, X_p)$. So how to pick the final model?

In any case, you should use your best judgement to rule out certain models. For example, if $Y$ is binary, a Gamma regression model doesn't make sense. If a covariate $X_j$ is binary, it doesn't make sense to expand it with polynomials or splines. If you know from the physics of your problem that $X_j$ influences $Y$, you should include it in the model. If you have only 20 data points, don't expand covariates to the 1000th order.

Usually, your judgement only gets you so far and you are left with a number of sensible models. *Model selection* refers to statistical procedures that help you make a final choice with a theoretical foundation. To set up the stage, we denote the candidate models $\mathcal{M}_1(\theta_1), \dots, \mathcal{M}_K(\theta_K)$, where $\theta_k$ is the collection of all parameters that characterize model $\mathcal{M}_k$. Everything that follows will be phrased in this abstract context, so let's consider a few concrete examples before we continue.

**Example 8.7.** *Suppose you want to choose between a Gamma-GLM $F^{\Gamma}_{g(\boldsymbol{\beta}^\top \boldsymbol{X}),\nu}$ and a Gaussian GLM $F^{\mathcal{N}}_{g(\boldsymbol{\beta}^\top \boldsymbol{X}),\sigma^2}$ (which is just usual the linear model). Then*

$$\mathcal{M}_1(\boldsymbol{\beta}_1, \nu)\colon F^{\Gamma}_{g(\boldsymbol{\beta}_1^\top \boldsymbol{X}),\nu}, \qquad \mathcal{M}_2(\boldsymbol{\beta}_2, \sigma^2)\colon F^{\mathcal{N}}_{g(\boldsymbol{\beta}_2^\top \boldsymbol{X}),\sigma^2}.$$

**Example 8.8.** *Consider the linear regression model $Y = \boldsymbol{\beta}_1^\top \boldsymbol{X} + \epsilon$. Suppose you want to use only one covariate, but you don't know which. Define*

$$\mathcal{M}_k(\boldsymbol{\beta}_k)\colon Y = \beta_{k,0} + \beta_{k,1} X_k + \epsilon, \quad k = 1, \dots, p.$$

*The extension to subsets of $\boldsymbol{X}$ of size larger than one is straightforward.*

**Example 8.9.** *Consider the spline GLM $F_{g(s_m(X)),\eta}$, where $m$ is the number of knots the spline function $s_m$ is defined on. To choose the tuning parameter $m$, define*

$$\mathcal{M}_k(\boldsymbol{\beta}_k)\colon F_{g(s_k(X)),\eta}, \quad k = 1, \dots, K,$$

*where $\boldsymbol{\beta}_k$ are the spline basis coefficients we need to estimate.*

Of course, all the examples above can be combined: you might want to choose between different types of GAMs, the covariates to include, and the smoothing parameter at the same time.

The two most popular criteria for model selection are *Akaike's information criterion (AIC)* and the *Bayesian information criterion (BIC)*. Both are based on the likelihood of a model. For model $\mathcal{M}_k$, let

- $\widehat{\theta}_k$ be the MLE of the model parameters $\theta_k$,
- $\ell_k(\widehat{\theta}_k)$ be the corresponding log-likelihood,
- $p_k$ the number of parameters of the model,
- $n$ the sample sized used for estimation.

Then

$$\begin{aligned} \mathrm{AIC}(\mathcal{M}_k) &= -2\ell_k(\widehat{\theta}_k) + 2p_k, \\ \mathrm{BIC}(\mathcal{M}_k) &= -2\ell_k(\widehat{\theta}_k) + \ln(n)p_k. \end{aligned}$$

Clearly, a good model has a large likelihood, so we want AIC or BIC to be as small (negative) as possible. The second term in the criteria is a penalty for model complexity. The intuition is that the more parameters we add to our model, the better is the likelihood we can achieve. But we don't want to include any unnecessary parameters and the penalties take care of that. The best model according to AIC/BIC is then the one that minimizes AIC/BIC. If $n > 8$, the BIC penalizes the number of parameters more strongly than AIC. Hence, the BIC tends to select less complex models.

Both criteria can be formally shown to select the "best" model, but they differ in what they see as best. Without going into too much detail: as $n \to \infty$ and $p_k \le \sqrt{n}$,

- AIC selects the best predictive model among a number of possibly misspecified models.
- BIC selects the true model (with minimal number of necessary parameters) if it is included in the candidate set.

So as a general rule of thumb: if the main goal is prediction, use AIC; if the main goal is identification of the truth, use BIC.

The two criteria can be used to select arbitrary statistical models with a likelihood, not just regression models. Furthermore, for the linear model, one commonly replaces $\ell_k$ by the residual sum of squares, which is equivalent to assuming Gaussian errors.

**Remark 8.6.** *There's one caveat when model parameters are not estimated by plain maximum-likelihood (like penalized splines). Then there is something called* effective number of parameters *or* effective degrees of freedom *that needs to be substituted for $p_k$ in the formulas above. Software usually takes care of that for you.*

# 9 Missing data

Missing data refers to situations where some of the objects or quantities that we measure are not or only partially observed. Missing data can be a problem if the observed sample gives a biased view on the whole population. Quite often, however, missingness can be accounted for by careful statistical modeling. This chapter gives an overview over different types of missingness and methods to address them.

## 9.1 Setup and notation

Assume that we analyze data drawn from a population with distribution $F_Y$ and density $f_Y$. Denote by $Y_1, \dots, Y_n$ the sample that we would have observed if there were no missing data. For each of these observations, denote $I_i = \mathbb{1}(Y_i \text{ is fully observed})$. We shall clarify later what we mean by *partially* observed data. For the moment you may assume that $I_i = 0$ means that we don't see the $Y_i$ at all.

## 9.2 Types of missing data

### Missing completely at random (MCAR)

MCAR refers to situations where $I_i$ is independent from $Y_i$. This means that whether or not we observe $Y_i$ has nothing to do with the actual value of $Y_i$.

**Example 9.1.** *Suppose an instrument is measuring short bursts of light occuring at random times. Each light burst takes around 0.1s, but the instrument is only measuring at a frequency of 1Hz. All bursts between measurements are missing.*

**Example 9.2.** *There's a large data base of nearby stellar objects and the survey is known to be complete. To get a sense of the data, you extract a random subset of the observations. All objects not in this subset are missing.*

This turns out to be the (rare) best case scenario. If data is MCAR, the usual statistical procedures remain valid.

As an instructive example, suppose we want to estimate the mean $\mathbb{E}[Y]$. The complete-data estimator would be just the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Let's assume only $m = \sum_{i=1}^n I_i < n$ of the data are actually observed. If we apply the sample mean to the incomplete data set, we get

$$\tilde{Y} = \frac{\sum_{i=1}^n Y_i I_i}{\sum_{i=1}^n I_i} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_i}{\frac{1}{n}\sum_{i=1}^n I_i} \to_p \frac{\mathbb{E}[Y_i I_i]}{\mathbb{E}[I_i]}, \tag{9.1}$$

where the last step follows from applying the law of large numbers to the numerator and denominator separately. Because $Y_i$ and $I_i$ are independent, $\mathbb{E}[Y_i I_i] = \mathbb{E}[Y_i]\mathbb{E}[I_i]$ and, thus, $\tilde{Y} \to_p \mathbb{E}[Y]$.

The same holds true for essentially all statistical methods, including tests and regression models. If we believe data is MCAR, there's no reason to worry about it any further. Unfortunately, that's rarely the case.

### Missing not and random (MNAR)

MNAR data is the worst case. It refers to situations where $Y_i$ and $I_i$ are *not* independent: the value of $Y_i$ influences whether or not we observe it.

**Example 9.3.** *Suppose a survey is measuring stellar mass, but cannot detect masses smaller than 1/100 of the sun's mass. Less massive stars are missing from the survey and massive stars are overrepresented.*

**Example 9.4.** *Suppose we measure Beryllium abundances in stars. Due to technical limitations the instrument can only measure abundances up to three times the sun's abundance. The observed data will then be truncated at this value. This is an example of partially observed data.*

MNAR is problematic because it leads to a biased sample. Consider again the sample mean based on incomplete observations. The computation in (9.1) remains valid. But because $Y_i$ and $I_i$ are dependent, we cannot simplify $\mathbb{E}[Y_i I_i] = \mathbb{E}[Y_i]\mathbb{E}[I_i]$. Consequently, we may (and usually will) have

$$\frac{\mathbb{E}[Y_i I_i]}{\mathbb{E}[I_i]} \neq \mathbb{E}[Y_i].$$

For example, when large $Y_i$ are less likely to be observed, we will underestimate the true mean.

MNAR is called *non-ignorable* because we have to do something about it to obtain valid inferences. More precisely, we have to come up with a model for mechanism leading to missing data. In general, this mechanism is not *identifiable*, meaning that it cannot be estimated from the observed data. To make (approximately) valid statistical inferences nevertheless, we need to come up with a model for the missingness mechanism. Optimally, we know a thing or two about how data are collected and can use domain knowledge to model the mechanism. If that's not the case, the best we can do is making educated guesses and be very careful in drawing conclusions from the results.

### Missing at random (MAR)

MAR is somewhere in between MCAR and MNAR and assumes having complete observations of additional covariates $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ (the 'complete' is important). MAR means that the probability of $Y_i$ being missing depends only on the fully observed covariates. More formally, data is MAR if $Y_i$ is independent of $I_i$ given $\boldsymbol{X}_i$: for all $y \in \mathbb{R}, i \in \{0, 1\}, \boldsymbol{x} \in \mathbb{R}^p$,

$$\mathbb{P}(Y_i \le y, I_i = i \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y_i \le y \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{P}(I_i = i \mid \boldsymbol{X} = \boldsymbol{x}).$$

**Example 9.5.** *Suppose a study estimates exoplanet masses from optical photometry data. If the planet doesn't emit light for accurate photometric measurements, it's mass is marked as missing. Here, only the photometric measurements determine missingness in mass. Thus, conditionally on the photometric outcome, the value of the mass and whether we observe it are independent.*

MAR is less problematic than MNAR, because the missingness mechanism is identifiable. More concretely, we can estimate a logistic regression model with response $I_i$ and covariates $Y_i$. This model tells us everything we need to know about the probabilities of an instances $Y_i$ being missing or not. As we shall see, this information can be used to correct statistical procedures for missingness.

## 9.3 Censoring and truncation

Censoring and truncation are mechanisms leading to *partially observed data*. Partially observed means: even though the actual value of $Y_i$ is not observed, we instead see another value $\tilde{Y}_i$ that carries some information about $Y_i$.

We call an observation $Y_i$ *right-truncated*, if all we see is $\tilde{Y}_i = \min\{Y_i, c\}$ for some fixed value of $c \in \mathbb{R}$. It is called *left-truncated* if we see $\tilde{Y}_i = \max\{Y_i, c\}$. It is called *doubly truncated* if $\tilde{Y}_i = \max\{c_1, \min\{Y_i, c_2\}\}$. Truncation appears frequently due to technical limitations of measurement instruments. It is unproblematic if we know the truncation value $c$. Then, the missingness mechanisms is known and statistical methods can be adjusted rather easily. Truncation appears frequently due to technical limitations of measurement instruments, see Example 9.4. It is unproblematic if we know the truncation value $c$. Then, the missingness mechanisms is entirely known and statistical methods can be adjusted rather easily.

*Censoring* is similar to truncation, but with random truncation point $C_i$. For example, $Y_i$ is right-censored if we observe $\tilde{Y}_i = \min\{Y_i, C_i\}$ and similarly for left- and doubly censored data. It's also possible to correct for this type of missingness if we additionally observe a censoring indicator $I_i = \mathbb{1}(Y_i \le C_i)$.

**Example 9.6.** *Censoring occurs most commonly when $Y_i$ is a time of some event of interest; for example the time until light is reflected back to earth. We can wait only a finite amount of time until this happens. If the light has not been reflected* *at that time, all we know is that (i) it has not yet been reflected back ($I_i = 0$), (ii) that the actual time $Y_i$ must be larger than $\tilde{Y}_i$ = time between emission from earth and the end of the study.*

## 9.4 Correcting for missingness

Dealing with missing data has a long history in statistics and there are many techniques to account for missing data. They can be broadly classified into three categories:

1. Likelihood methods: explicit modifications of the likelihood function, often involving integration.

2. Imputation: Missing observations $Y_i$ are substituted with 'plausible' values $Y_i^*$ generated from some model.

3. Inverse probability weighting (IPW): Complete observations are upweighted by the inverse probability of them being observed. (That's what you did in the assignment.)

We shall only consider the third category here, because it is both easy to implement and very general.

Let's start with a general setup. Suppose we have a model for the probability $\pi(Y_i, \boldsymbol{X}_i) = \mathbb{P}(I_i = 1 \mid Y_i, \boldsymbol{X}_i)$. Here, $I_i = \mathbb{1}(Y_i \text{ is fully observed})$ such that $I_i = 0$ if $Y_i$ is not or only partially observed. If no covariates $\boldsymbol{X}_i$ are available, they can be omitted in the formulas, i.e., $\pi(Y_i, \boldsymbol{X}_i) = \pi(Y_i)$.

The idea is as follows: first, we throw away all incomplete observations (including partially observed ones). Now for all $y, \boldsymbol{x}$ with $\pi(y, \boldsymbol{x}) < 1$, observations with $(Y_i, \boldsymbol{X}_i) = (Y, \boldsymbol{x})$ are underrepresented in the remaining data; a complete data set would contain $1/\pi(y, \boldsymbol{x})$ times more of such observations. We correct for this by up-weighting the complete observations by this factor. This assumes that the quantity we compute is a sum or average. But as I've said earlier in the semester: almost everything in statistics is an average or well approximated by one.

To make this more concrete, reconsider the example of estimating the mean $\mathbb{E}[Y_i]$ from incomplete (MNAR or MAR) data. The IPW version of the sample mean is

$$\bar{Y}^{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i I_i}{\pi(Y_i, \boldsymbol{X}_i)}.$$

The $I_i$ in the numerator is responsible for "throwing away all incomplete data". The $\pi(Y_i, \boldsymbol{X}_i)$ in the numerator is up-weighting the complete cases. By the law of large numbers, we have

$$\bar{Y}^{IPW} \to_p \mathbb{E}\left[\frac{Y_i I_i}{\pi(Y_i, \boldsymbol{X}_i)}\right].$$

Using the Tower rule, we get

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_i I_i}{\pi(Y_i, \boldsymbol{X}_i)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{Y_i I_i}{\pi(Y_i, \boldsymbol{X}_i)} \mid Y_i, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\frac{Y_i}{\pi(Y_i, \boldsymbol{X}_i)}\mathbb{E}\big[I_i \mid Y_i, \boldsymbol{X}_i\big]\right] \\
&= \mathbb{E}\left[\frac{Y_i}{\pi(Y_i, \boldsymbol{X}_i)}\pi(Y_i, \boldsymbol{X}_i)\right] \\
&= \mathbb{E}[Y_i].
\end{aligned}$$

The second equality is due to the fact that $Y_i, \boldsymbol{X}_i$ are fixed numbers if we condition on $Y_i, \boldsymbol{X}_i$. Hence, the IPW version of the sample mean is a consistent estimator.

The same arguments apply whenever we rely on estimating one or more expectations of the form $\mathbb{E}[g(Y_i)]$ for some function $g(Y_i)$. That covers almost everything we learned in this course, including empirical CDFs, histograms, sample quantiles, sample variances, maximum-likelihood estimators, etc. In practice, the probabilities $\pi(Y_i, \boldsymbol{X}_i)$ are rarely known. If we observe the indicator $I_i$, we can estimate them with a regression model. If the indicator $I_i$ is unobserved, we have to use domain knowledge and EDA to postulate a plausible model.

## 9.5 Takeaways

1. Missing data problems are common and it is important to think about them carefully. What type of missingness do you face? What is the reason/mechanism for incomplete observations?

2. There is a large arsenal of well-established methods to account for missing data. As soon you have identified what type you are facing, search for a method that's suitable for your specific problem. IPW will work in most cases, but sometimes it's worth investigating other options.

# 10 Bayesian inference

This chapter gives a very brief introduction to *Bayesian inference*. Bayesianism is an entire school of statistics and deserves its on course. However, Bayesian methods have become quite popular in astronomy, so it's worthwhile understanding the basics. The goal of this chapter is to become familiar with the core ideas and concepts. You can learn more about Bayesian methods in the MSc course *Modern Astrostatistics*.

## 10.1 Frequentism vs. Bayesianism

The statistical methods and related results we've seen so far all operate under the *frequentist* paradigm. They focus on long-run frequencies of events. For example, consistency requires long-run frequencies of "being away from the true parameter" to vanish. Confidence intervals are set such "they cover the true parameter" with prescribed frequency. Statistical procedures are then designed to adhere to these targets. You may notice that I used the term 'frequency' instead of 'probability' in the last sentences. And that's the clue:

> *For a frequentist, probabilities are the long-run limit of frequencies.*

We already faced an important consequence in Section 5.4.1 and Section 6.5: because the true parameters of our probability model are fixed numbers, we cannot make meaningful probabilistic statements about them. They attain their true values with frequency/probability 100% and any other value with frequency/probability 0%.

The Bayesian paradigm is fundamentally different.

> *For a Bayesian, probabilities express degree of belief.*

An immediate consequence is that a Bayesian *can* make meaningful probability statements about unknown parameters — even before having seen any data. If I ask you how likely it is that I woke up before 8 AM today, you may answer 70%. That's a Bayesian probability: Either I woke up before 8 AM (100% frequency) or I didn't (0% frequency), but 70% is nevertheless a valid description of your belief. You think it's roughly twice as likely that I woke before 8 than after 8. Because you haven't seen any data on my morning routine, the 70% is called a *prior probability* (as in 'belief prior to seeing any data'). Bayesian statistical procedures are then designed to update our belief optimally after seeing some data.

As you can see, frequentism vs. Bayesianism is a matter of philosophy. There has been quite some dispute over the right or wrong over the last decades and some hold strong opinions. Nowadays, the majority of statisticians take a rather neutral stance and use whatever is most convenient in a given situation.

## 10.2 Bayesian model setup

As usual, suppose we want to make inferences about an unknown parameter $\theta$ from repeated observations of a random variable $X$. Bayesian procedures all follow the same recipe.

1. Choose a **prior** probability density $\pi(\theta)$ that expresses our beliefs about the unknown parameter $\theta$.

2. Choose a **statistical model** $f(x \mid \theta) = f(x; \theta)$ that reflects our beliefs about the behavior of $X$ if the true parameter would be $\theta$.

3. Calculate the **posterior** density $f(\theta \mid X_1, \dots, X_n)$, which reflects our updated belief after having observed data $X_1, \dots, X_n$.

Note that step 1 only makes sense if we view the unknown parameter as a random variable $\Theta$. In that view, the true parameter $\theta$ is the realization of this random variable in our universe.[1] Accordingly we write the statistical model in step 2 as $f(x \mid \theta)$. It is our model for the data conditional on the event $\Theta = \theta$. In practice, this model is formulated just as in the frequentist paradigm.

**Example 10.1.** *Suppose we want to make inferences about the parameter $p \in (0, 1)$ of a Bernoulli distribution. In the Bayesian paradigm, we view this parameter as a random variable $P$. If we take an agnostic view on the distribution of $P$, we would specify a flat prior, where all values in $(0, 1)$ are equally likely. That is, our prior is the uniform distribution, $\pi(p) = 1$ for $p \in (0, 1)$. Conditional on $P = p$, we belief that our data $X_1, \dots, X_n$ are* iid *and have* Bernoulli($p$) *distribution. Using Bayesian conventions, the entire model is written succinctly as*

$$X_1, \dots, X_n \sim \text{Bernoulli}(p), \quad p \sim \text{Uniform}(0, 1).$$

## 10.3 Bayesian updating

The big question is how we come up with the posterior in step 3. And that's where the name *Bayesian* comes from. Suppose for the moment that both $X$ and

[1] Most Bayesians do not distinguish random variable $\Theta$ and realization $\theta$ in notation and just write $\theta$ for both.

and $\Theta$ are discrete. Then Bayes theorem (Theorem 2.25) and the law of total probability (Theorem 2.26) give

$$\begin{aligned} f(\theta \mid x) = \mathbb{P}(\Theta = \theta \mid X = x) &= \frac{\mathbb{P}(X = x \mid \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X = x \mid \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x \mid \Theta = \theta)\mathbb{P}(\Theta = \theta)} \\ &= \frac{\mathbb{P}(X = x \mid \Theta = \theta)\pi(\theta)}{\sum_\theta \mathbb{P}(X = x \mid \Theta = \theta)\pi(\theta)} \end{aligned}$$

If $X$ and $\Theta$ are continuous, the analogous statement is

$$f(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta}.$$

The above rule gives the optimal update of our belief $\pi(\theta)$ having seen a single observation $X = x$. If we see multiple *iid* observations $X_1, \dots, X_n$, we replace the likelihood of a single observation $f(x \mid \theta)$ by the joint likelihood

$$L_n(\theta) = f(X_1, \dots, X_n \mid \theta) = \prod_{i=1}^n f(X_i \mid \theta).$$

This gives,

$$f(\theta \mid X_1, \dots, X_n) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)d\theta}.$$

The denominator $\int L_n(\theta)\pi(\theta)d\theta$ is called the *marginal likelihood* of the data. It is a normalizing constant not depending on $\theta$ and usually irrelevant for inference. Thus the main take away is

$$f(\theta \mid X_1, \dots, X_n) \propto L_n(\theta)\pi(\theta)$$

(posterior is proportional to likelihood times prior). This posterior reflects our updated belief — coming from prior belief $\pi(\theta)$ and having seen data $X_1, \dots, X_n$.

In the vast majority of cases, the posterior density is complex and we use simulation for inference (more on that later). It is instructive to study at least one case where the posterior has a simple form.

**Example 10.2.** *We continue with the model from Example 10.1:*

$$X_1, \dots, X_n \sim \text{Bernoulli}(p), \quad p \sim \text{Uniform}(0, 1).$$

*Using our main finding, we get*

$$f(p \mid X_1, \dots, X_n) \propto L_n(p)\pi(p) = L_n(p) = p^{\sum_{i=1}^n X_i}(1-p)^{n - \sum_{i=1}^n X_i}.$$

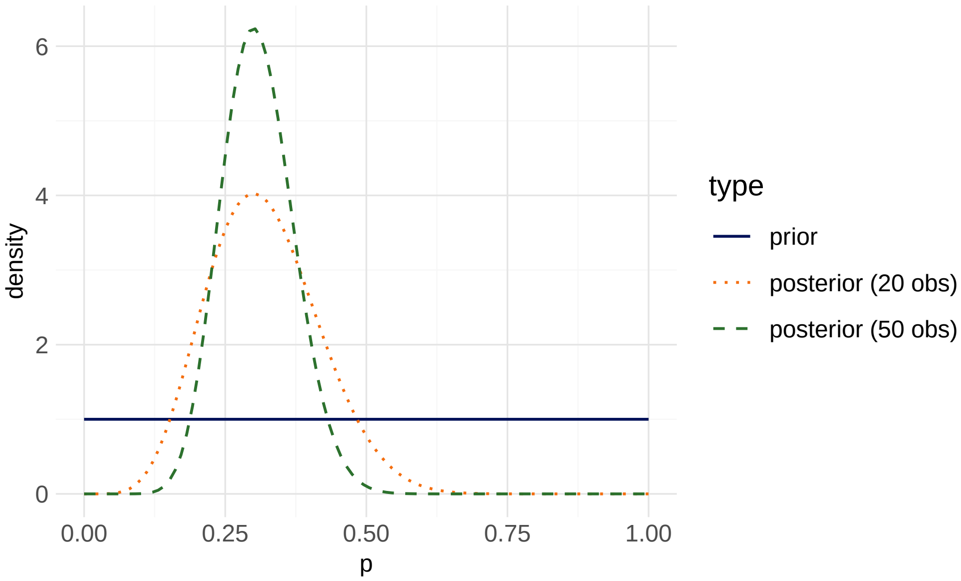


Figure 10.1: Prior and posterior densities in Example 10.2

*A seasoned statistician would immediately realize that this is proportional to the density of a* $\mathrm{Beta}(\sum_{i=1}^n X_i + 1, n - \sum_{i=1}^n X_i + 1)$ *random variable ('proportional' only because we threw away the normalizing constant). Hence, our posterior belief about the unknown parameter* $p$ *is expressed as*

$$p \mid X_1, \dots, X_n \sim \mathrm{Beta}\left(\sum_{i=1}^n X_i + 1, n - \sum_{i=1}^n X_i + 1\right).$$

*Now suppose that in reality* $p = 0.3$ *and we observe* $n = 20$ *observations such that* $\sum_{i=1}^n X_i = 6$. *Fig. 10.1 shows the prior,* $\pi(p) = 1$, *as a solid line and the posterior as dotted line. Intuitively, we saw* $6/20 = 0.3$ *observations with* $X_i = 1$ *so our updated beliefs should reflect that. Indeed, we see that the posterior concentrates on a region near* $p = 0.3$. *If we get 30 additional data points (so* $n = 50$*) with* $\sum_{i=1}^n X_i = 15$, *we can strengthen our posterior beliefs and this is reflected by a higher degree of concentration (dashed line).*

## 10.4 Bayesian inference

Bayesian methods for parameter estimation and uncertainty quantification are all based on the posterior density $f(\theta \mid X_1, \dots, X_n) \propto L_n(\theta)\pi(\theta)$. The standard estimator for the unknown $\theta$ is the *posterior mean*

$$\bar{\theta}_n = \mathbb{E}[\Theta \mid X_1, \dots, X_n] = \int \theta f(\theta \mid X_1, \dots, X_n) d\theta.$$

To quantify uncertainty, we can construct a Bayesian *credible interval* $C$ such that

$$\mathbb{P}(\theta \in C \mid X_1, \ldots, X_n) = \int_C f(\theta \mid X_1, \ldots, X_n) d\theta = 1 - \alpha.$$

The set $C$ is called credible interval instead of confidence interval to emphasize the difference in paradigm. The above *is* a probability statement about the unknown parameter $\theta$ (which would be meaningless under the frequentist paradigm): Given the data we have, our belief that $\theta \in C$ is described by the probability $1 - \alpha$.

If the posterior density has a simple form, posterior means and credible sets are easy to extract.

**Example 10.3.** *Let's continue where we left off in* (10.2)*:*

$$p \mid X_1, \ldots, X_n \sim \mathrm{Beta}\left(\sum_{i=1}^n X_i + 1, n - \sum_{i=1}^n X_i + 1\right).$$

*The mean of a* Beta$(\alpha, \beta)$ *is* $\alpha/(\alpha + \beta)$*, so the posterior mean is*

$$\bar{p} = \frac{\sum_{i=1}^n X_i + 1}{n + 2} = \frac{n}{n + 2}\bar{X}_n + \left(1 - \frac{n}{n + 2}\right)\frac{1}{2}.$$

*Hence, the posterior mean is a weighted average between the sample mean* $\bar{X}_n$ *(which is the frequentist MLE) and the prior mean,* $1/2$*. As* $n \to \infty$*, the weight for the sample mean,* $n/(n + 2)$*, tends to one. This is a general phenomenon: for large samples, Bayesian and frequentist estimates are very similar. On small samples, the prior contribution matters however.*

## 10.5 The choice of prior

The prior plays an important role in Bayesian methods. First and foremost, it makes them inherently subjective. Militant Bayesians see this is as the main advantage, militant frequentists as the main disadvantage. In any case, the choice of prior can be important, and choosing a bad prior can set us up for failure.

### 10.5.1 Bad priors

To illustrate how bad priors can go wrong, we'll go back to the origins of Bayes' theorem. Bayes' motivation for his framework was to model how rational people should make decisions under uncertainty. One such decision is to vote for our political leaders. Suppose there are only two options, candidates $A$ and $B$. People vote for the candidate that they believe to be most suitable for the role. Let's assume that all people are rational and update their beliefs according to Bayes' theorem. In reality, candidate $A$ is a huge idiot and everything he does reflects that. What happens if some people have the prior belief $\mathbb{P}(\text{A is best}) = 1$?

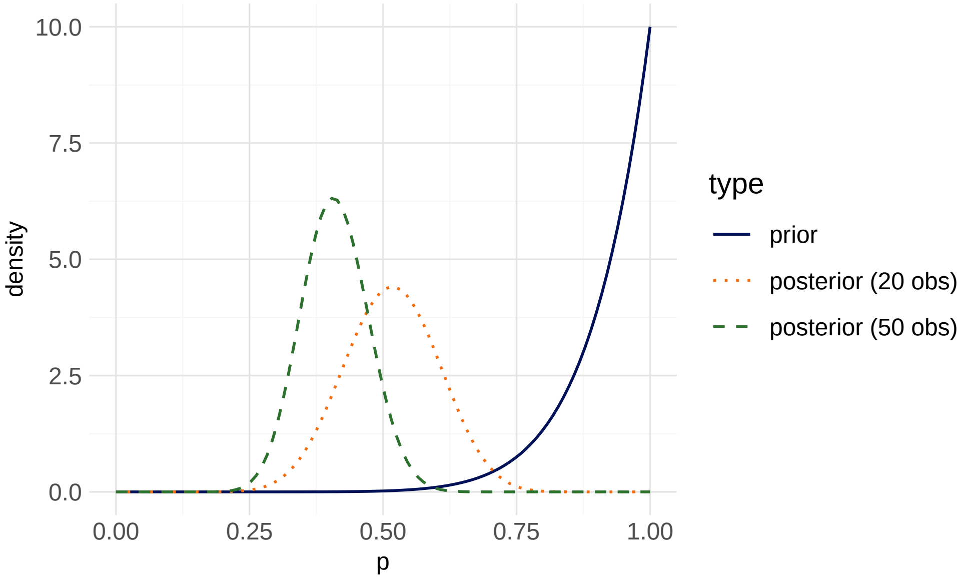


Figure 10.2: The effect of a biased prior on the posterior densities in Example 10.2

Well, no matter the evidence, if they apply Bayes theorem to update their beliefs, they will end up with $\mathbb{P}(\text{A is best} \mid \text{data}) = 1$. (You can do the calculation yourself.) They will vote for the idiot, no matter how outrages his actions. This is a very extreme example, but a weaker phenomenon occurs if a voter's prior is strongly biased towards one candidate. The stronger the bias, the more contradictory evidence the voter must see before changing their opinion.

So what does that mean for Bayesian procedures? First, if we assign prior probability 0 to some region $C$ of the parameter space, the posterior probability for this region will inevitably be 0. Thus, we should make sure that the prior density assigns positive mass to all possible outcomes. Second, if our prior heavily favors a certain region, this bias will only gradually fade out from the posterior. This is illustrated in Fig. 10.2, where we chose a prior that is heavily biased towards large values of $p$. Given the same data as in Fig. 10.1, the posteriors preserve the bias for large values of $p$ and only gradually move towards the true value $p = 0.3$.

### 10.5.2 Improper and non-informative priors

To avoid unreasonable bias in the posterior, we may be tempted to assign all values equal prior probability. This is what we've done in Example 10.1 by choosing a *flat prior*, $\pi(\theta) \propto \text{const}$. If the parameter range for $\theta$ is unbounded, flat priors cannot be proper probability densities (because $\int \pi(\theta) d\theta = \int \text{const}. d\theta = \infty \neq 1$). In that case, we say that the prior is *improper*. Improper priors are usually unproblematic, the posterior will be proper density nevertheless.

A more subtle question is whether flat priors are really as agnostic as they seem. Spoiler: they are not. The key insight is that flat priors are not transformation invariant. Suppose that $g$ is some nonlinear function. If we reparametrize our model by $g(\theta)$ instead of $\theta$, the flat prior for $\theta$ turns into a non-flat prior for $g(\theta)$ (this follows from the transformation-of-densities theorem, Theorem 2.58). A better choice for non-informative priors is *Jeffrey's prior*

$$\pi(\theta) = I(\theta)^{1/2},$$

where $I(\theta)$ is the Fisher information from Section 6.3. This prior can indeed be shown to be transformation invariant, but is often hard to compute.

### 10.5.3 Conjugate priors

Conjugate priors are a particularly convenient choice. A prior is called *conjugate* (relative to the statistical model $f(x \mid \theta)$), if both the prior $\pi(\theta)$ as well as the posterior $f(\theta \mid X_1, \dots, X_n)$ belong to the same parametric family. An example is the Beta distribution in Example 10.1. If we choose $p \sim \mathrm{Beta}(\alpha, \beta)$ as our prior, one can show that the posterior is

$$p \mid X_1, \dots, X_n \sim \mathrm{Beta}\biggl(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta\biggr).$$

(The flat prior $\pi(\theta) \equiv 1$ is a special case with $\alpha = \beta = 1$.) Conjugate priors are nice because they allow to do everything in closed form. However, there is only a handful of rather simple statistical models for which conjugate priors are known.

### 10.5.4 The good news: a frequentist perspective

We may ask whether Bayesian methods are reasonable in terms of their frequentists properties.[2] In summary, the answer is 'yes' as long the prior isn't plain stupid. The only condition we need is that the prior assigns positive mass to the true value $\theta$. Then posterior means are consistent and asymptotically normal and $(1 - \alpha)$-credible intervals cover the true parameter with probability $1 - \alpha$. And even if our prior is biased, the bias will disappear at the order $1/n$ and is therefore asymptotically negligible (compared to the usual $1/\sqrt{n}$ contribution of the variance).

## 10.6 Bayesian inference via simulation

For non-conjugate models, the default strategy for Bayesian inference relies on simulation from the posterior. As mentioned earlier, posterior densities are often complex and only known up to the normalizing constant. So how can we simulate from them? The solution is a simulation technique called *Markov Chain Monte Carlo* or MCMC for short. Let me say in advance that you rarely have to

[2] We need to switch to the frequentists paradigm to objectively assess the quality of estimators. The Bayesian view is always subjective through the prior.

implement this yourself, there are excellent libraries (like `emcee` and `PyStan` for Python). So we'll just quickly brush over the main idea.

The goal is to simulate from the posterior density $f(\theta \mid X_1, \dots, X_n) \propto L_n(\theta)\pi(\theta)$, optimally without knowing the normalizing constant. We do this by constructing a *stationary Markov chain* $\Theta_1, \dots, \Theta_T$. The sequence of random variables $\Theta_1, \dots, \Theta_T$ is called stationary if all $\Theta_t$ have the same distribution. In our case, we want this distribution to be the posterior. However, random variables in a Markov Chain are not independent. They only need to satisfy the Markov property

$$f(\Theta_t \mid \Theta_{t-1}, \dots, \Theta_1) = f(\Theta_t \mid \Theta_{t-1}).$$

Hence, the distribution of $\Theta_t$ depends on the past realizations $\Theta_1, \dots, \Theta_{t-1}$, but only through the most recent element $\Theta_{t-1}$.

The dependence in a stationary Markov chain is weak enough for the law of large numbers to hold. So if we are able to simulate such a stationary Markov chain $\Theta_1, \dots, \Theta_T$, the posterior mean can be approximated by

$$\widehat{\theta} = \frac{1}{T} \sum_{t=1}^{T} \Theta_t$$

and a $(1 - \alpha)$ credible interval can be computed from corresponding $\alpha/2$ and $(1-\alpha/2)$ sample quantiles. The quality of these approximations depends primarily on the length of the chain $(T)$ and the strength of dependence between consecutive elements. It is customary to throw away a good portion (maybe 5-10%) of the first elements of the chain (the so-called *burn-in period*), because most simulation algorithms take a while to stabilize.

So how do we simulate from the desired Markov chain? This has been and still is a very active field of research, so let's only have a quick look at the simplest method of all, the *Metropolis-Hastings algorithm.*

1. Pick some density $q(\cdot \mid \Theta_{t-1})$ that we can easily simulate from, called *proposal distribution.* For example we may take $q(\cdot \mid \Theta_{t-1}) \sim \mathcal{N}(\Theta_{t-1}, \sigma^2)$ if $\theta \in \mathbb{R}$ or a suitable Beta density if $\theta \in (0, 1)$.

2. Set $\Theta_0$ to an arbitrary value.

3. For $t = 1, \dots, T$:

   (i) Simulate a proposal value $\Theta_t \sim q(\cdot \mid \Theta_{t-1})$.

   (ii) Compute
   $$R = \frac{L_n(\Theta_t)\pi(\Theta_t)q(\Theta_{t-1} \mid \Theta_t)}{L_n(\Theta_{t-1})\pi(\Theta_{t-1})q(\Theta_t \mid \Theta_{t-1})}.$$

   (iii) Simulate $U \sim \text{Uniform}(0, 1)$. If $R > U$, set $\Theta_t = \Theta_{t-1}$.

Note that we do not require the normalizing constant $\int L_n(\theta)\pi(\theta)$ anywhere in the algorithm. One can show that the algorithm indeed produces the desired

Markov chain if $q$ assigns positive probability to all values in the parameter range of $\theta$. How much dependence there is between $\theta_t$ and $\theta_{t-1}$ depends on two factors: how much the proposal density $q(\cdot \mid \Theta_{t-1})$ concentrates around $\Theta_{t-1}$ and how often we set $\Theta_t = \Theta_{t-1}$ in step 3(iii). The dependence will be weaker, the closer $q(\cdot \mid \theta_{t-1})$ is to the posterior density $f(\theta \mid X_1, \dots, X_n)$.

## 10.7 Concluding remarks

You should now be familiar with the core ideas behind Bayesian inference. There is also a Bayesian variant of hypothesis testing. Here, so-called *Bayes factors* play a similar role to frequentist $p$-values. But Bayes factors actually do quantify the probability of a certain hypothesis given the data (and our prior belief). That was impossible in the frequentist paradigm.

The Bayesian paradigm has many advantages and equally many disadvantages over the frequentist one. For some people, the main advantage is that we can incorporate prior information in a principled manner. Sometimes this is important, but more often than not specifying the prior is treated as ancillary task. Some find the Bayesian philosophy more intuitive. Being able to make probability statements about unknown parameters and hypotheses is a nice consequence. There is another advantage that I think is the driving force behind its popularity: it is user friendly. No matter how complex a model is, we always do the same: write down the likelihood, specify priors for all parameters, and throw them into some Python function. The function returns samples from the posterior density and we get parameter estimates and uncertainty quantification for free (more or less).

All these points have a flipside. First and foremost, we absolutely must specify a prior. For complex models with many parameters, coming up with a justifiable choice is hard. More importantly, priors make our inferences inherently subjective. While we can make probability statements about unknown parameters and hypotheses, someone else holding different priors would end up with different answers. Which one is more valid scientifically? Hard to say. Furthermore, throwing every model into an MCMC algorithm is convenient but computationally demanding. Especially for very large samples, this can be annoying or even prohibitive.

In the end, both philosophies have their merits. Empirically, well-designed Bayesian and frequentists methods give almost the same answers (there's even mathematical theory to back that up). In my opinion, modern statisticians should have both in their toolbox and decide on a case-by-case basis what suits the problem better.

Lastly, a word of caution. There is a lot of nuance to the theory that was left out intentionally. Before you go all Bayesian with your data, make sure to consult other textbooks[3].

[3] I heard ‘Statistical rethinking’ by Richard McElreath is nice for applications.