Statistics Lab
Lecture Notes

Verena Wolf[1]

July 13, 2020

[1] verena.wolf@uni-saarland.de

Contents

Preface

The goal of this course is to guide students to a thorough understanding of probability and statistics, which play a very prominent role for most state-of-the-art machine learning tools.
Throughout the course, a large number of examples and exercises will help to get an intuitive understanding of the covered topics. Whenever appropriate, Python code will be included to explore important concepts using concrete data sets.
Before using these lecture notes, please carefully read the following hints:

- We use hyperlinks to Wikipedia pages to encourage students to explore further the context of this course. However, we would like to point out that Wikipedia entries are mostly written by laymen and should not be used as a primary information source.
- The Python code will be provided as Jupyter Notebooks and we use Python 3. Please make sure that you have all the necessary packages installed and up to date.

Chapter 1

Data Description

Descriptive statistics refers to a set of tools that is typically used in a first step of data analysis. By examining a number of statistical quantities and graphical displays, a preliminary understanding of the data is achieved.

Chapter learning objectives:

- distinguish different types of data (e.g. ordinal vs. nominal)
- describe data by means of different summary statistics for a data set (e.g. mean, variance, skewness)
- apply and understand standardization of data

1.1 Important Terms

We consider a data set and refer to it as a *sample* or *sample set* from a (typically larger) *population*, i.e., a population is a set of elements which is so large or even infinite that we can only analyze subsets of it. Hence, we draw random subsets from it which are called samples. Examples of population are all the citizens of a country, all cancer patients, all sentences of a certain language, etc. We typically hope that our sample set is representative for the population in the sense that statistical quantities of the sample set are similar to those of the population. Often we are interested in certain *attributes or factors* of the population elements such as the height of a person or the frequency of a word.

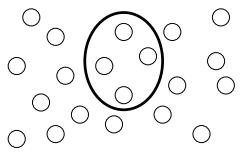


Figure 1.1: A sample from a population.

Quantitative data

Attributes may refer to some quantity of something and have numerical values such as the attribute height, for instance. Quantitative data may have a discrete range (e.g. subset of $\mathbb{N}$) or a continuous range (e.g. subset of $\mathbb{R}_{\geq 0}$). Attributes such as height, weight, length, and time are *continuous data* while counts of days, successes, etc are *discrete data*.

Qualtitative data

Qualitative data are any type of data that are not numerical. Examples are name, gender, country, social security number, etc. If a qualitative attribute has an order it is called *ordinal*. For instance, assume that we consider a person's educational experience (with values such as elementary school graduate, high school graduate, and college graduate). These also can be ordered and we can assign numbers accordingly such as elementary school (1), high school (2), and college (3). Even though we can order these from lowest to highest, the spacing between the values may not be the same (e.g. the gap between elementary school and high school may be larger than the gap between high school (2), and college). Another example for an ordinal attribute is shoe size where there is a clear order but size 42 is not twice as large as size 21.
If attribute values do not have an implied ordering, the attribute is called *nominal*. For instance, the attribute *color* is a nominal one because we want to distinguish blue, red, yellow, etc but do not order the colors in a certain way.

1.2 Data description

1.2.1 Frequency

Assume that we are interested in an attribute that has a finite number of possible values $a_1, a_2, \ldots, a_k$. Assume now that we have a sample set of size n and that $x_i \in \{a_1, a_2, \ldots, a_k\}$ is the attribute value of the i-th element in our sample set (e.g. x_1=blue, x_2=red, x_3=red, etc.). Then the number of x_i's in the sample set with value a_j is called the *absolute frequency* of a_j, written

$$h_j := h(a_j).$$

The *relative frequency* is then

$$f_j := h_j/n.$$

Hence

$$\sum_{j=1}^{k} h_j = n, \quad \sum_{j=1}^{k} f_j = 1.$$

Later, we will discuss frequency tables and how to generate them using Python.

Example 1: Absolute and relative frequency

Assume that we have a sample set of size $n = 3$ with the attribute 'color' taking values in $A = \{red, blue, green\}$ and the samples are $x_1 = blue$, $x_2 = red$, $x_3 = red$. Then, the absolute frequencies are

$$\begin{aligned} h_{blue} &= 1 \\ h_{red} &= 2 \\ h_{green} &= 0 \end{aligned}$$

and the relative frequencies are

$$\begin{aligned} h_{blue} &= 1/3 \\ h_{red} &= 2/3 \\ h_{green} &= 0 \end{aligned}$$

1.2.2 Measures of Location

To find out where the center of the data is, the following measures of location are common.
Assume that we consider a numerical attribute (values may be real or integer numbers) and that x_i is the value of the i-th element in our sample of size n. In other words, we have data $x_1, \ldots, x_n$. Then the arithmetic mean or sample mean is given by

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

Note that the mean is a very natural measure, but it is sensitive to extreme values. In the case of highly skewed data, the mean is not a good measure of location of the data.
The (sample) median is a value that splits the data into two parts of equal size. To compute the median, we sort the data $x_1, x_2, \ldots, x_n$ in increasing order. Given the sorted list $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and define

$$x_{med} := \begin{cases} x_{(\frac{n+1}{2})} & : \ n \text{ is uneven}, \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} & : \ n \text{ is even}, \end{cases}$$

Example 2: Median and Mean

For instance, if the sorted data is 1,2,2,3,4,4,5,5,5,5,6,7 then the median is $\frac{4+5}{2} = 4.5$ *while the arithmetic mean is* $\frac{49}{12} \approx 4.08$.

Compared to the mean, the median is harder to compute (sorting necessary!) but more resistant to extreme values.

Question:
Is it possible to modify the sequence of integers in the above example so that the median roughly gives a better measure of location compared to the mean?

Another measure of location which can also be used if an attribute is nominal, is the mode of the data given by the attribute value that appears most often in our data.

Example 3: Mode

The mode of the data set 1,2,2,3,4,4,5,5,5,5,6,7 is 5. We illustrate the mode together with mean and median in Figure 1.2.

In general, the mode and the median are less sensitive to outliers than the sample mean. If the last value in our example data set would be 70 (instead of 7), the mode and the median would be unchanged, while the sample mean increases to approximately 9.33.

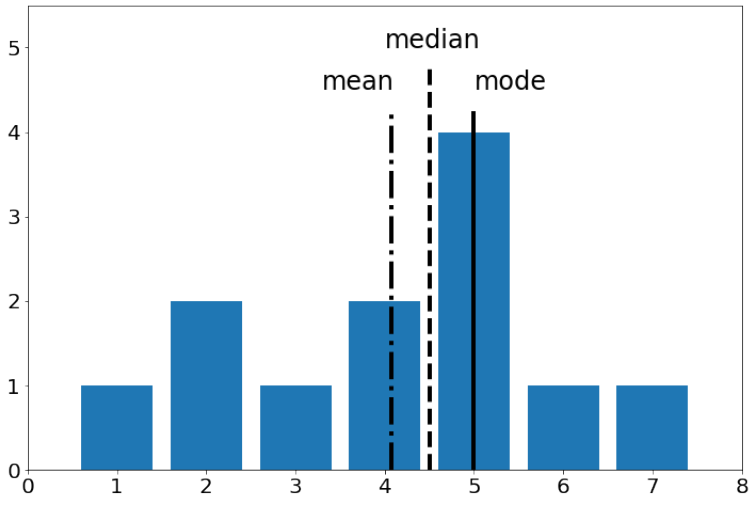


Figure 1.2: Plot of the data including mode, median, and mean.

1.2.3 Measures of Dispersion

Next we consider quantities that measure how much the data spreads (around its mean). The most common measure it the *sample variance*

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

and the *sample standard deviation*

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

To compute the sample variance, it is often helpful to exploit the identity

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

Example 4: Sample Variance

The sample variance of the data set 1,2,2,3,4,4,5,5,5,5,6,7 is

$$s^2 = \frac{1}{11}\left(1^2 + 2\cdot 2^2 + 3^2 + 2\cdot 4^2 + 4\cdot 5^2 + 6^2 + 7^2\right) - \frac{12}{11}\left(\frac{49}{12}\right)^2 \approx 3.17$$

Intuitively, the sample variance approximates the average squared distance of the observations from the sample mean. By taking its root, we scale back to the measurement units of the original data.
Also the *range* of the data is a measure of dispersion. It is defined as the difference between the largest and the smallest value, i.e., if $x_{max} = \max_i x_i$ and $x_{min} = \min_i x_i$ then the range is

$$x_{max} - x_{min}.$$

Example 5: Range

The range of the data set 1,2,2,3,4,4,5,5,5,5,6,7 is 7-1=6. We illustrate the range together with mean and range in Figure 1.3.

The Python code for the sample mean, sample variance and range for our example data set can be found below. We give a naive implementation for the mean and variance, as well as the built-in functions from numpy. We always `import numpy as np`. Note that the variance function from numpy has an optional parameter ddof, which we choose as 1 at this point. Later, we will discuss this issue in more detail.

```
# list with data
x=[1,2,2,3,4,4,5,5,5,5,6,7]
# mean
xbar = sum(x)/len(x)
# sample variance
s2 = sum([(i-xbar)**2 for i in x])/(len(x)-1)
#range
r = max(x)-min(x)

#built-in function from numpy (np)
#mean
xbar = np.mean(x)
#sample variance
s2 = np.var(x,ddof=1)
```

Later, we will also discuss bounds such as Chebychev's Rule that consider how much of the data will fall past a certain distance (e.g. k standard deviations) from the mean.

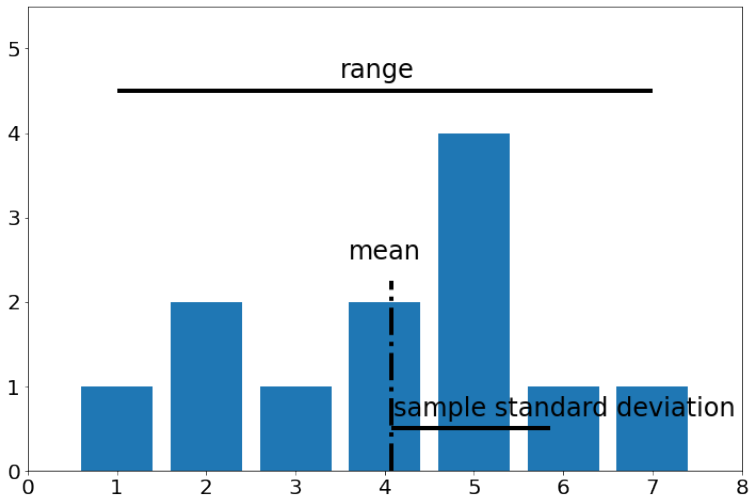


Figure 1.3: Plot of the data including mean, sample standard deviation, and range.

1.2.4 Measures of Shape

The distribution of the data may be symmetric, skewed to the right or left. In Figure1.4 we show the left- and right-skewed case (plots taken from [2]).

For the sample skewness, different definitions exist, for reasons that will become clear when we discuss important properties of estimators such as unbiasedness. For now, we only consider the most convenient formula to

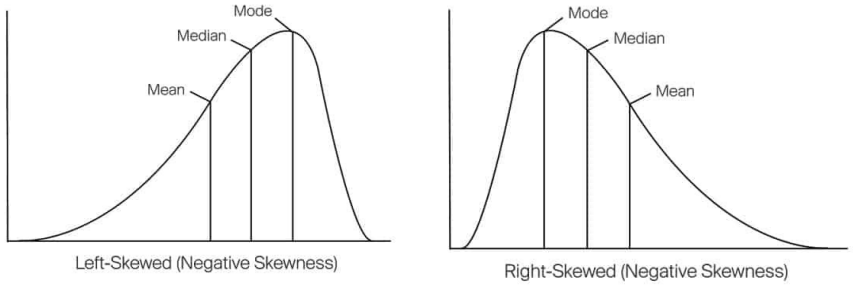


Figure 1.4: Left- and right-skewed data (here as a distribution plot).

estimate the skewness of the data:

$$g_1 = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

The skewness can be any value, $-\infty < g_1 < \infty$ and its sign indicates the direction of skewness (negative values correspond to skewness to the left and positive values to skewness to the right). If $g_1 \approx 0$, then the distribution is (nearly) symmetric.

1.2.5 Standardization of Data

Assume you have two or more sets of data from different sources and would like to compare the data. Then, it is helpful to transform (*standardize*) each sample set $x_1, \ldots, x_n$ as follows:

$$z_i := \frac{x_i - \bar{x}}{S},$$

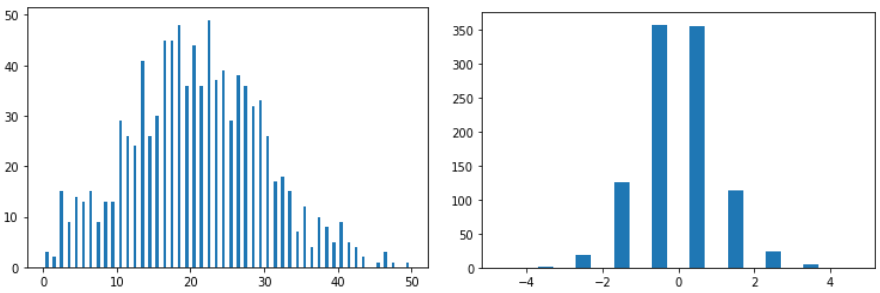


Figure 1.5: Histogram plot of data before (left) and after standardization (right).

i.e., consider the differences of a data point x_i to the mean $\bar{x}$ of the corresponding sample set *relative* to the standard deviation S.
Standardized data sets have mean zero and standard deviation 1 and can then be compared to other standardized data sets.
In Figure 1.5 we show histogram plots of some data before (left) and after standardization (right). Note that the data is shifted such that the mean becomes zero and scaled so that we have a standard deviation of one.

Chapter 2

Probabilities and Combinatorics

> "Was ein Punkt, ein rechter Winkel, ein Kreis ist, weiss ich schon vor der ersten Geometriestunde, ich kann es nur noch nicht präzisieren. Ebenso weiss ich schon, was Wahrscheinlichkeit ist, ehe ich es definiert habe." (Hans Freudenthal)

This section gives a short primer on probabilities combinatorics. The probability models that we discuss next are used to describe *chance experiments*. They allow to analyze systems of the real world that are subject to uncertainty - even if no data is available. The system may even be a hypothetical one or no measurements have been made. For instance, if we toss a fair coin three times, we can determine the probability of getting head only once. However, if we would like to know whether the coin is fair or not, a number of sample tosses can give us statistical evidence for or against fairness of the coin.

Chapter learning objectives:

- get familiar with basic concepts of probability: sample space, events, probability functions, etc.
- understand conditional probability, independence (including the corresponding laws and rules for computing (conditional) probabilities)
- determine numbers of possible outcomes by applying urn models

2.1 Probabilities

Let us consider chance experiments with a countable number of possible outcomes $\omega_1, \omega_2, \ldots$. The set $\Omega = \{\omega_1, \omega_2, \ldots\}$ of all outcomes is called the

sample space. Subsets of Ω are called *events* and by 2^Ω we denote the set of all events.

Example 6: Rolling a die and tossing a coin

If we roll a die, the set of possible outcomes is $\Omega = \{1, 2, \ldots, 6\}$. *The event "number is even" is given by* $E = \{2, 4, 6\}$.
If we toss a coin and count the number of trials until a head turns up for the first time, $\Omega = \{1, 2, \ldots\}$. *The event "number of trials greater than 10" is given by* $E = \{11, 12, \ldots\}$.

There are typically many different possibilities to define Ω. In the above example of a sequence of coin tosses, we already chose Ω such that it fits to the question that we have in mind (what is the probability of more than 10 trials). If, however, we were interested in the probability of having tails in the first three tosses, we would rather define Ω as

$$\{(\omega_1, \omega_2, \omega_3) | \omega_1, \omega_2, \omega_3 \in \{H, T\}\}$$

and ignore later tosses. Obviously, we could also encode H as 1 and T as 0. We define what a *probability* is using Kolmogorov's axioms.

Definition 1: Probability
Assume Ω *is a discrete (i.e. finite or countably infinite) and non-empty sample space. Let* P *be a function such that* $P : 2^\Omega \to [0, 1]$. *The value* $P(A)$ *is called the* probability *of event* A *if* P *is such that*

1. $P(\Omega) = 1$,
2. *for pairwise disjoint events* $A_1, A_2, \ldots$ *it holds that*

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots.$$

When we reason about the probability of certain events, we can use many arguments from set theory. For instance, if the events A, B, and C are as illustrated in Figure 2.1, it holds that

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ P(A \cup C) &= P(A) + P(C) - P(A \cap C) \\ P(\Omega \setminus B) &= P(\Omega) - P(B) \\ &= 1 - P(B) = P(\bar{B}) \\ P(A \setminus C) &= P(A) - P(A \cap C). \end{aligned}$$

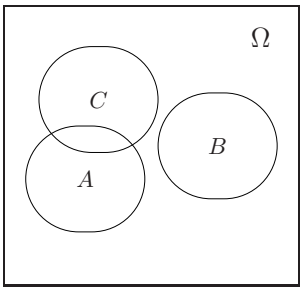


Figure 2.1: Using set arguments for the calculation of event probabilities.

Here, $\bar{B}$ denotes the *complement* $\Omega \setminus B$ of the set B. In addition, it is easy to show that $P(\emptyset) = 0$ and $P(A) \leq 1$ for all events A.

Example 7: Rolling a die

The probability of the event $\{1, 2, \ldots, 6\}$ *is one. Moreover, for each* $n \in \{1, 2, \ldots, 6\}$ *it holds that* $P(\{n\}) = 1/6$. *Thus,* $P(\{2, 4, 6\}) = 1/6 + 1/6 + 1/6 = 1/2$. *Similarly,* $P(\{1, 3, 5\}) = 1 - P(\{2, 4, 6\}) = 1/2$.

Example 8: Tossing a coin until heads for the first time

Consider a chance experiment, where a fair coin is tossed until a head comes up for the first time. We count the number of trials needed. $\Omega = \{1, 2, \ldots\}$. *The probability of the events "exactly n trials" and "more than three trails" are*

$$P(\{n\}) = (1/2)^n \text{ and}$$

$$P(\{4, 5, \ldots\}) = P(\Omega \setminus \{1, 2, 3\}) = 1 - (1/2 + 1/4 + 1/8) = 1/8.$$

Note that $P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\}) = 1/2 + 1/4 + 1/8 + \ldots = 1$.

The triple $(\Omega, 2^\Omega, P)$ is called a *discrete probability space*.

2.1.1 Conditional Probabilities

It is possible to restrict our reasoning to certain conditions, e.g. we consider rolling a die but restrict to those cases where the number is even. Under this condition, we could ask what the probability of getting number 2 is. Hence, we already *know* that the number is even and consider only even outcomes. Then, clearly with 1/3 we get two pips.

Definition 2: Conditional Probability

Let A, B *be events and* $P(B) > 0$. *Then*

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

is called probability of A under the condition B. *Clearly, this implies that* $P(A \cap B) = P(A|B) \cdot P(B)$.

Sometimes, we simply write $P_B(A)$ instead of $P(A|B)$. It is easy to show that P_B is a probability in the sense of Definition 1.

Example 9: Lung Cancer

We define the events
A: person gets lung cancer with $P(A) = \frac{72}{200000} = 0.00036$,
B: person is a smoker with $P(B) = 0.25$.
From the people that get lung cancer, 90% are smokers. The experiment consists in choosing a person at random. For the probability of getting lung cancer under the condition of being a smoker, we calculate

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{0.9 \cdot 0.00036}{0.25} = 0.001296.$$

For the probability of getting lung cancer under the condition of not *being a smoker, we get*

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(\bar{B}|A)P(A)}{P(\bar{B})} = \frac{(1-0.9) \cdot 0.00036}{0.75} = 0.000048.$$

Thus, the chance of getting lung cancer is around 30 times higher for smokers compared to non-smokers.

Muliplication rule

Let $A_1, A_2, \ldots, A_n$ be events. Then the following *multiplication rule* holds:

$$\begin{aligned} & P(A_1 \cap A_2 \cap \ldots \cap A_n) \\ = \; & P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \cdots \cdot P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \end{aligned}$$

Example 10: Drawing from an urn

Suppose you have an urn with 3 red, 2 blue and 1 green ball. You draw three times without placing the drawn balls back into the urn. The probability of drawing blue, green, red (in that order) is

$$\frac{2}{6} \cdot \frac{1}{5} \cdot \frac{3}{4}.$$

Note that we will discuss scenarios like this in more detail later.

Law of total probability

Assume that we have a finite or countably infinite number of events $A_1, A_2, \ldots$ that are pairwise disjoint. Assume further $\Omega = A_1 \cup A_2 \cup \ldots$ and $P(A_i) > 0$ for all i. Often, the probability of some event B is unknown, but the conditional probabilities $P(B|A_i), \ldots$ are known. In this case, $P(B)$

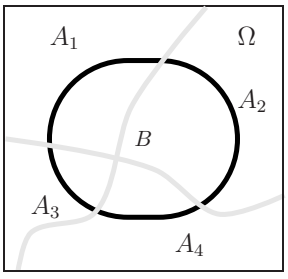


Figure 2.2: The law of total probability.

can be computed using the *law of total probability*, which states that

$$P(B) = \sum_i \underbrace{P(B|A_i) \cdot P(A_i)}_{=P(B \cap A_i)}.$$

The corresponding partitioning of Ω is illustrated in Fig. 2.2.

Example 11: Weather Forecast

In a city on 70 from 100 days the weather is good (G) and on 30 from 100 days the weather is bad ($\bar{G}$). The local meteorologist can predict good weather with 90% accuracy and bad weather with 60% accuracy. According to the law of total probability the forecast is correct with a probability of

$$P(M) = P(M|G) \cdot P(G) + P(M|\bar{G}) \cdot P(\bar{G}) = 0.9 \cdot 0.7 + 0.6 \cdot 0.3 = 0.81.$$

2.1.2 Bayes' rule

The law of total probability is helpful, if we are interested in $P(B)$. If, however, $P(A_i|B)$ is the probability of interest, one can use *Bayes' theorem* to compute it. With Bayes' theorem we can for any event A with $P(B) > 0$ express $P(A|B)$ by $P(B|A)$, that is,

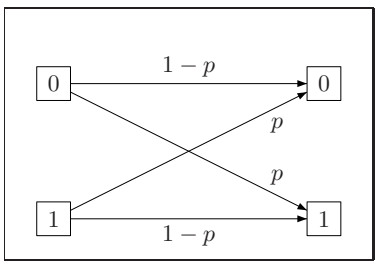
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

Thus, for the above problem of computing $P(A_i|B)$ for pairwise disjoint events A_i, this gives

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)}.$$

Example 12: Noisy Channel

Consider the noisy binary channel illustrated on the right.
If the channel is noise-free ($p = 0$), zero is transmitted from the upper left node to the upper right node. Similarly, one is transmitted from the lower left node to the lower right node.



If the channel is noisy, with probability $p > 0$ one is transmitted instead of zero and zero instead of one. Assume that the probability of sending

zero is π_0 and the probability of sending one is $\pi_1 = 1 - \pi_0$.
We define the events

- *A_0: send 0 with $P(A_0) = \pi_0$,*
- *A_1: send 1 with $P(A_1) = \pi_1 := 1 - \pi_0$,*
- *B_0: receive 0,*
- *B_1: receive 1,*

Then, $P(B_1|A_0) = P(B_0|A_1) = p$ and $P(B_1|A_1) = P(B_0|A_0) = 1 - p$.
We calculate

$$\begin{aligned} P(B_1) &= P(B_1|A_0) \cdot P(A_0) + P(B_1|A_1) \cdot P(A_1) = p \cdot \pi_0 + (1-p) \cdot \pi_1, \\ P(B_0) &= P(B_0|A_0) \cdot P(A_0) + P(B_0|A_1) \cdot P(A_1) = (1-p) \cdot \pi_0 + p \cdot \pi_1. \end{aligned}$$

and the receiver can determine the probability that the transmission was correct as

$$\begin{aligned} P(A_1|B_1) &= \frac{P(B_1|A_1) \cdot P(A_1)}{P(B_1)} = \frac{(1-p) \cdot \pi_1}{p \cdot \pi_0 + (1-p) \cdot \pi_1} \\ P(A_0|B_0) &= \frac{P(B_0|A_0) \cdot P(A_0)}{P(B_0)} = \frac{(1-p) \cdot \pi_0}{(1-p) \cdot \pi_0 + p \cdot \pi_1} \end{aligned}$$

Sometimes, it is advantageous to consider an equation for the ratio (the odds) $P(A|C)/P(B|C)$, which is straightforwardly derived from Bayes' theorem.

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

Example 13: Bayesian Odds

There are two urns, one containing 7 red and 3 blue balls, the other containing 3 red and 7 blue balls. We flip a fair coin to determine, from which urn we draw 12 balls with replacement. As a result, we get 8 times a red and 4 times a blue ball. What is the probability that it was the first urn with predominantly red balls?
Let U_1 (U_2) be the event of selecting the first (second) urn, respectively. Further, let A be the event of getting 8 times a red and 4 times a blue ball. Some concepts from combinatorics are necessary to determine the probabilities $P(A|U_1)$ and $P(A|U_2)$. They will be discussed in the next section. For now, just take it as a fact that

$$\begin{aligned} P(A|U_1) &\approx 0.23 \\ P(A|U_2) &\approx 0.01 \end{aligned}$$

Moreover, as we choose each urn with $1/2$, $P(U_1) = P(U_2) = 1/2$. *We are interested in the ratio between* $P(U_1|A)$ *and* $P(U_2|A)$, *because this tells us, whether* U_1 *or* U_2 *is more likely, given the result* A*:*

$$\frac{P(U_1|A)}{P(U_2|A)} = \frac{P(A|U_1)}{P(A|U_2)}\frac{P(U_1)}{P(U_2)} \approx 29.642.$$

Hence, given A *it is about 29 times more likely that we drew from urn 1. We get* $P(U_1|A)$ *by exploiting* $P(U_1|A)+P(U_2|A) = 1$ *and the above ratio:*

$$\begin{aligned} & P(U_1|A) = 1 - P(U_2|A) = 1 - \tfrac{P(U_1|A)}{29.642} \\ \iff \quad & P(U_1|A) \approx 0.967 \end{aligned}$$

Thus, with nearly 97%, *we drew from urn 1, which is intuitive since we got many more red balls than blue ones.*

2.1.3 Independent Events

Assume that we toss a coin twice. The sample space is $\{HH, HT, TH, TT\}$ and $P(B) = 2/4 = 1/2$ if B is the event that the first toss is H. If A is the event that the second toss is H, $P(A) = 1/2$ and $P(B \cap A) = 1/4$. If we now condition on B and compute probabilities that reason only about the second toss, we should not see any difference.

$$P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{1/4}{2/4} = 1/2 = P(A).$$

The information that the first toss is H has no bearing on the probability that the second toss is H. Morover, if we condition on $\bar{A}$, for the same reason, we get

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{1/4}{2/4} = 1/2 = P(A).$$

Definition 3: Independence
Let $0 < P(B) < 1$. *The event* A *is called* independent of B *if*

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

An equivalent condition for independence is that $P(A|B) = P(A)$.
Thus, the event A has the same probability no matter whether we condition on B or on $\bar{B}$. Our intuition tells us that in this case, event B "has nothing to do" with the event A. Another example than the coin tosses would be

that one rolls, say, two dice, one is red and one is blue. Event A is "red die gives six" and event B is "blue die gives even number". Then, clearly, A is independent of B and vice versa. However, the following example shows that independence can also occur between events that are highly related.

Example 14: Independent Events

Consider the case $A = \Omega$. The event Ω is independent of any event B with $0 < P(B) < 1$ since

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1 = \frac{P(\Omega \cap \bar{B})}{P(\bar{B})} = P(\Omega|\bar{B}).$$

Assume now that $\Omega = \{1, 2, \ldots, 6\}$, $A = \{5, 6\}$, and $B = \{2, 4, 6\}$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{6\})}{P(\{2,4,6\})} = 2 \cdot \tfrac{1}{6} = 1/3$$

and

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(\{5\})}{P(\{1,3,5\})} = 2 \cdot \tfrac{1}{6} = 1/3,$$

which means that $A = \{5, 6\}$ is independent of $B = \{2, 4, 6\}$.

Note that one can use the alternative (and equivalent) condition

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

to define independence. In particular, this condition can also be checked if $P(B) = 0$.
In summary, if $P(B) > 0$ we have

$$\begin{aligned} (1) \quad P(A|B) &= P(A|\bar{B}) \\ (2) \quad P(A \cap B) &= P(A) \cdot P(B) \\ (3) \quad P(A|B) &= P(A) \end{aligned}$$

From (2) we see that independence is a symmetric property. So, if $P(A) > 0$, further equivalent conditions can be derived where the roles of A and B are reversed. In other words: If the events A and B are independent then

- A and $\bar{B}$ are independent,
- $\bar{A}$ and B are independent,
- $\bar{A}$ and $\bar{B}$ are independent.

Remark:

In information theory, the negative log probability $I(A) := -\log P(A)$ of an event A is interpreted as the information content or level of surprise (the smaller $P(A)$, the larger $-\log P(A)$). For two independent events, the information content of the combined event simply adds up since

$$\begin{aligned} I(A \cap B) &= -\log P(A \cap B) \\ &= -\log(P(A) \cdot P(B)) \\ &= -\log P(A) - \log P(B) \\ &= I(A) + I(B) \end{aligned}$$

Later, we will take a closer look at information contents, entropy and other concepts from information theory.

Pairwise and Mutual Independence

In the case of more then two events, we distinguish pairwise independence and mutual independence. For events A, B, C, we say that A, B, and C are *pairwise independent* if event A is independent of event B, event A independent of event C, and B independent of event C.

Question:
How can the above definition be extended to n events?

Example 15: Pairwise independent events

We consider a roulette wheel with 36 numbers colored in red (R) or black (B) according to the following pattern:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
R	*R*	*R*	*R*	*R*	*B*	*B*	*B*	*B*	*B*	*R*	*R*	*R*	*R*	*B*	*B*	*B*	*B*
36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19

Now, let A and B denote the event that a spin of the wheel yields a red number and an even number, respectively. In addition, let C be the event that the number is smaller than 19. We check whether A, B, and C are pairwise independent. We have $P(A) = P(B) = 1/2$ since half of the numbers are red and also, half of the numbers are even. Moreover, $P(C) = 1/2$ since the numbers from 1 to 18 are the first half of all the numbers. From the above table, we see that

$$\begin{aligned} A \cap B &= \{2, 4, 10, 12, 24, 26, 32, 34, 36\}, \\ A \cap C &= \{1, 2, 3, 4, 5, 10, 11, 12, 13\}, \\ B \cap C &= \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, \end{aligned}$$

and thus

$$P(A \cap B) = 9/36 = 1/4 = 1/2 \cdot 1/2 = P(A) \cdot P(B),$$
$$P(A \cap C) = 9/36 = 1/4 = 1/2 \cdot 1/2 = P(A) \cdot P(C),$$
$$P(B \cap C) = 9/36 = 1/4 = 1/2 \cdot 1/2 = P(B) \cdot P(C),$$

which shows pairwise independence.

In the above example, it is the case that

$$P(A \cap B \cap C) = 4/36 = 1/9 \neq 1/8 = P(A) \cdot P(B) \cdot P(C).$$

Hence, a stronger notion of independence is necessary, to describe the relation between events where the probability of the intersection can be determined by multiplying the probabilities of all individual events.

Definition 4: Mutual Independence

Let K be a finite set of indices. We say that the events A_k, $k \in K$, are mutually independent if and only if for all $L \subset K$

$$P(\cap_{\ell \in L} A_\ell) = \prod_{\ell \in L} P(A_\ell).$$

Example 16: Mutual independent events

In the example above with three events, we considered $P(A \cap B), P(A \cap C), P(B \cap C)$ to check for pairwise independence. To determine whether A, B, and C are mutually independent, we additionally need check if

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

which is not the case as already stated above.

Question:
Does pairwise independence follow from mutual independence?

2.2 Combinatorics

To determine the probability of certain events, combinatorial considerations are often useful. We first focus on the elementary events, that is, all singleton sets $\{\omega\}$ where ω is an outcome. The union of these events forms Ω and in many chance experiments, all elementary events have the same probability. This is the case, for instance, in dice games, lottery, and other

games of chance. Assume now that we are interested in a certain event $A = \{\omega_1, \omega_2 \ldots, \omega_k\}$ which consists of k outcomes. Then, if all outcomes in Ω have equal probability, the probability $P(A)$ is given as the ratio of the number n_A of outcomes favorable to A to the number of all possible outcomes N,

$$P(A) = \frac{n_A}{N} = \frac{|\text{outcomes favorable to } A|}{|\text{all possible outcomes}|}.$$

This probability is also called *Laplace probability*. If an experiment has two parts, and there are n_1 ways for the first part of the experiment to happen, and n_2 ways for the second part to happen, then there are $n_1 \cdot n_2$ ways for the whole experiment to happen (*multiplication principle or rule of product*). This generalizes in the obvious way to more than two parts. The examples presented in the sequel are inspired by those mentioned here [3].

Example 17: Laplace Probability

We flip a coin and roll a six-sided die. The coin can land in two ways (heads or tails). The die can land in six ways (showing $1, 2, 3, 4, 5,$ *or* 6*). The total number of possible outcomes in the sample space is therefore* $2 \times 6 = 12$*. If we assume that the outcomes are equally likely, the probability for any single outcome, such as the coin landing heads and the die showing* 4*, is therefore* $1/12$*.*
In a different experiment, we flip a coin six times and roll a six-sided die. The total number of possible outcomes in the sample space is therefore $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 6 = 2^6 \times 6 = 384$*. Let* A *be the event that exactly one flip lands heads and the die shows* 1*. The number of outcomes in* A *is* 6*, since there are six possibilities for which flip is a head and only one way for the die to show* 1*. Hence, if we assume that the outcomes are equally likely,* $P(A) = 6/384$*.*

Many chance experiments can be mapped to one of the following urn problems, where we have an urn containing n distinguishable balls. We draw a ball from the urn k times. We can do this in two ways: we might replace the ball drawn each time before drawing the next ball, or we might not replace the ball (in which case k cannot be bigger than n). We may also consider the order of balls drawn to matter, or we may consider the draws to be unordered.

Drawing with Replacement, Ordered Result

Since we replace the balls drawn, each draw can pick any of the n balls. Since we draw k times, the multiplication principle says that the number of possible outcomes is n multiplied by itself k times, which is n^k.

Example 18: With Replacement, Ordered Result

We draw with replacement two times from an urn containing three balls - red, green, and blue. There are $3^2 = 9$ *possible outcomes: RR, RG, RB, GR, GG, GB, BR, BG, BB.*

Question:
Suppose we have a neural network with k layers, where each layer has n nodes. How many ways are there to traverse the network if all layers are fully connected?

Question:
In tic tac toe there are 9 fields, which can either be empty or filled with an X or an O. How many possible configurations are there?

Drawing without replacement, ordered result

When we do not replace the balls, the number of possible ball choices goes down by one after each draw. The multiplication principle then says that the total number of possible outcomes with k draws from an urn with n balls is

$$n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}.$$

Example 19: With Replacement, Ordered Result

We draw without replacement two times from an urn containing three balls - red, green, and blue. There are $3 \times 2 = 6$ *possible ordered outcomes: RG, RB, GR, GB, BR, BG.*

If $k = n$*, we are drawing some permutation of the balls. From the formula above, the number of possible permutations is* $n!$ *(remembering that* $0! = 1$*).*

Drawing without replacement, ordered result

When we don't replace the balls, the number of possible ball choices goes down by one after each draw. The multiplication principle then says that the total number of possible outcomes with k draws from an urn with n balls is

$$n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}.$$

Example 20: Drawing without replacement, ordered result

We draw without replacement two times from an urn containing three balls - red, green, and blue. There are $3 \times 2 = 6$ *possible ordered outcomes: RG, RB, GR, GB, BR, BG.*

If $k = n$, we are drawing some permutation of the balls. From the formula above, the number of possible permutations is $n!$ (remembering that $0! = 1$).

Drawing without replacement, unordered result

If we draw k balls without replacement, and don't look at the order of the balls drawn, the number of possible results of the experiment decreases, compared to the result above when we do look at the order. There are $k!$ ways of ordering k balls (the number of permutations of k items), so the result with ordering over-counts by this factor. So, dividing the number of outcomes with ordering by this factor, we get that the number of ways of drawing k balls without replacement ignoring order from an urn with n balls is

$$\frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

This is called "n choose k". Note that

$$\binom{n}{k} = \binom{n}{n-k} \quad \text{and} \quad \binom{n}{0} = \binom{n}{n} = 1.$$

Example 21: Drawing without replacement, unordered result

We draw without replacement two times from an urn containing three balls - red, green, and blue. There are $(3 \times 2)/(1 \times 2) = 3$ *possible unordered outcomes:* $\{R,G\}$, $\{R,B\}$, $\{G,B\}$.

Example 22: Drawing without replacement, unordered result

A popular german lottery game is "6 aus 49", where six balls are drawn from an urn that contains 49 enumerated balls. In total there are

$$\binom{49}{6} = 13\,983\,816$$

possible outcomes.

Drawing with replacement, unordered result

We can use the result above to find how many possible outcomes there are when drawing k balls with replacement, when we don't look at the order the balls are drawn in. Since we don't look at the order, all that matters is how many times each of the n balls in the urn are drawn. We can represent a count as a sequence of Os, with the number of Os being equal to the count. We can represent the counts of how many times each of the n balls were drawn by putting together the sequences of Os representing the counts for each ball, separating them with Xs. (We choose some order for the n balls; it doesn't matter which order.)
For example, if $n = 3$, with the balls labelled red, green, and blue, and $k = 6$, one possible outcome is 2 red, 1 green, and 3 blue. Ordering the balls as red, green, blue, these counts can be represented by the sequence OOXOXOOO.
Every set of counts will correspond to a sequence of $k + (n - 1)$ Xs and Os, in which the number of Xs is exactly $n - 1$, and every such sequence will correspond to a set of counts. The correspondence is one-to-one, so we can count the number of outcomes of the experiment by counting how many sequences there are of length $k + n - 1$ with $n - 1$ of the positions being occupied by Xs.
The number of ways of putting $n - 1$ Xs down in a sequence of length $k + n - 1$ is the same as the number of ways of choosing $n - 1$ balls without replacement from an urn with $k + n - 1$, ignoring order. We figured that out above that this is $k + n - 1$ choose $n - 1$, i.e.,

$$\binom{k+n-1}{n-1}.$$

This is the same as the number of ways of choosing places for the k Os out of the $k + n - 1$ positions, which is $k + n - 1$ choose k.

Example 23: Drawing with replacement, unordered result

We draw with replacement two times from an urn containing three balls - red, green, and blue. This gives us two Os and two Xs. For instance, OOXX means that both balls are red, XOXO represents one green and one blue ball, etc. There are

$$\binom{2+3-1}{2} = 6$$

possible outcomes if we ignore order: $\{R,R\}$, $\{R,G\}$, $\{R,B\}$, $\{G,G\}$, $\{G,B\}$, $\{B,B\}$.

Example 24: Processor jobs

A computer has 6 processors. It is regularly used to run jobs of 4 kinds. It always runs 6 jobs at a time, so that all the processors will be used, but there won't be any processor contention between jobs. The performance of the computer may depend on what kinds of jobs it is running (e.g., it may go slowly if two jobs that both access the disk a lot are running). We're therefore interested in how many possible job mixes there are, since we may need to evaluate performance for each job mix.
We can treat this as a problem where we draw $k = 6$ *balls (jobs) with replacement from an urn with* $n = 4$ *balls (kinds of jobs), and we care only about the numbers of jobs of each kind (there's no order to jobs). The answer is therefore*

$$\binom{6+4-1}{4-1} = \binom{9}{3} = 84$$

possible job mixes.

The above models can be used to calculate Laplace probabilities. Recall that if we assume equally-likely outcomes, for an event A,

$$P(A) = \frac{n_A}{N} = \frac{|\text{outcomes favorable to } A|}{|\text{all possible outcomes}|}.$$

Both, the number of outcomes favorable to A and the number of all possible outcomes may be computed using one of the formulas above.

Example 25: Birthday problem

A famous probability problem is to find how likely it is that, at a party with n *people, at least two people have the same birthday. Let* A *be the event that two or more of the* n *people have the same birthday.* A^c *is the event that all birthdays are distinct. We'll find* $P(A^c)$*, and then get* $P(A)$ *as* $1 - P(A^c)$.

We assume equal probability for each day of the year. Hence, $P(A^c) = \#(A^c)/\#(S)$. The number of outcomes in the sample space is $\#(S) = 365^n$, the same as the number of ways of drawing n balls with replacement from an urn with 365 balls, paying attention to the order. Using the multiplication principle, the number of outcomes with no birthdays on the same day is $\#(A^c) = 365 \cdot 364 \ldots (365 - n + 1)$, which is the same as the number of ways of drawing n balls from an urn with 365 balls without replacement, paying attention to the order. We use these numbers to compute $P(A) = 1 - \#(A^c)/\#(S)$. For example, if $n = 5$, then $P(A) = 2.7\%$.
The following Python code computes the probabilties of the birthday problem (while preventing overflow problems due to large integers).

```
plt.plot([1 - np.prod([(356 - i) / 356
                       for i in range(n)])
          for n in range(100)])
```

Try it out for different n!

Indistinguishable Balls

So far, we assumed that the urn contains n distinguishable balls. But what if the urn contains r red balls and $n - r$ blue balls and balls of the same color cannot be distinguished?

Question:
Consider the case of drawing multiple balls (possibly more than one ball of the same color) with replacement, where balls with the same color can not be distinguished. Assume that we do not care for the order in which the balls are drawn. Does the number of possible outcomes differ from the case with only one ball of each color? Why (not)?
Is your claim also true for the probabilities of the outcomes?

Example 26: With Replacement

Consider an urn with $n = 9$ balls, where 5 are red and 4 are blue. For k draws the number N of possible outcomes is $2 \cdot 2 \cdot \ldots \cdot 2 = 2^k$. However, the probability of, say, two times red and once blue for $k = 3$ is $(5/9)^2 \cdot (4/9)$ in the ordered case and $3 \cdot (5/9)^2 \cdot (4/9)$ in the unordered case as there are three possible combinations.

Without replacement things are a little bit more complicated. However, for small examples it is possible to manually count the possible outcomes:

Example 27: Without Replacement

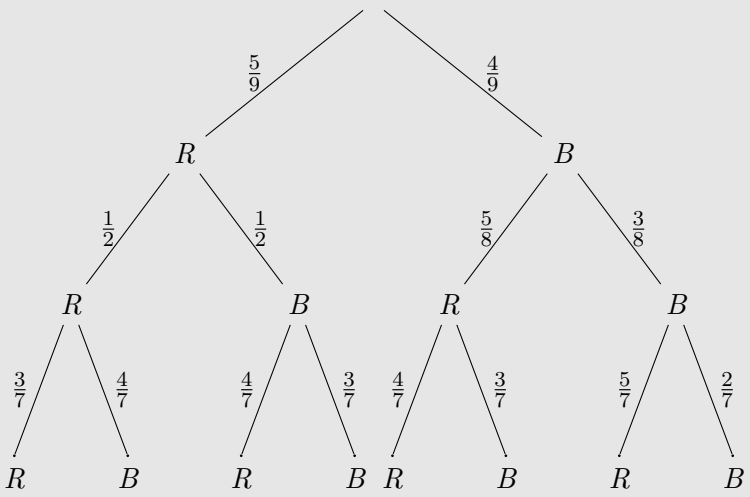
Consider an urn with $n = 9$ balls, where 5 are red and 4 are blue. For k draws the number N of possible outcomes, ignoring the order, is

k	1	2	3	4	5	6	7	8	9
N	2	3	4	5	5	4	3	2	1

If we now consider more than two different colors or care for the order, things get even more complicated. Luckily, if we are only interested in the probability of the outcomes (and not in the total number of possible outcomes), we can use probability trees to easily calculate the probabilities.

Example 28: Without Replacement

Consider again an urn with $n = 9$ balls, where 5 are red and 4 are blue. Assume we draw three times from the urn without replacement ($k = 3$). The probability of a certain event in the case with considering the order but without replacement can be obtained from the following probability tree that lists all possible cases of the three draws.



To get the probability of a certain event, we traverse the tree along the path that corresponds to the event (from top to bottom) and multiply the respective probabilities. For example, for the event RBR we get $\frac{5}{9} \cdot \frac{1}{2} \cdot \frac{4}{7}$. If we do not care for the order, we go through all suitable paths and add the resulting probabilities. For $\{R, R, B\}$, i.e., two red and one blue ball in arbitrary order, we have the paths RRB, RBR and BRR which yield the total probability $\frac{5}{9} \cdot \frac{1}{2} \cdot \frac{4}{7} + \frac{5}{9} \cdot \frac{1}{2} \cdot \frac{4}{7} + \frac{4}{9} \cdot \frac{5}{8} \cdot \frac{4}{7}$.
Note that if we consider the case with replacement, every edge that leads to R has the probability $\frac{5}{9}$ and every edge that leads to B the probability $\frac{4}{9}$.

Chapter 3

Discrete Random Variables

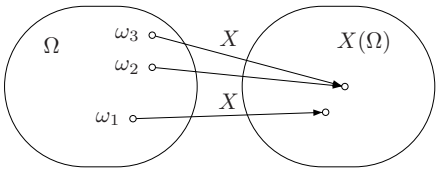
In this chapter, we discuss one of the most important concepts of the course: random variables. To keep things simple, we first stick to discrete random variables, which take values in a finite or countably infinite support set.
The relationship to the previously explained idea of having a theoretical population and a random sample is the following: We will assume that our population corresponds to a certain probability distribution p and a sample can then be described as a random variable that follows this distribution. The distribution p is usually unknown and our goal is to approximate it (e.g. with some theoretical distribution) and we describe samples drawn from the population as random variables with distribution p. For example, let $x_1, \ldots, x_{20}$ be the age of 20 first-year computer science students. This is the information that we have at hand and the theoretical population is the set of all first-year computer science students. We approximate it with, say, a binomial distribution[1]. Hence, if X is the random variable that follows the corresponding binomial distribution, we assume that $x_1, \ldots, x_{20}$ are realizations of X.

Chapter learning objectives:

- understand the mathematical construction underlying discrete random variables in all its depths
- define appropriate discrete random variables for a variety of problems
- computation of basic properties of discrete random variables: expectation, variance, etc.
- combination and transformation of discrete random variables
- kowledge of the most important discrete probability distributions

[1] Later in the course, you will learn how to infer the corresponding parameters of the binomial distribution.

3.1 Discrete Random Variables and Probability Distributions



A random variable is used to represent an outcome of an experiment. Technically, the fact that this variable is "random" (that is, it takes values with a certain probability), is realized by using a mapping (as illustrated in the figure on the left). But why are random variables useful if we have sample spaces, events and probabilities?
Often, it is difficult to explicitly define a probability space for a specific problem. For example, consider the experiment where we throw a die 100 times. Then, we are probably not only interested in the frequencies of the pips, but also in events such as the sum of the pips being equal to 85. Such transformations become simple when we work with random variables.

Definition 5: Discrete Random Variable
Let $(\Omega, 2^\Omega, P)$ *be a discrete probability space. A function*

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

with

$$\omega \mapsto X(\omega) \in \mathbb{R}$$

is called a discrete (real-valued) random variable *on* $(\Omega, 2^\Omega, P)$.

The value $x := X(\omega)$ is called realization of the random variable.
Now, that we mapped the outcome of an experiment to $\mathbb{R}$, we need to assign probabilities to the subsets of $X(\Omega) = \{a \in \mathbb{R} \mid \exists \omega \in \Omega : X(\omega) = a\}$. Since we already have probabilities for events $A \subseteq \Omega$, we define a function $P_X : 2^{X(\Omega)} \to [0, 1]$ such that, for $A \in 2^{X(\Omega)}$,

$$P_X(A) := P(X^{-1}(A)) = P(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Then $(X(\Omega), 2^{X(\Omega)}, P_X)$ is a discrete probability space.

Example 29: Rolling two dice

We want to define a random variable for the sum of the two numbers on the two dice. Let $\Omega = \{(\omega_1, \omega_2) \mid \omega_1, \omega_2, \in \{1, \ldots, 6\}\}$ *and* $X : \Omega \to \mathbb{R}$ *is such that* $X(\omega_1, \omega_2) = \omega_1 + \omega_2$.
Then, the probability of having the number 10 *is*

$$\begin{aligned} P_X(\{10\}) &= P(X^{-1}(\{10\})) = P(\{(\omega_1, \omega_2) \in \Omega \mid X(\omega_1, \omega_2) = 10\}) \\ &= P(\{(4,6)\}) + P(\{(6,4)\}) + P(\{(5,5)\}) = 1/12. \end{aligned}$$

In the sequel, we use the following "shortcuts" to refer to subset of Ω:

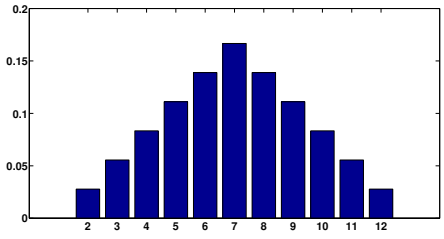


Figure 3.1: Discrete probability distribution (rolling two dice).

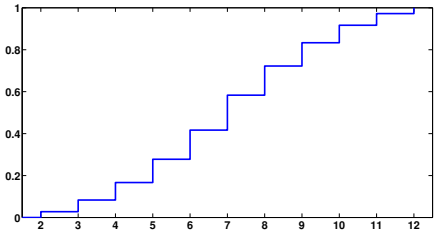


Figure 3.2: Cumulative probability (rolling two dice).

- "$X = a$" stands for the set $\{\omega \in \Omega \mid X(\omega) = a\}$
- "$X \leq a$" stands for the set $\{\omega \in \Omega \mid X(\omega) \leq a\}$
- "$X < a$" ...

Thus, for instance,

$$P(X \leq a) = P(\{\omega \in \Omega \mid X(\omega) \leq a\}) = \sum_{c \in X(\Omega), c \leq a} P(X = c).$$

The function $f : X(\Omega) \to [0, 1]$ with $f(a) = P(X = a)$ is called the discrete probability distribution of X (or probability mass function). It tells us the probability of each possible event of the form "$X = a$". Of similar importance is the following function related to a random variable X.

Definition 6: Cumulative Probability Distribution
Let X be a discrete (real-valued) random variable. *The function $F : \mathbb{R} \to [0, 1]$ with*

$$F(x) := P(X \leq x) = \sum_{a \in X(\Omega), a \leq x} P(X = a).$$

is called the cumulative probability distribution of X.

Example 30: Rolling two dice

Assume that X is defined as in Example 29. The discrete probability distribution and the cumulative probability distribution of X are shown in Figures 3.1 - 3.2, respectively.

The cumulative probability distribution is monotonically increasing and, as $x \to \infty$, $F(x)$ approaches 1. The most common discrete probability distributions will be discussed later.

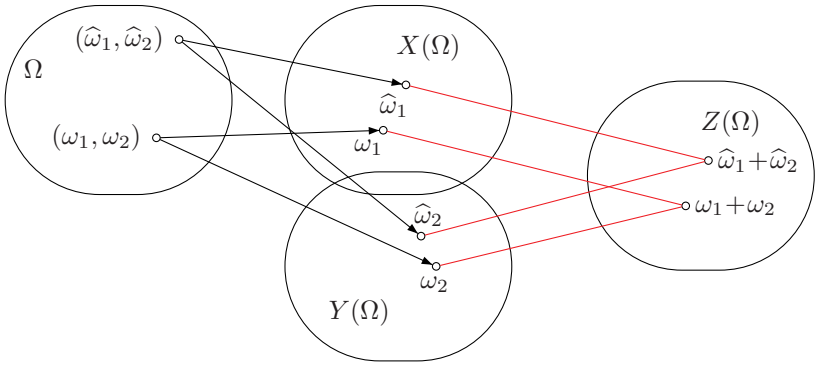


Figure 3.3: Sum of two random variables (rolling two dice).

3.2 Combining and Transforming Random Variables

It is possible to combine random variables on the same sample space Ω using operations such as $+$, $-$, $\cdot$, etc. For instance, we define $Z = X + Y$ as a random variable on Ω by

$$Z(\omega) = X(\omega) + Y(\omega) \text{ for all } \omega \in \Omega.$$

The probability $P(Z = z) = P(X + Y = z)$ is then well-defined since

$$P(Z = z) = P(\omega \in \Omega \mid Z(\omega) = z) = \sum_{\omega \in \Omega, X(\omega)+Y(\omega)=z} P(\{\omega\}).$$

Moreover,

$$\sum_{z \in Z(\Omega)} P(Z = z) = \sum_{z \in Z(\Omega)} \sum_{\omega \in \Omega, X(\omega)+Y(\omega)=z} P(\{\omega\}) = P(\Omega) = 1.$$

Example 31: Rolling two dice

Assume that $\Omega = \{1, 2, \ldots, 6\}^2$ *and* $X, Y : \Omega \to \mathbb{R}$ *are such that, for* $(\omega_1, \omega_2) \in \Omega$,

$$X(\omega_1, \omega_2) = \omega_1, \; Y(\omega_1, \omega_2) = \omega_2.$$

Then for $Z = X + Y$ *we have*

$$Z(\omega_1, \omega_2) = X(\omega_1, \omega_2) + Y(\omega_1, \omega_2) = \omega_1 + \omega_2$$

and (see illustration in Fig. 3.3)

$$P(Z=z) = \sum_{\substack{(\omega_1,\omega_2)\in\Omega,\\ X(\omega_1,\omega_2)+Y(\omega_1,\omega_2)=z}} P(\{(\omega_1,\omega_2)\}) = \sum_{\substack{(\omega_1,\omega_2)\in\Omega,\\ \omega_1+\omega_2=z}} P(\{(\omega_1,\omega_2)\}).$$

Remark:

In the sequel, we often work with combinations of only two random variables. However, all these properties and definitions carry over in a straightforward way to more than two, i.e. a finite number of random variables. Exceptions or additional concepts for the case of more than two random variables are explicitly mentioned (e.g. for independence we have pairwise and mutual independence).

General Transformations of X

In general, we can consider a function $g : \mathbb{R} \to \mathbb{R}$ and tranform a discrete random variable $X : \Omega \to \mathbb{R}$ with probability mass function f_X by defining $Y : \Omega \to \mathbb{R}$ with $Y(\omega) := g(X(\omega))$ for all $\omega \in \Omega$. Then, if the inverse of g exists, the probability mass function f_Y of Y is obtained as follows:

$$\begin{aligned} f_Y(y) &= P(Y=y) = P(g(X)=y) \\ &= P(\{\omega\in\Omega \mid g(X(\omega))=y\}) \\ &= P(\{\omega\in\Omega \mid X(\omega)=g^{-1}(y)\}) \\ &= P(X=g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \end{aligned}$$

Examples for transformations:

$$2\cdot X, \qquad \frac{X-1}{2}, \qquad e^X$$

Example 32: Diameter and volume of a soccer ball

As manufactured products, soccer balls are subject to small fluctuations in size. Suppose that in order to be sold a soccer ball must meet a certain volume criteria. Unfortunately the manufacturer is only able to measure the diameter of the ball with an accuracy up to 1 mm, which is a discrete random variable denoted by D. Since diameter and volume are directly related, we can transform the original random variable D to a new RV

V, which describes the volume via

$$V = \frac{4}{3}\pi \left(\frac{D}{2}\right)^3.$$

Hence, for a concrete realization $D(\omega)$, the volume can be directly determined by the above transformation.

In the sequel, we will often use transformations to shift or scale random variables. More examples will be discussed in the context of continuous random variables.

3.2.1 Joint Probability Distribution and Independence

Now, having two random variables on the same probability space, we can define a joint distribution as follows.

Definition 7: Joint Probability Distribution
Let X, Y be discrete (real-valued) random variables on the same probability space $(\Omega, 2^\Omega, P)$. The function $P_{X,Y} : X(\Omega) \times Y(\Omega) \to [0,1]$ with

$$P_{X,Y}(a,b) := P(X = a \wedge Y = b) = P(X^{-1}(\{a\}) \cap Y^{-1}(\{b\})).$$

is called the joint probability distribution of X and Y.

Similar to the definition of independence between events we can define when two random variables are independent.

Definition 8: Independence of Random Variables
Let X, Y be discrete (real-valued) random variables on the same probability space $(\Omega, 2^\Omega, P)$. We call X and Y independent iff for all $a \in X(\Omega)$, $b \in Y(\Omega)$

$$P(X = a \mid Y = b) = P(X = a).$$

(Or equivalently, X and Y are independent iff $P(X = a \wedge Y = b) = P(X = a) \cdot P(Y = b)$.)

Note that we use $\wedge$ to consider the intersection of two events, i.e. $X = a \wedge Y = b$ refers to all outcomes that are mapped to a by X and to b by Y.

Example 33: Rolling two dice

Assume that X and Y are defined as in Example 31. Then X and Y are

independent, since for any $a, b \in \{1, \ldots, 6\}$,

$$\begin{aligned} P(X = a \wedge Y = b) &= P(X^{-1}(\{a\}) \cap Y^{-1}(\{b\})) \\ &= P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_1 = a\} \\ &\quad \cap \{(\omega_1, \omega_2) \in \Omega \mid \omega_2 = b\}) \\ &= P(\{(a, b)\}) = 1/36 \end{aligned}$$

and

$$\begin{aligned} P(X = a) \cdot P(Y = b) &= P(X^{-1}(\{a\})) \cdot P(Y^{-1}(\{b\})) \\ &= P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_1 = a\}) \\ &\quad \cdot P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_2 = b\}) \\ &= 6/36 \cdot 6/36 = 1/36. \end{aligned}$$

We remark that the above definition can be extended to more than two random variables by formulating conditions for mutual and pairwise independence.
If $X_1, \ldots, X_n$ are (real-valued) random variables on the same probability space, then we call $X_1, \ldots, X_n$ *(mutually) independent* iff for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$

$$P(X_1 = a_1, \ldots, X_n = a_n) = P(X_1 = a_1) \cdot \ldots \cdot P(X_n = a_n).$$

We call the random variables $X_1, \ldots, X_n$ (pairwise) independent iff all pairs $X_i, X_j, i \neq j$ are independent in the sense of Definition 8.
In the sequel, we will often consider random variables that are *independent and identically distributed (i.i.d.)*. This means that the considered random variables all follow the same probability distribution and are (mutually) independent.

3.3 Expectation and Variance

Besides the probability distribution of a discrete random variable X, other values related to X are of interest. The most important ones are the expectation and the variance of X which are defined as

- $E(X) = \sum_{x \in X(\Omega)} x \cdot P(X = x)$
- $V(X) = \sum_{x \in X(\Omega)} (x - E(X))^2 \cdot P(X = x)$

respectively. (Note that the sum might not converge, in which case the expectation/variance does not exist.) The standard deviation σ_X of X is given by $\sigma_X = \sqrt{V(X)}$. Note that $V(X)$ (and σ_X) are always non-negative.

Example 34: Exp. and Var. of Special Random Variables

Let $c \in \mathbb{R}$. Assume $X(\omega) = c$ for all $\omega \in \Omega$. Then $E(X) = c$ and $V(X) = 0$.

Let $A \subseteq \Omega$ and define a random variable I_A where

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then one can compute that $E(I_A) = 1 \cdot P(A) + 0 \cdot P(\bar{A}) = P(A)$ and $V(I_A) = P(A)(1 - P(A))$.

Often the expectation of a random variable is denoted by the Greek letter μ while the variance is often denoted by a σ^2 and σ denotes the standard deviation. The latter measures the variation using the same "units" as X and μ while σ^2 measures in squared units.

One might think that $E[X]$ is something like "the best guess for X" that we can make. However, note that even if X is discrete (e.g. $X \in \{1, 2, \ldots\}$) its expectation may not be a valid outcome (e.g. 2.6) since it is a real number. In addition, it might be a bad guess if the probability distribution of X is, for instance, bimodal as illustrated in Fig. 3.4.

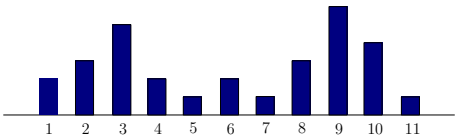
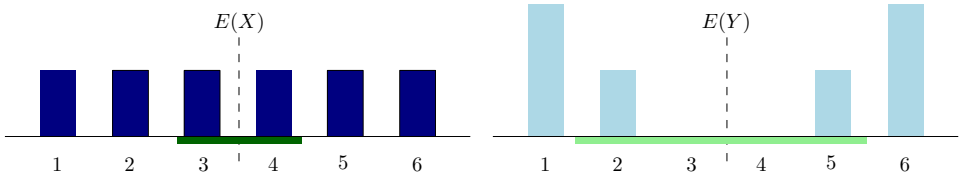


Figure 3.4: Bimodal discrete probability distribution.

Example 35:

We consider first a standard fair die with 6 sides. For such a die the expectation is given by $E(X) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = 3.5$ and the variance is $V(X) = \sum_{k=1}^{6} (k - 3.5)^2 \cdot \frac{1}{6} = 2.9$. Now let us consider the second fair die where number on sides 3 and 4 are substituted by 1 and 6 (so the die has sides $1, 2, 1, 6, 5, 6$). For this die the expectation is 3.5 as before but the variance is much larger, $V(Y) = 4.9$. Distribution of both dice are shown on the figure below (blue colors) together with the corresponding standard deviations (green colors give expectation +/− standard deviation).



3.3.1 Properties of the Expectation and Variance Operator

We are now able to state some properties of the expectation and variance.

Properties of the expectation:

Let X, Y be discrete random variables on the same probability space and assume that $E(X)$ and $E(Y)$ exist. Then, for $a, b \in \mathbb{R}$

1. $E(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\})$,
2. $E(a \cdot X + b) = a \cdot E(X) + b$,
3. $E(X + Y) = E(X) + E(Y)$.
4. If X and Y are independent, then $E(X \cdot Y) = E(X) \cdot E(Y)$.

Proof. (1.)

$$\begin{aligned} E(X) &= \sum_{x \in X(\Omega)} x \cdot P(X = x) && \text{(definition of } E(X)) \\ &= \sum_{x \in X(\Omega)} x \cdot P(\{\omega \mid X(\omega) = x\}) && \text{(definition of } P(X = x)) \\ &= \sum_{x \in X(\Omega)} x \cdot \sum_{\omega \in \Omega \wedge X(\omega) = x} P(\{\omega\}) && \text{(2nd axiom of Def. 1)} \\ &= \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) && (X \text{ is a function}) \end{aligned}$$

The remaining proofs are left as an exercise. □

Remark:

We saw that expectation is linear. However, in general expectation does not factor nicely $E(XY) \neq E(X)E(Y)$!

Properties of the variance:

We list the properties of the variance operator without proof: Let X, Y be discrete random variables on the same probability space and assume that $E(X)$, $E(Y)$, $V(X)$, and $V(Y)$ exist. Then

1. $V(a \cdot X + b) = a^2 \cdot V(X)$ for $a, b \in \mathbb{R}$,
2. $V(X) = E(X^2) - E(X)^2$
3. $V(X + Y) = V(X) + V(Y) + 2 \cdot (E(X \cdot Y) - E(X) \cdot E(Y))$.
4. If X and Y are independent, then $V(X + Y) = V(X) + V(Y)$.

For more than two independent random variables, we generalize the above properties of the expectation and variance, we have: If $X_1, \ldots, X_n$ are (mutually) independent, then

$$E(X_1 \cdot \ldots \cdot X_n) = E(X_1) \cdot \ldots \cdot E(X_n)$$

and

$$V(X_1 + \ldots + X_n) = V(X_1) + \ldots + V(X_n).$$

3.3.2 Conditional Expectations

In Section 2 we have seen that conditional probabilities are probabilities, i.e., they fulfil the axioms of probability. Hence, we can consider the corresponding "conditional" probability space and compute also expectations using conditional probabilities. Suppose we have two discrete random variables X and Y. Assume that the event $X = x$ has positive probability, i.e. $P(X = x) > 0$. Then, then expectation of Y conditioned on $X = x$ is defined as

$$E[Y|X = x] = \sum_y y \cdot P(Y = y \mid X = x).$$

In the same way, we could compute the mean of X given $Y = y$ (just switch the roles of X and Y).

Example 36: Conditional Expectation

Consider the rolling of two 6-sided fair dice D_1 and D_2 and two random variables X and Y where X = value of $D_1 + D_2$ and Y = value of D_2. We compute

$$\begin{aligned} E[X \mid Y = 6] &= \textstyle\sum_x x \cdot P(X = x \mid Y = 6) \\ &= \tfrac{1}{6}(7 + 8 + \ldots + 12) = 57/6 = 9.5. \end{aligned}$$

Note that this intuitively makes sense, as it is equal to E[value of D_1] + $E[Y \mid Y = 6] = 3.5 + 6$.

Note that the properties of the expectation carry over to conditional expectations as we are only considering a different probability function (namely, the conditional probability, which is also a probability in the sense of Def. 1).

We remark that it is possible to construct new random variables by leaving the value in the condition unspecified, i.e. considering $Z = E[X \mid Y]$. The conditional expectation as a random variable is very useful and popular construct, but beyond the scope of this course.

3.4 Higher Order Moments, Covariance and Correlation

In the previous section we have used different operators (e.g. +) to combine several random variables and considered the expectation of combinations of random variables. Thus, we already used the following general formula for the expectation of a function g

$$E(g(X)) = \sum_x g(x) \cdot P(X = x).$$

(Again the sum might not converge, in which case $E(g(X))$ does not exist.) Note that this formula can also be applied if g is not a one-to-one function, e.g. if $g(x) = x^2$. In the case $g(x) = x^i$, we call $E(X^i)$ the i-th moment of X. Note that the moments of a distribution are related to its skewness and its kurtosis. The former characterizes the degree of asymmetry while the latter characterizes the flatness or peakedness of the distribution.

3.4.1 Covariance and Correlation

While expectation, variance and standard deviation describe properties of a single random variable, the covariance and the correlation measure the level of dependence between variables:

- The covariance of two random variables X and Y is defined by
$$COV(X, Y) = E([X - E(X)][Y - E(Y)]).$$
- Assuming $VAR(X), VAR(Y) > 0$, the Pearson correlation coefficient of X and Y is a value between -1 and 1 and defined by
$$COR(X, Y) = \frac{COV(X, Y)}{\sqrt{VAR(X)}\sqrt{VAR(Y)}}.$$

The covariance is the expected product of the deviations of X and Y from their respective means. The correlation can be seen as a scaled version of the covariance having the same sign. Note that both are symmetric, i.e.

$$COV(X, Y) = COV(Y, X)$$

and

$$COR(X, Y) = COR(Y, X).$$

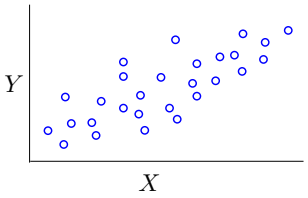


Figure 3.5: Positively correlated X and Y.

If X and Y are positively correlated, $COR(X,Y) > 0$, large values of X (positive deviations from $E(X)$) generally correspond to large values of Y. Similarly, small values of X imply small values of Y. If X and Y are negatively correlated, $COR(X,Y) < 0$, large values of X generally correspond to small values of Y (example is given in Fig. 3.5). If X and Y are uncorrelated, $COR(X,Y) = 0$, we know that there is no linear dependency between X and Y. This does, however, not imply that they are independent since other forms of dependence are possible as the following example shows.

Example 37: Uncorrelated Variables

Assume that X is a discrete random variable such that $P(X = -1) = P(X = 0) = P(X = 1) = \frac{1}{3}$ and $Y = X^2$. Then $E(X) = 0$ and $E(Y) = 2/3$. Thus,

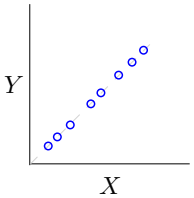
$$\begin{array}{rcl} COV(X,Y) & = & \sum_{x,y} P(X = x \wedge Y = y)(x - E(X))(y - E(Y)) \\ & = & \frac{1}{3}(-1-0)(1-\frac{2}{3}) + \frac{1}{3}(1-0)(1-\frac{2}{3}) \\ & = & 0 \end{array}$$

Thus, X and Y are uncorrelated even though Y is a function of X (the strongest form of dependence). Clearly, X and Y are not independent since, for instance,

$$P(X = 1 \wedge Y = 1) = \frac{1}{3} \neq P(X = 1)P(Y = 1) = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}.$$

Next, we list some important properties of the covariance:

- $COV(X,Y) = COV(Y,X)$
- $COV(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$
- $COV(X,X) = VAR(X)$



Finally, we note that since the correlation is scaled and has the property $-1 \leq COR(X,Y) \leq 1$, we have a maximal positive (negative) correlation at $COR(X,Y) = 1$ ($COR(X,Y) = -1$). In this case all possible realizations for (X,Y) lie on a straight line. Thus, given $X = x$ the value $Y = y$ is fixed by the line. The figure on the right illustrates this case.
Note that other correlation tests exist such as the Spearman rank correlation, which tests whether one variable is a monotone function of the other.

3.4.2 Moment Generating Function

When we discussed transformations of random variables, one example was the transformation $g(X) = e^X$. This is a very popular transformation as it can be turned into the powerful tool of moment generating functions, which are useful for deriving equations for the moments of many popular random variables.

The moment generating function for a random variable X is defined by

$$M_X(t) = E[e^{tX}] = E\left[1 + \frac{tX}{1!} + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \ldots\right], \quad t \in \mathbb{R}.$$

Here, the variable t is basically a placeholder that determines the degree k (from $(tX)^k$). For the moment generating function, the coefficient of t^k is the kth moment $E[X^k]$:

$$M_X(t) = E[e^{tX}] = E\left[\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}\right] = \sum_{k=0}^{\infty} E[X^k]\frac{t^k}{k!}$$

If $M_X(t)$ exists on an open interval around $t = 0$, then the kth moment equals the kth derivative of the moment generating function, evaluated at $t = 0$:

$$E[X^k] = \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0}$$

Another usefull equation for computing moment generating functions is

$$M_X(t) = \sum_{x \in X(\Omega)} P(X = x) \cdot e^{tx}.$$

Proof.

$$\begin{aligned}
M_X(t) &= E[e^{tX}] && \text{(definition of } M_X(t)\text{)} \\
&= E\left[\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}\right] && \text{(definition of } e^{tX}\text{)} \\
&= \sum_{k=0}^{\infty} E[X^k]\frac{t^k}{k!} && \text{(linearity of } E\text{)} \\
&= \sum_{k=0}^{\infty} \left(\sum_{x \in X(\Omega)} x^k \cdot P(X=x)\right)\frac{t^k}{k!} && \text{(definition of } E[g(X)]\text{)} \\
&= \sum_{x \in X(\Omega)} \left(\sum_{k=0}^{\infty} x^k \cdot P(X=x)\frac{t^k}{k!}\right) && \text{(reorder sums)} \\
&= \sum_{x \in X(\Omega)} P(X=x)\sum_{k=0}^{\infty} \frac{(xt)^k}{k!} && \text{(rearrange)} \\
&= \sum_{x \in X(\Omega)} P(X=x)\cdot e^{tx} && (e^x = \textstyle\sum_{k=0}^{\infty} \frac{x^k}{k!})
\end{aligned}$$

□

Example 38: $M_X(t)$ OF THE BERNOULLI DISTRIBUTION

A random variable X is Bernoulli-distributed with parameter $p \in [0,1]$ if $P(X=1) = p$ and $P(X=0) = 1-p$. Thus, it is simply the indicator function of the event "X=1". We consider its moment generating function:

$$\begin{aligned}
M_X(t) &= E[e^{tX}] \\
&= P(X=0)\cdot e^{t\cdot 0} + P(X=1)\cdot e^{t\cdot 1} \\
&= (1-p)e^0 + pe^t \\
&= 1 - p + pe^t
\end{aligned}$$

Next, we consider

$$E[X] = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = pe^0 = p \text{ and } E[X^2] = \left.\frac{d^2M_X(t)}{d^2t}\right|_{t=0} = pe^0 = p.$$

Thus, $E[X] = p$ and $V[X] = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$.

An important property of the moment generating function is that if two distributions have the same moment generating function, then they have

the same distribution, i.e. $M_X = M_Y$ implies $P_X = P_Y$.

3.5 Important Discrete Probability Distributions

Let us now consider some important discrete probability distributions.

- **Bernoulli distribution**: Consider again an experiment where there are only two possible outcomes (e.g. flipping a coin). This is called a Bernoulli trail. Let us use 0 (failure) and 1 (success) for the two possible outcomes. The probability of 1 is $P(X = 1) = p$ for some fixed $p \in [0, 1]$ and $P(X = 0) = 1 - p$. We say that X is Bernoulli distributed with parameter p, shorthand

  $$X \sim Bernoulli(p).$$

  As we have seen above, the mean of the Bernoulli distribution is $E(X) = 0(1 - p) + 1(p) = p$.

- **Binomial distribution**: We consider a sequence of n independent Bernoulli trials and count the number X of successes. Then X follows a binomial distribution, i.e. for $k \in \{0, 1, \ldots, n\}$

  $$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

  where the first factor counts all possible orderings for k successes among n trials, the second one gives the probability that in the independent trials we have k successes and the last factor describes the probability of (the remaining) $n-k$ failures. Note that one can express X as the sum of n independent Bernoulli variables $X_1, \ldots, X_n$:

  $$X = X_1 + \ldots + X_n.$$

  Thus, $E(X) = E(X_1) + \ldots + E(X_n) = p + \ldots + p = n \cdot p$. *Please use the Wikipedia link to view the plot of the distribution and learn more about the Binomial distribution.*

- **Geometric Distribution**: Consider again an experiment where the probability of an event A is $P(A) = p$. We repeat this experiment until A occurs for the first time. Let X be the random variable that describes the number of trials until A occurs for the first time. Then X is called geometrically distributed. We have $P(X = i) = (1 - p)^{i-1} \cdot p$ and $E(X) = 1/p$. The variance of X is $V(X) = (1 - p)/p$. *Please use the Wikipedia link to view the plot of the distribution and learn more about the geometric distribution.*

- **Poisson Distribution** Consider a call center where on average $\mu = 6$ calls per minute arrive. Let X be the random variable that represents the number of calls in the next minute and assume $P(X = k) = \frac{\mu^k}{k!}e^{-\mu}$. Then X is called Poisson distributed and has expectation

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{\mu^k}{k!}e^{-\mu} = \mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu} e^{\mu} = \mu$$

and variance $V(X) = \mu$ (without proof). *Please use the Wikipedia link to view the plot of the distribution and learn more about the Poisson distribution.*

Note that the Poisson distribution is the limit of the binomial distribution, when n is large and p is small. This can be shown as follows:
Let $\mu = np$. Starting from a binomial distribution we then get

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} \frac{\mu^k}{k!} \frac{\left(1 - \frac{\mu}{n}\right)^n}{\left(1 - \frac{\mu}{n}\right)^k} \\ &\approx \frac{\mu^k}{k!} e^{-\mu}, \end{aligned}$$

where we used the large n approximations $\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} \approx 1$, $\left(1 - \frac{\mu}{n}\right)^k \approx 1$ and $\left(1 - \frac{\mu}{n}\right)^n \approx e^{-\mu}$.

Example 39: Binomial vs Poisson distribution

Suppose a chip is defective in 10% of the time. You have 10 chips and the number of defective chips is described by the RVs Y or X. We would like to know the probability that no more than 1 chips are defective. For the binomial distribution $Y \sim binomial(10, 0.1)$ we get

$$\begin{aligned} P(Y \leq 1) &= P(Y = 0) + P(Y = 1) \\ &= \binom{10}{0} 0.1^0 0.9^{10} + \binom{10}{1} 0.1^1 0.9^9 \approx 0.7361. \end{aligned}$$

For the Poisson distribution $X \sim Poisson(\mu = 10 \cdot 0.1 = 1)$ on the other hand, we get

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1^1}{1!} \approx 0.7358. \end{aligned}$$

Note that even for a moderate size of $n = 10$ *the results are already very similar.*

Chapter 4

Continuous Random Variables

So far, we considered only random variables that take discrete values such as $0, 1, 2, \ldots$. However, many chance experiments can only be described by means of continuous random variables. For instance, consider a dartboard with one foot in radius. The experiment consists of throwing a dart at the board and the outcome is the point at which it hits. Then, any point in

$$S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$

is a possible outcome. Hence, the sample space is uncountably large. In this case, the defintion of random variables (in fact even that of a probability) gets more complicated.

Chapter learning objectives:

- understand the mathematical construction underlying continuous random variables in all its depths
- define appropriate continuous random variables for a variety of problems
- computation of basic properties of discrete random variables: expectation, variance, etc.
- kowledge of the most important continuous probability distributions
- extension of all these concepts to the multivariate case

4.1 σ-algebras

Consider a chance experiment where the sample space Ω contains uncountably many elements. In this case, assigning probabilities to all elements in

2^Ω poses problems. Assume, for instance, we randomly choose a real number in $[0, 1]$. If all numbers are equally likely to occur, we have to assign probability zero to each since their "sum" must be one (note that the sum over uncountably many nonzero values is undefined). Instead of giving a new definition of probabilities, we restrict ourselves to a set of events, called σ-algebra, for which we can define probabilities as in the discrete case.
If we define a random variable as a function $X : \Omega \to \mathbb{R}$, we want to reason about the probability of events such as $X = x$ for any $x \in \mathbb{R}$ or $a < X \leq b$ for any interval $(a, b]$. From the properties of probabilities (see Def. 1), we then know the probability of the disjoint union of countably many of such sets as well as the probability of the complement of such a set[1].

Definition 9: σ-ALGEBRA
A set $\mathcal{F} \subseteq 2^\Omega$ is called a σ-algebra *if*

1. $\Omega \in \mathcal{F}$,
2. $A \in \mathcal{F}$ *implies* $\bar{A} \in \mathcal{F}$.
3. *If $A_1, A_2, \ldots \in \mathcal{F}$ is a sequence of sets then*

$$A_1 \cup A_2 \cup \ldots \in \mathcal{F}.$$

The most important example of a σ-algebra, which is needed in the sequel, is the σ-algebra that is generated by a set $\mathcal{E} \subseteq 2^\Omega$. We define the smallest σ-algebra that contains $\mathcal{E}$ by

$$\sigma(\mathcal{E}) := \cap\{\mathcal{F} \supset \mathcal{E} : \mathcal{F} \text{ is a } \sigma\text{-algebra}\}.$$

Example 40: GENERATED σ-ALGEBRA

For simplicity, we consider the finite set $\Omega = \{1, 2, \ldots, 6\}$. Let $\mathcal{E} = \{\{2, 6\}, \{5, 6\}\}$. Then

$$\sigma(\mathcal{E}) = \{\{2, 6\}, \{5, 6\}, \{1, 3, 4, 5\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}, \{6\}, \ldots\}.$$

If Ω is finite, the idea is to construct a sequence of sets in an iterative fashion, where we start with $\mathcal{E}$ and obtain the next set from the previous one by joining and complementing elements of the current set. If no new element can be constructed by union or complement operations, the current set equals $\sigma(\mathcal{E})$.

Net we assume $\Omega = \mathbb{R}$ and

$$\mathcal{E} = \{(a, b] : a, b \in \mathbb{R}, a \leq b\}.$$

[1] From the two conditions in Def. 1, one can easily derive that $P(\bar{A}) = 1 - P(A)$.

The σ-algebra $\mathcal{B} := \sigma(\mathcal{E})$ is called the Borel algebra on the reals. It contains all subsets, called *Borel sets*, of $2^{\mathbb{R}}$ that can be obtained from $\mathcal{E}$ by countable union and complement operations. Note that also intervals of the form (a, b) or $[a, b]$ are Borel sets. A similar construction is possible for $\Omega = \mathbb{R}^n$. Intuitively, the Borel sets are those sets, for which we can assign a "volume", "area" or "size". Subsets of $\mathbb{R}^n$ that are not Borel sets are only of theoretical interest since for practical applications they are not of importance.
We are now able to give a more general definition of a probability space (i.e. also for sample sets with uncountably many elements).

Definition 10: Probability Space
Let Ω be a nonempty set and let $\mathcal{F} \subseteq 2^{\Omega}$ be a σ-algebra. A probability space *is a triple $(\Omega, \mathcal{F}, P)$ where the* probability measure $P : \mathcal{F} \to [0, 1]$ *is such that*

- $P(\Omega) = 1$ *and,*
- *if $A_1, A_2, \ldots \in \mathcal{F}$ is a sequence of pairwise disjoint sets, then*

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots.$$

Example 41: Discrete Probability Space

Any discrete probability space $(\Omega, 2^{\Omega}, P)$ fulfills Def. 10, since 2^{Ω} is a σ-algebra and P is as in Def. 1.

Example 42: One-dimensional Interval

Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$ and $0 \leq a < b$, $a, b \in \mathbb{R}$. Consider a probability measure $P : \mathcal{F} \to [0, 1]$ such that

$$P(\{\omega \in \Omega \mid x < \omega \leq y\}) = \begin{cases} \frac{\min(y,b) - \max(x,a)}{b-a} & \text{if } \max(x, a) < \min(y, b) \\ 0 & \text{otherwise.} \end{cases}$$

Similar to how we extended the set $\mathcal{E} = \{(a, b] : a, b \in \mathbb{R}, a \leq b\}$ to a σ-algebra, we can show that if P is a probability measure and defined as above for all half-open intervals, its value for the remaining sets in $\mathcal{B}$ is uniquely determined.

4.2 Continuous Random Variables

Definition 11: Real-valued Random Variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real-valued random variable *on $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$ such that for all $A \in \mathcal{B}$*

$$X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}.$$

The above definition ensures that if we want to know the probability that X is in some Borel set A, we can consider the inverse image of A with respect to X, for which we know its probability.

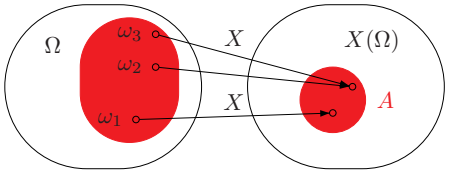


Figure 4.1: Is the inverse image of A w.r.t. X an element of $\mathcal{F}$?.

Clearly, we can define a probability measure $P_X : \mathcal{B} \to [0, 1]$ by setting $P_X(A) := P(\{\omega \mid X(\omega) \in A\}) = P(X^{-1}(A))$ and use similar notations as in the discrete case (e.g. $P(a < X \leq b)$).

Definition 12: Cumulative Probability Distribution

Let X be a real-valued random variable on $(\Omega, \mathcal{F}, P)$. The function $F : \mathbb{R} \to [0, 1]$ with $x \mapsto F(x) := P(X \leq x)$ is called the cumulative probability distribution (CDF) of X.

Example 43: One-dimensional Interval

Assume that X is a randomly chosen point in the interval $[a, b]$ and $(\Omega, \mathcal{F}, P)$ is as in Ex. 42. Then

$$F(y) = P(X \leq y) = \begin{cases} \frac{y-a}{b-a} & \text{if } y \in [a, b], \\ 1 & \text{if } y > b, \\ 0 & \text{otherwise.} \end{cases}$$

We call a random variable X on $(\Omega, \mathcal{F}, P)$ *discrete* if $X(\Omega)$ is a discrete set (finite or countably infinite). We call X *continuous* if $X(\Omega)$ contains uncountably many elements and there exists a non-negative and integrable function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$, called *density*, with

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

Clearly,

$$\int_{-\infty}^{\infty} f(y)\, dy = 1$$

since $F(\infty) = 1$, but note that $P(X = y) \neq f(y)$.

Example 44: One-dimensional Interval

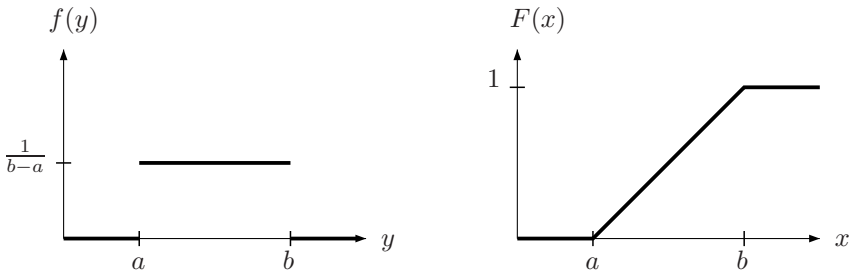


Figure 4.2: Density and cumulative probability distribution of a random variable X.

Assume that X is as in Ex. 43. Then X is a continuous random variable since the (constant) function f with

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{if } y \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

is the density of X. We can verify this by calculating

$$F(x) = \int_{-\infty}^{x} f(y)\ dy = \int_{a}^{x} \frac{1}{b-a}\ dy = \left[\frac{y}{b-a}\right]_{a}^{x} = \frac{x-a}{b-a} = P(X \leq x)$$

for $x \in [a,b]$, $F(x) = 0$ for $x < a$, and $F(x) = 1$ for $x > b$. Figure 4.2 shows a plot of the functions f and F.

The distribution in the above example is called the (continuous) uniform distribution.
We summarize the most important properties of the CDF $F(x)$ and the density $f(x)$ in the following table:

	$F(x)$	$f(x)$
domain	$\mathbb{R}$	$\mathbb{R}$
codomain	$[0,1]$	$\mathbb{R}_{\geq 0}$
monotonicity	increasing	not necessarily
relation	$= \int_{-\infty}^{x} f(y)dy$	$= \frac{dF(x)}{dx}$

The expectation and variance of a continuous random variable X with density f are defined as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\ dx \text{ and } V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\ dx.$$

Note that these integrals may not exist, in which case the expectation/variance is undefined.

What from the discrete case carries directly over to the continuous case?

Many results for discrete random variables carry over to the continuous setting. For instance, properties of the expectation and variance (e.g. $E(X+Y) = E(X) + E(Y)$) or results concerning the combination/transformation of random variables, joint distributions, stochastic independence, etc. Also, we define conditional expectation, covariance, correlation, and higher-order moments for continuous random variables in an equivalent way. All results of the previous sections carry over to the continuous case (as long as the corresponding integrals exist). We do not repeat these definitions here as the only things that change compared to the discrete case are that we replace the sum by an integral and the probability of a value x by the density $f(x)$.

4.3 Important Continuous Distributions

Besides the uniform distribution that we have seen above, there are some other important continuous distributions:

Uniform Distribution

We already described the uniform distribution in the examples above. In summary we get the following properties if X is uniformly distributed on the interval (a, b) (denoted by $X \sim U(a, b)$):

- Its density is constant on the interval (a, b), that is, $f(x) = 1/(b-a)$ for $x \in (a, b)$ and $f(x) = 0$ otherwise.
- If we want to know the probability that X falls into an subinterval of (a, b) of length d we get

  $$P(X \in (c, c+d)) = d/(b-a), \qquad (c, c+d) \subseteq (a, b)$$

  which means that it is proportional to the length d (and independent of c, cf. Figure 4.2).
- The expectation is given by the midpoint of the interval $E(X) = (a+b)/2$.

Exponential Distribution

Let $\lambda > 0$. We say that a continuous random variable X is exponentially distributed with parameter λ (denoted by $X \sim Exp(\lambda)$) if the density of X

is such that, for $t \in \mathbb{R}$,

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative probability distribution of X is then given by

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(t)\ dt = \int_{0}^{x} \lambda \cdot e^{-\lambda t}\ dt = 1 - e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The exponential distribution is often used to describe waiting times or interarrival times since in many real-world systems where a sequence of rare events plays an important role, the assumption that the time between these events is exponentially distributed is used. For instance, the time of radioactive decay is exponentially distributed. This is related to the fact that these events are assumed to occur *spontaneously* and that the exponential distribution has the so-called memoryless property:
Assume that the random variable X describes a waiting time and we already know that $X \geq t$ and ask for the probability that $X \geq t + h$ (we have to wait for additional h time units). Then, if X is exponentially distributed we can verify that

$$P(X \geq t + h \mid X \geq t) = P(X \geq h), \qquad t, h > 0.$$

The exponential distribution is the only continuous distribution that is memoryless. In fact, it is possible to derive from the memoryless property that the distribution of X must be exponential. Similarly, one can show that the geometric distribution is the only discrete distribution that is memoryless.

Example 45: Bus waiting paradox

Assume that Bob arrives at a bus stop where buses arrive on average every 10 min. The interarrival times X of the buses are exponentially distributed. Bob asks one of the people waiting there how long he is already waiting. Whatever the answer of that person is (e.g. $t = 20$ min or 1 min), it does not change the probability that Bob has to wait for h minutes for the next bus.

Normal distribution

The normal distribution is one of the most important continuous distributions since it naturally arises when we sum up many random variables. Therefore, measurement errors and fluctuations are often (approximately) normally distributed. Also measurements of the length or size of certain objects give often normally distributed results.

A normally distributed random variable X with mean μ and variance σ^2 (denoted by $X \sim N(\mu, \sigma^2)$) has the density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2},$$

The density of the standardized normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is then defined as

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

and its cumulative distribution function is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz.$$

Assume we are interested in the probability that $X \sim N(\mu, \sigma^2)$ falls into a certain interval (a, b). Since the solution of the above integral is difficult to compute, we can standardise X and use the fact that we have a table with the values for Φ.

Definition 13: Standardized Random Variables.
For any random variable X with finite $E(X) = \mu$ and finite $VAR(X) = \sigma^2$, we define its standardised version

$$X^* = \frac{X - \mu}{\sigma}.$$

Then $E(X^*) = 0$ and $VAR(X^*) = 1$ since

$$E(X^*) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0$$

and

$$VAR(X^*) = VAR\left(\frac{X-\mu}{\sigma}\right) = \frac{VAR(X)}{\sigma^2} = 1.$$

For $X \sim N(\mu, \sigma^2)$ this means that we can find the cumulative distribution of X^* in a standard normal table and use it to compute that of X. Using $X^* = \frac{X-\mu}{\sigma} \iff X = X^*\sigma + \mu$ we get

$$P(X \leq x) = P(X^*\sigma + \mu \leq x) = P(X^* \leq \frac{x-\mu}{\sigma}).$$

Example 46: Computing Normal Probabilities

The average salary of a full-time employee in Germany is $\mu = 45240$ Euros per year (in 2017, brutto). Assuming a normal distribution and a

standard deviation of $\sigma = 10000$ Euros, we can compute the proportion of employees that earn between, say 30 000 and 50 000 Euros per year as follows:

$$\begin{aligned} P(30000 < X < 50000) &= P\left(\tfrac{30000-\mu}{\sigma} < \tfrac{X-\mu}{\sigma} < \tfrac{50000-\mu}{\sigma}\right) \\ &= P(-1.524 < X^* < 0.476) \\ &= \Phi(0.476) - \Phi(-1.524) \\ &\approx 0.68439 - 0.06426 = 0.6203. \end{aligned}$$

Next, we can compute the income limit of the poorest 3% of the employees, i.e. find x such that $P(X < x) = 0.03$:

$$P(X < x) = P\left(X < \tfrac{x-\mu}{\sigma}\right) = \Phi(\tfrac{x-\mu}{\sigma}) = 0.03$$

Thus,

$$x = \mu + \sigma\Phi^{-1}(0.03) = 45240 + 10000 \cdot (-1.88) = 26440$$

where we used that $\Phi(-1.88) = 0.03$.

Finally, we note that if a random variable X is normally distributed with expectation $\hat{\mu}$ and variance $\hat{\sigma}^2$, written $X \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, then, for $a, b \in \mathbb{R}$, $aX + b$ has distribution $\mathcal{N}(a\hat{\mu} + b, (a\hat{\sigma})^2)$.

Gamma distribution

The Gamma distribution generalizes a number of other distributions (e.g. exponential) and has a rather flexible shape compared to other continuous distributions. Therefore it can be used to fit very different types of data.
A random variable X is Gamma distributed with parameters α and β (written $X \sim Gamma(\alpha, \beta)$) if its density is given by

$$f(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} x^{\alpha-1} \exp\left(\frac{-x}{\beta}\right), \qquad x > 0, \alpha > 0, \beta > 0$$

and $f(x) = 0$ whenever $x \leq 0$. Thus, it assigns positive probability only to positive values of X and also the parameters α (for the shape) and β (for the scaĺe) must be positive. The gamma function $\Gamma(\alpha)$ is used as a normalization factor of the density. Often α is a positive integer and then $\Gamma(\alpha) = (\alpha - 1)!$ which means that the density (and its integral) is easy to evaluate (for plot of the Gamma densities please visit https://upload.wikimedia.org/wikipedia/commons/e/e6/Gamma_distribution_pdf.svg).

How do the parameters α and β influence the mean and the variance of the distribution? It can be shown that the mean of $X \sim Gamma(\alpha, \beta)$ is $\mu = \alpha\beta$ and the variance is $\sigma^2 = \alpha\beta^2$. Further interesting properties of the Gamma distribution are:

- For large α and β, the distribution is very similar to a normal distribution (with mean $\mu = \alpha\beta$ and variance $\sigma^2 = \alpha\beta^2$)
- If $\alpha = 1$ and $\beta = 1/\lambda$, then $X \sim Exp(\lambda)$.
- If $\alpha = \nu/2$ and $\beta = 2$ then the distribution is a chi-squared distribution with ν degrees of freedom (where ν is a positive integer). In that case, we have the distribution of a sum of the squares of ν independent standard normal random variables. The chi-squared distribution is often used for hypothesis testing (goodness of fit tests).

4.4 Multivariate Random Variables

So far we only considered one dimensional RVs, i.e. the outcome of the mapping X leads to a one dimensional result. Imagine now, that during a measurement you obtain multiple quantities at once. For example during a health screening, you obtain the age, height, weight, ... for each participant. In order to properly describe the outcome, we need the following definition:
Definition 14: Multivariate Random Variable
Let Ω be a sample space and

$$X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$$

n one dimensional RVs. The vector $X = (X_1, X_2, \ldots, X_n)$ is then called an n-dimensional random variable *or an n-dimensional* random vector. *A vector of realizations of the corresponding one dimensional RVs $x = (x_1, x_2, \ldots, x_n)$ is called* realization *of X.*

If the sample space of $X = (X_1, \ldots, X_n)$ is countable, we call X a discrete n-dimensional RV. Since each entry X_i in the vector of our multivariate RV X is a one dimensional RV, we can adopt Definition 7 for the joint distribution of the vector entries. The generalization to more than two entries is straightforward.
In a similar fashion to the one dimensional case, we can also define the cumulative probability distribution for multivariate RVs.
Definition 15: Cumulative Probability Distribution
The cumulative distribution function $F : \mathbb{R}^n \to [0, 1]$ *of a RV $X = (X_1, \ldots, X_n)$ is defined as*

$$F(x) = P(X_1 < x_1, \ldots, X_n < X_n).$$

In an analogous way we lift properties such as expectation and covariance

to the multi-dimensional case.
To ease the notation, we consider the case $n = 2$ in the following and call the two one dimensional RVs X and Y, i.e., our multivariate RV is (X, Y). Suppose now, that there are N possible realizations $x_1, \ldots, x_N$ of X and K possible realizations $y_1, \ldots, y_K$ of Y. The possible values $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, K\}$ then form the index sets I and J. Note that N and K might be (countably) infinite and that N and K might in general be different.
Besides the probability of a certain realization $P(X = x_i, Y = y_j)$ obtained from the joint distribution, other interesting quantities are $P(X = x_i)$ and $P(Y = y_j)$, i.e, we keep one of the variables fixed, while the other still changes. This leads to the following definition.

Definition 16: Marginal Distribution
Given a RV (X, Y) and its joint distribution $P(X = x, Y = y)$, the marginal distribution *of X is defined as*

$$P(X = x_i) = \sum_{j \in J} P(X = x_i, Y = y_j)$$

for all $i \in I$. In a similar fashion the marginal distribution of Y is given by

$$P(Y = y_j) = \sum_{i \in I} P(X = x_i, Y = y_j)$$

for all $j \in J$. For more than two dimensions, we sum over all variables but one. For two dimensional RV a well-arranged representation for the joint and marginal distribution is a so-called contingency table.

Example 47: Contingency Table

Consider the RV (X, Y) where X has two possible realizations x_1 and x_2 and Y has the three possible realizations y_1, y_2, y_3.

Y \ X	x_1	x_2	
y_1	*0.1*	*0.25*	*0.35*
y_2	*0.1*	*0.4*	*0.5*
y_3	*0.05*	*0.1*	*0.15*
	0.25	*0.75*	

The values form the joint distribution are in the inside of the table, for example $P(X = x_2, Y = y_2) = 0.4$. The sum of each column gives the marginal distribution for X, the sum of each row for Y. For example $P(X = x_1) = 0.1 + 0.1 + 0.05 = 0.25$ and $P(Y = y_3) = 0.05 + 0.1 = 0.15$. Note that the sum of all entries within the table, as well as the sum of the respective marginal probabilities is 1.

In the continuous case (i.e. Ω is not discrete), we have a continuous n-dimensional RV with an n-dimensional joint density $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and,

compared to the discrete case, we replace summations by integrals in the definitions given above.

Example 48: Bivariate continuous distribution

Consider the function

$$f(x,y) = \begin{cases} c \cdot \exp(-(x+2y)), & \text{if } 0 \leq x, y \leq \infty, \\ 0, & \text{else.} \end{cases}$$

How do we have to choose c such that f becomes a density? We determine c such that the integral becomes 1, i.e.

$$\begin{aligned} 1 &= \int\int f(x,y)dxdy = \int_0^\infty \int_0^\infty c \cdot \exp(-(x+2y))dxdy \\ &= c\int_0^\infty e^{-2y}dy = \frac{c}{2} \Rightarrow c = 2. \end{aligned}$$

We can now compute the marginal distributions

$$\begin{aligned} f_X(x) &= 2\int_0^\infty \exp(-(x+2y))dy = e^{-x}\int_0^\infty 2e^{-2y}dy = e^{-x}, \\ f_Y(y) &= 2\int_0^\infty \exp(-(x+2y))dx = 2e^{-2y}\int_0^\infty e^{-x}dx = 2e^{-2y}, \end{aligned}$$

and the expectations

$$\begin{aligned} E(X) &= \int_0^\infty \int_0^\infty 2x\exp(-(x+2y))dxdy = \int_0^\infty xe^{-x}dx = 1 \\ E(Y) &= \int_0^\infty \int_0^\infty 2y\exp(-(x+2y))dxdy = \int_0^\infty 2ye^{-2y}dy = 0.5 \\ E(XY) &= \int_0^\infty \int_0^\infty 2xy\exp(-(x+2y))dxdy \\ &= \int_0^\infty xe^{-x}dx\int_0^\infty 2ye^{-2y}dy = E(X)\cdot E(Y) = 0.5. \end{aligned}$$

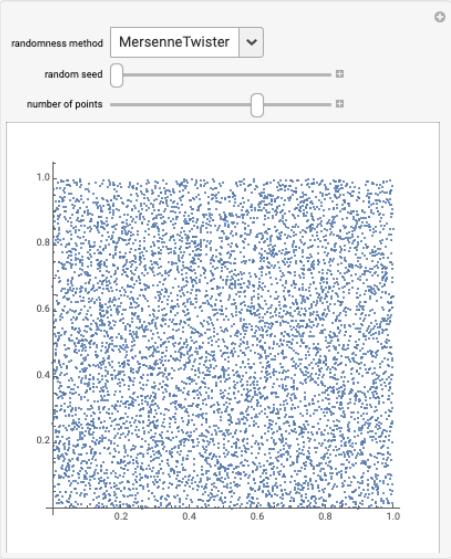
From this it is obvious that

$$Cov(X,Y) = 0$$

as expected, since X and Y are independent, because $f(x,y) = f_X(x) \cdot f_Y(y)$.

Chapter 5

Generation of Random Variates



The generation of independent samples, i.e. random variates, that follow a specific distribution is an important part of the popular sampling methods in statistics. All common programming languages provide methods to generate pseudo random numbers, i.e. numbers that look as if they random, but are generated using a deterministic algorithm, a so-called pseudo random number generator. On the right, you see the results of the Mersenne Twister, a popular pseudo random number generator. It generates numbers that are uniformly distributed on the area $(0, 1) \times (0, 1)$. The screenshot is taken from the WOLFRAM Demonstrations Project.
In the sequel, we will not discuss algorithms to generate pseudo random numbers but focus on the transformation of uniformly distributed numbers such that the results are variates that follow a certain distribution.

Chapter learning objectives

- how to exploit relationships between random variables for the generation of random variates
- understand the concept of the inverse transform method for the con-

tinuous and the discrete case

- understand the concept of rejection-based approaches
- implemention of concrete random variate sampling algorithms

5.1 Generating Discrete Random Variates

Assume we want to generate random numbers that are realizations of a discrete random variable X. For discrete distributions there are two possibilities.

a) We either use a specific relationship between X and $U \sim U(0,1)$ or

b) We split the interval $(0,1)$ into disjoint intervals and do a case distinction as explained below.

Example 49: Generating Bernoulli distributed numbers

Let $U \sim U(0,1)$ and $p \in (0,1)$. We define

$$X = \begin{cases} 1 \text{ if } U < p, \\ 0 \text{ if } U \geq p. \end{cases}$$

Then $P(X=1) = P(U<p) = p$ and $P(X=0) = 1-p$. Thus, X follows the Bernoulli distribution with probability p.
Algorithmically, we first generate U, set X accordingly and return X.
A Python code for this is (after setting the value for p):

```
1  import numpy as np
2  U = np.random.rand()
3  X = int(U < p)
```

Example 50: Generating binomially distributed numbers

Let $U_1, \ldots, U_n$ be independent and $U(0,1)$-distributed. Moreover, let $p \in (0,1)$. We define

$$X_i = \begin{cases} 1 \text{ if } U_i < p, \\ 0 \text{ if } U_i \geq p, \end{cases}$$

and $Y = \sum_{i=1}^{n} X_i$. Then $X_1, X_2, \ldots, X_n$ are Bernoulli distributed (with parameter p) and thus their sum Y is binomially distributed with parameters n and p (see Section 3.5).
Algorithmically, we generate $U_1, \ldots, U_n$, set $X_1, X_2, \ldots, X_n$ accordingly and return $Y = \sum_{i=1}^{n} X_i$.
A Python code for this is (after setting the values for n and p):

```
import numpy as np
U = np.random.rand(n)
X = np.sum(U < p)
```

Example 51: GENERATING GEOMETRICALLY DISTRIBUTED NUMBERS

Let $U_1, U_2, \ldots$ be independent and $U(0,1)$-distributed. Moreover, let $p \in (0,1)$. We define

$$X = \min_i \{U_i < p\}.$$

Then, clearly, $P(X = 1) = p$ and it is easy to see that $P(X = k) = (1-p)^{k-1}p$.
Algorithmically, we set $X = 1$ and perform a while-loop in which we generate U and check whether $U > p$ (loop condition). If this is the case we set $X = X + 1$. After the loop we return X. A Python code for this is (after setting the value for p):

```
import numpy as np
X = 1
while (np.random.rand() > p):
X += 1
X
```

5.1.1 Interval Method

The splitting of the interval $(0,1)$ can be done for any discrete random variable X. Assume that X takes the values $x_0, x_1, \ldots$ and $P(X = x_0) = p_0$, $P(X = x_1) = p_1, \ldots$. Then we divide the interval $(0,1)$ into the subintervals

$$\begin{array}{lllll} A_0 & = & (0, p_0), & & \\ A_1 & = & & [p_0, p_0 + p_1), & \\ A_2 & = & & & [p_0 + p_1, p_0 + p_1 + p_2). \\ \ldots & & & & \end{array}$$

Then we let $X = x_i$ if U falls into interval A_i. Note that since A_i has length p_i the probability that U falls into A_i is exactly p_i. Hence X has the desired distribution.
Algorithmically, the naive approach to find the index i with

$$p_0 + \ldots + p_{i-1} \leq U < p_0 + \ldots + p_i$$

is performed as follows:

1. initialize $c = p_0$ for the cumulative value,
2. iteratively set $c = c + p_k$ in the k-th execution of a while loop that is only executed if $U \geq c$.

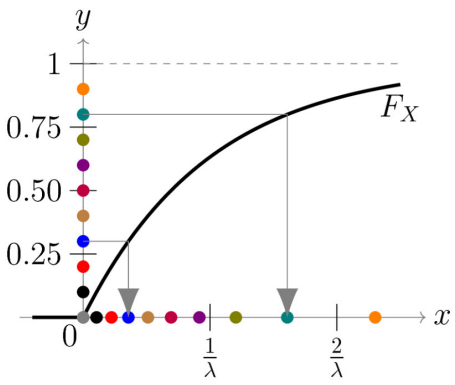
If $U < c$ holds after i iterations, then $X = x_i$.
A Python code for this is (when p_i is stored in $p(i)$ as in Python array indices start at 0):

```
1    import numpy as np
2    c = p[0] # set to p_0
3    i = 0
4    U = np.random.rand()
5    while (U >= c):
6    i += 1
7    c += p[i] # set to p_i and add to c
8    X = i
```

There is a simple way to improve the efficiency of the above approach. If we sort the values $x_0, x_1, \ldots$ according to their probability $p_0, p_1, \ldots$ we can ensure that large intervals (high probabilities) are considered first. Hence it is more likely that the number of iterations is small since U is more likely to fall in one of the first intervals.
In the case that X can take infinitely many values with positive probability, the intervals can be constructed on the fly. However, care has to be taken in how the intervals are traversed in order to ensure convergence of the while loop.
We remark that Python also offers functions for generating random variates for many well-known continuous and discrete distributions (e.g. package numpy.random).

5.2 Inverse Transform Method

Several approaches exist for the generation of independent samples that follow certain probabilities distributions.
Here, we present the most common one: inverse transform sampling, also called the inverse transform method. It can be used for any distribution (discrete or continuous) whenever the inverse F^{-1} of the cumulative distribution function (CDF) F exists.



The main idea of the method is based on the following intuition: If we plot the CDF F_X of a RV X and randomly choose a point y on the $[0, 1]$-y-axis (all points are equally likely), then the corresponding value $F_X^{-1}(y) = x$ on the x-axis will have distribution F_X. In the above plot (source: Wikipedia), this is illustrated for the exponential distribution with parameter λ.

Example 52: Generating exponentially distributed numbers

We have the CDF $F(x) = 1 - e^{-\lambda x}$ for $X \sim Exp(\lambda)$. Thus, $y = F(x)$ is a number between 0 and 1. Solving for x yields

$$\begin{aligned} & y := F(x) &=& \; 1 - e^{-\lambda x} \\ \iff & 1 - y &=& \; e^{-\lambda x} \\ \iff & \ln(1-y) &=& \; -\lambda x \\ \iff & x &=& \; -\tfrac{1}{\lambda}\ln(1-y) = F^{-1}(y). \end{aligned}$$

Thus, if U is uniformly distributed on $(0,1)$, then the random variable $X = F^{-1}(U) = -\frac{1}{\lambda}\ln(1-U)$ is exponentially distributed with parameter λ. Moreover, since $1-U$ has the same distribution as U, the random variable $X = -\frac{1}{\lambda}\ln(U)$ must also be exponentially distributed with parameter λ.

Next, we formulate the inverse transform sampling algorithm for any given CDF F_X with inverse F_X^{-1}:

1. Generate $U \sim U(0,1)$.
2. Return $X = F^{-1}(U)$.

Next, we prove that $X = F^{-1}(U)$ has the cumulative probability distribution F as desired: First observe that

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x))$$

where we used that F is monotone in the last step. Finally, we note that because U is uniformly distributed, $P(U \leq y) = y$ for any $y \in [0,1]$ and in particular for $y = F(x)$. Hence, $P(U \leq F(x)) = F(x)$ and thus $P(X \leq x) = F(x)$ holds.

5.2.1 Inverse transform sampling: the discrete case

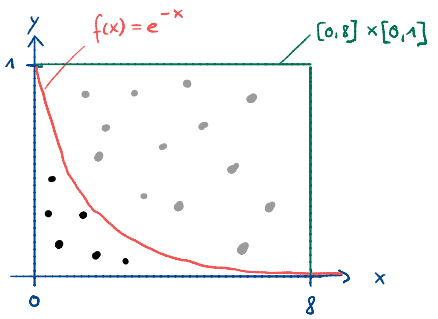
In Section 5.1.1 we developed an interval method to sample discrete distributions. How does this relate to the idea of inverse transform sampling? For the cumulative distribution function F_X of a discrete random variable X we define $F_X^{-1}(y) = \inf\{x \mid F_X(x) \geq y\}$. But then, inverse transform sampling means that we look for the interval into which $U \sim U(0,1)$ falls on the y-axis (which corresponds to the while-loop in Section 5.1.1) and choose the smallest x such that $F_X(x) \geq U$. Hence, inverse transform sampling for discrete distributions is identical to the interval method discussed before.

5.3 Rejection Sampling

Random variates can also be generated using acceptance-rejection sampling. The idea is similar to that of Monte-Carlo methods for computing complicated integrals and is best explained starting with the example of an exponential distribution with parameter $\lambda = 1$.

Example 53: GENERATING EXPONENTIALLY DISTRIBUTED NUMBERS

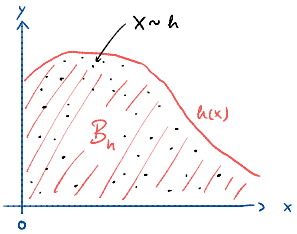
Assume we want to generate $X \sim exp(1)$. The corresponding density is $f(x) = e^{-x}$. Now, consider a rectangle, which includes most of f (if we choose, say, $[0, 8] \times [0, 1]$ as illustrated on the right). We now repeatedly draw uniformly distributed numbers $U_1 \sim U(0, 1)$ and $U_2 \sim U(0, 8)$.



The numbers U_1 and U_2 give use points (U_2, U_1) that are uniformly distributed within the 2-dimensional rectangle. Whenever the pair $(x, y) = (U_2, U_1)$ is below the density f (black points), i.e. $y \leq f(x)$, we keep x as a sample and otherwise, we reject x (illustrated as grey pairs (x, y)). Observe that if x is small, $f(x)$ is large and it is very likely that we keep x as the point (x, y) falls below the red line. Indeed, it is easy to show that the samples that we keep have the desired distribution $f(x) = e^{-x}$ (up to the approximation we made by truncating the x-axis at 8).

For the general method, we first define what the *body* B_h of a nonnegative integrable function h on $\mathbb{R}^d$ is.

$$B_h = \{(x, y) : x \in \mathbb{R}^d, 0 \leq y \leq h(x)\}.$$



Now, observe the following: if we sample pairs (X, Y) that are uniformly distributed on B_h, then X has a density that is proportional to h. This is simply because in those regions where h is large, we will have more samples (X, Y) and only few where h is small (see illustration on the right). Moreover, if we generate $U \sim U(0, 1)$ and X, independently of U such that its density is proportional to h, then $(X, Uh(X))$ is uniformly distributed on B_h. Intuitively, this is because in regions where h is large, we have many realizations for X and we use U to choose one of them (uniformly distributed).

Now, consider the probability density f of the random variable for which

we want to sample variates. We choose another function g that majorizes f, i.e. that is at least as large as f at all points $x \in \mathbb{R}$, $g(x) \geq f(x)$ for all x. Let $c > 0$ be the scaling constant such $\frac{1}{c}g(x)$ is a density *(you should know how to determine c!)*. We then generate $U \sim U(0,1)$ and X such that it has density $\frac{1}{c}g(x)$. If $Ug(X) \leq f(X)$ then we keep X. Otherwise, we reject it. Using pseudocode, we summarize the algorithm as follows:

i) Generate $U \sim U(0,1)$.

ii) Generate X, independent of U, distributed according to density $\frac{1}{c}g(x)$.

iii) If $Ug(X) \leq f(X)$ then return X ('accept'); otherwise go back to i) ('reject');

Using the arguments before, we note that we sample uniformly points $(X, Ug(X))$ on B_g, but partition B_g into B_f and its complement $B_g \setminus B_f$. We accept whenever we hit the area B_f. Hence, the x-components of our samples are distributed according to f. Moreover, the acceptance probability is

$$\begin{aligned} P(\text{acceptance} \mid X = x) &= P\left(Ug(X) \leq f(X) \mid X = x\right) \\ &= P\left(U \leq \tfrac{f(X)}{g(X)} \mid X = x\right) \\ &= \tfrac{f(x)}{g(x)} \end{aligned}$$

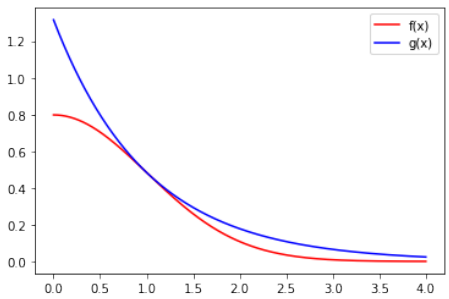
Unconditioning and recalling that X has density $\frac{1}{c}g(x)$ yields

$$P(\text{acceptance}) = \int_{-\infty}^{+\infty} \frac{f(x)}{g(x)} \frac{1}{c} g(x) dx = \frac{1}{c} \int_{-\infty}^{+\infty} f(x) dx = \frac{1}{c}.$$

Hence, in order to achieve a high probability for acceptance, we must choose g such that only a small scaling constant c is needed, i.e. g must majorize f but only slightly such that only little scaling to a density is needed.

Example 54: GENERATING NORMALLY DISTRIBUTED NUMBERS

Assume we want to generate $X \sim N(\mu, \sigma)$. Since by shifting and scaling, we can express X as $X = \sigma Z + \mu$, where $Z \sim N(0,1)$, it is enough to concentrate on the standard normal distribution.



Another simplification is that it is enough, if we can sample the absolute value $|Z|$ because the density is symmetric. We can simply sample the sign by choosing $+$ and $-$ each with probability $1/2$. (Sample $U \sim U(0,1)$ and choose $+Z$ if $U < 1/2$ and $-Z$ if $U \geq 1/2$.) The density of the non-negative RV $|Z|$ is

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, x \geq 0.$$

Next, we need a function g that majorizes f and is proportional to a density, from which we can sample easily. We choose the density $r(x) = e^{-x}$ because we can use the inverse transform method to sample $X \sim exp(1)$ and it has the same exponential form as f. Now, we need to scale up r to a function g that majorizes f. Observe that the ratio $f(x)/r(x) = \sqrt{2/\pi} \cdot e^{x-x^2/2}$ has its maximal value at $x = 1$ (you find this when you set the derivative of $f(x)/r(x)$ to zero!). *Then, we determine $c = \sqrt{2e/\pi}$ by setting $f(1) = c \cdot r(1)$ because at $x = 1$ we want that $f(x) = g(x)$ (the curves coincide at this point). Checking*

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2} \leq \sqrt{2e/\pi} \cdot e^{-x} = c \cdot r(x), for\ all\ x \geq 0$$

we find a function $g(x) = c \cdot r(x)$ that majorizes f and apply the following rejection algorithm to generate $Z \sim N(0,1)$:

i) Generate $U_1 \sim U(0,1)$.

ii) Generate X exponentially distributed with $\lambda = 1$ by first generating $U_2 \sim U(0,1)$ and then setting $X = -\ln(U_2)$.

iii) If $U_1 g(X) \leq f(X)$ then return $|Z| := X$ ('accept'); otherwise go back to i) ('reject');

iv) Generate $U_3 \sim U(0,1)$. Set $Z := |Z|$ if $U_3 < 1/2$ and $Z := -|Z|$ otherwise.

Note that the algorithm can be simplified when noting that $\frac{f(X)}{g(X)} = e^{-(X-1)^2/2}$ and thus

$$\begin{aligned} Ug(X) \leq f(X) \quad &\Longleftrightarrow \quad U \leq \tfrac{f(X)}{g(X)} = e^{-(X-1)^2/2} \\ &\Longleftrightarrow \quad -\ln(U) \geq (X-1)^2/2. \end{aligned}$$

Since $-\ln(U)$ is exponentially distributed with rate 1, we simply need two of those generated in step ii), say X_1 and X_2 and only have to check whether $X_2 \geq (X_1 - 1)^2/2$ in step iii) and set $|Z| := X_1$ if so.

Chapter 6

Laws of Large Numbers

In this chapter, we will discuss some fundamental results of probability theory: the laws of large numbers. They formalize and generalize what is intuitive for most of us: if we consider many independent replications of the same chance experiment, we can, based on the outcomes, approximate the characteristics of the average outcome. E.g. consider flipping a fair coin very often. Then, we expect that approximately half of the time we see heads. More formally, consider n realizations $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, ..., $x_n = X(\omega_n)$ of a random variable X with finite expectation μ. They could, for instance, be generated by repeating an experiment n times under equal conditions. Intuitively, for large n the sample mean approximates μ, i.e.,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mu.$$

In this chapter, we discuss important aspects related to this approximation.

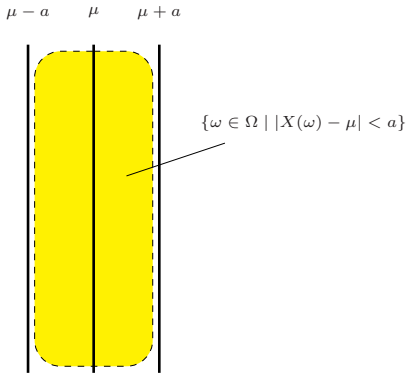
Chapter learning objectives

- understand the consequences of the laws of large numbers and the central limit theorem
- apply Chebyshev's inequality
- apply the central limit theorem

6.1 Chebyshev's inequality

In order to determine the quality of the above approximation, we recall *Chebyshev's Inequality*.

Let Y be a random variable with finite expectation $E(Y)$ and finite variance $VAR(Y)$. Assume that besides $E(Y)$ and $VAR(Y)$, nothing is known about Y (e.g. the cumulative probability distribution). Our aim is to reason about the deviation of Y from its expectation. According to Chebyshev's Inequality (see figure on the right for an illustration), for any $a > 0$



$$P(|Y - E(Y)| \geq a) \leq \frac{VAR(Y)}{a^2}. \tag{6.1}$$

Note that the proof of the theorem is quite straight forward.

6.2 Weak Law of Large Numbers

In order to apply the inequality above, we assume that $x_1, x_2, \ldots, x_n$ are realizations of independent random variables $X_1, X_2, \ldots, X_n$, all having the same distribution as X. Define

$$Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then $E(Z_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$ and, if X has finite variance $VAR(X) = \sigma^2$,

$$VAR(Z_n) = VAR\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} VAR(X_i) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(See also page 36 for the properties of the variance operator.) Thus, Eq. (6.1) gives us, for any $\epsilon > 0$,

$$P(|Z_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n \cdot \epsilon^2}.$$

If ϵ is given, we can make $\frac{\sigma^2}{n \cdot \epsilon^2}$ arbitrarily small by increasing n. Thus, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|Z_n - \mu| \geq \epsilon) = 0 \quad \text{(or, equivalently, } \lim_{n \to \infty} P(|Z_n - \mu| < \epsilon) = 1),$$

which is known as the *weak law of large numbers*[1]. Here, the sequence $\{Z_n\}_{n \geq 1}$ of random variables converges "weakly" since only the corresponding probabilities converge.

[1] Note that, as opposed to the strong law of large numbers, for the weak law of large numbers pairwise independence of $X_1, \ldots, X_n$ is already sufficient and (mutual) independence is not necessary.

Example 55: Bernoulli Trials

Consider a sequence of n Bernoulli trials and for $1 \leq j \leq n$, let X_j be the random variable that is 1 if the j-th trial is success and 0 otherwise. Moreover, let $P(X_j = 1) = p$ for all j. Then $E(X_1) = E(X_2) = \ldots = E(X_n) = p$. According to the weak law of large numbers, for any $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{j=1}^{n} X_j - p\right| < \epsilon\right) \to 1$$

as $n \to \infty$.

6.3 Strong Law of Large Numbers

An even stronger result than the weak law of large numbers provides the *strong law of large numbers*, which states that

$$P\left(\lim_{n\to\infty} |Z_n - \mu| = 0\right) = 1.$$

Here, we measure the probability of all outcomes $\omega \in \Omega$ with $\lim_{n\to\infty} |Z_n(\omega) - \mu| = 0$ and find that this event occurs with probability one. The strong law of large numbers implies the weak law of large numbers.
We have seen two laws that match our initial intuition that

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mu.$$

In the following we will see that even more can be derived about sums of random variables.

6.4 Central Limit Theorem.

Recall that $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where the X_i are independent and identically distributed random variables with finite expectation μ and finite variance $\sigma^2 > 0$. According to the central limit theorem, the distribution of

$$Z_n^* = \frac{Z_n - E(Z_n)}{\sqrt{VAR(Z_n)}} = \frac{Z_n - \mu}{\sigma/\sqrt{n}}.$$

converges to the normal distribution with expectation 0 and variance 1 (standard normal distribution). Thus, for $x, y \in \mathbb{R}$, $x < y$,

$$\lim_{n\to\infty} P\left(x < Z_n^* < y\right) = \frac{1}{\sqrt{2\pi}}\int_x^y e^{-0.5t^2} dt$$

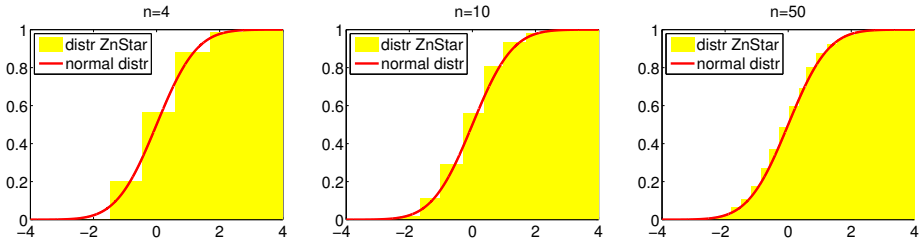


Figure 6.1: The distribution of Z_n^* approaches the standard normal distribution as n increases.

and Z_n is approximately normally distributed with expectation μ and variance σ^2/n, since $Z_n = \sigma/\sqrt{n} \cdot Z_n^* + \mu$ and $Z_n^* \sim \mathcal{N}(0,1)$.

Example 56: Approximation of the binomial distribution

We have already seen that for a sequence of n independent Bernoulli trials the number S_n of successes follows a binomial distribution. We can write S_n as

$$S_n = X_1 + \ldots + X_n$$

where $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter p. If n is large, we can approximate the binomial distribution of S_n by a normal distribution. The central limit theorem tells us that $Z_n = \frac{1}{n} S_n$ is approximately normally distributed with mean p (of the Bernoulli random variables) and variance $\frac{p(1-p)}{n}$. Therefore S_n must approximately follow the distribution $N(np, np(1-p))$. Fig. 6.1 shows a plot of the cumulative probability distribution of Z_n^*. It can be seen that the distribution (in yellow) approaches $\mathcal{N}(0,1)$ (red curve) as n increases.

Chapter 7

Parameter Estimation

We consider an unknown distribution of which we have a collection of samples and we know the family of the distribution (e.g. Poisson) but some of the parameters are unknown and have to be estimated. This problem is an instance of statistical inference, which differs from the methods of descriptive statistics (see Section 1) in that we do not compute numbers to describe our data but we infer (i.e., estimate) the parameters of the process (here this is our theoretical distribution) that is generating our data.
Values that we estimate based on the data that we have are called statistical estimators. In this chapter, we will consider estimators for the mean of the distribution and for its standard deviation. They are random variables computed from a collection of data and are subject to a sampling error. We will also construct random intervals that contain the true value of the parameter with high probability.

Chapter learning objectives

- understand the concept of an estimator and important properties of estimators
- understand and apply popular methods for estimating parameters: method of moments, maximum likelihood estimation, Bayesian inference
- determine standard deviations of the corresponding estimators
- know the advantages and disadvantages of the different estimation methods

Example 57: Photon Count

Consider a telescope observing the photon flux of a certain star during certain time intervals of fixed length. We assume that the distribution of

the number of observed photons in the interval $[t, t+h]$ is constant in t, i.e. we have the same distribution for all intervals of length h. Assume further that n independent measurements are performed for certain intervals of length h, i.e. we have n numbers $x_1, \ldots, x_n$ for the photon flux within intervals of length h. Clearly, we will in reality never have exact and truly independent measurements but let us assume for now that

a) the measurements have been taken over n days for, say, $h = 1$ hour every day,

b) the telescope gives a very accurate measurement.

From a) we get approximate independence and from b) we derive that measurement errors may be neglected. Random photon counts can well be described by a Poisson distribution with some *parameter λ.*
We know that the mean of the Poisson distribution is $E(X) = \lambda$ and the variance is $VAR(X) = \lambda$ as well. Shall we now use the sample mean

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

to fit $E(X) = \lambda$ or shall we use the sample variance

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

as defined in Section 1.2.3 to fit $VAR(X) = \lambda$? Which of the two will give us a better estimate for λ? What does "better" mean in this context? We give answers to these questions in the next sections.

Recall that we use a lower case 'x' for concrete realizations (real values!) of a random variable X. When we discuss independent observations that follow the distribution of X, then we can take two different views:

View 1: We have random samples of X, i.e. independent realizations $x_1, \ldots, x_n \in \mathbb{R}$. If we compte the corresponding mean $\bar{x}$, then $\bar{x}$ is a real value.

View 2: We look at the observation process in a more abstract view and describe the n independent observation samples as *random variables* $X_1, \ldots, X_n$ with the following properties:

- they are independent because we ensured independence during the observation process. E.g. assume you consider data for the height of a

set of individuals. If these individuals are relatives, your independence assumption is no longer true as this might bias the data towards a higher or lower height than you have on average in the total population that you consider.

- the follow the same distribution, namely the distribution of the random variable X which represents the random variable that describes a single data point (e.g. height of a person). Hence, the all have, for instance, the same expectation:

$$E[X] = E[X_1] = \ldots = E[X_n].$$

- any combination or transformation of the X_i is again a random variable. E.g. their mean

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

is a *random variable*! This is intuitive, if you imagine that you repeat your observation process several times. Your data might change and the mean of your data as well. In the sequel, the properties (such as its mean and variance) of such an *estimator* will be an important aspect.

7.1 Method of Moments

Given data points $x_1, \ldots, x_n$ of n independent measurements (view 1!), the idea of the method of moments is to adjust the moments of the distribution such that they fit the sample moments of the data. In general, the k-th sample moment is defined as

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k.$$

Note that we often write $\bar{x}$ for m_1. For $k > 1$ it is often a good alternative to consider the k-th sample central moments

$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

since non-central moments may become very large. Note that for the sample variance two definitions exist, the one for $\bar{m}_2$ and the one for s^2. The difference is that for $\bar{m}_2$ we divide the sum by n while for s^2 we divide it by $n-1$. The reason is that if we view the sample variance as an estimator of the true variance of a random variable X based on the data points $X_1, \ldots, X_n$

(view 2!) where the X_i are i.i.d. (same distribution as X with mean μ and variance σ^2), then the estimator

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \text{ where } \bar{X} = \frac{1}{n}\sum_i X_i,$$

is unbiased ($E(S^2) = \sigma^2$) while

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is not ($E(\tilde{S}^2) \neq \sigma^2$). We will come back to this issue when we discuss the pros and cons of the method of moments.

Now we assume that we consider a theoretical distribution with exactly K unknown parameters $\theta_1, \ldots, \theta_K$. Then the theoretical moments $\mu_1, \mu_2, \ldots$ of the distribution (often called *population moments*) are functions of these parameters, i.e. $E(X^k) = \mu_k(\theta_1, \ldots, \theta_K)$. For example, the exponential distribution has a parameter $\theta_1 = \lambda > 0$ that determines the mean $E[X] = \mu_1(\theta_1) = \lambda^{-1}$.
We may also simply write μ instead of $\mu_1(\theta_1, \ldots, \theta_K)$. Clearly, the central moments $E((X-\mu)^k) = \bar{\mu}_k(\theta_1, \ldots, \theta_K)$, $k = 2, 3, \ldots$ are also a function of $\theta_1, \ldots, \theta_K$.
To determine these parameters we construct an equation system

$$\begin{aligned}
\mu_1(\hat{\theta}_1, \ldots, \hat{\theta}_K) &= m_1 \\
\bar{\mu}_2(\hat{\theta}_1, \ldots, \hat{\theta}_K) &= \bar{m}_2 \\
&\ldots \\
\bar{\mu}_K(\hat{\theta}_1, \ldots, \hat{\theta}_K) &= \bar{m}_K
\end{aligned}$$

and solve for $\hat{\theta}_1, \ldots, \hat{\theta}_K$. Note that it is also possible to equate the non-central moments instead. Also note that in general, it may be necessary to add equations for higher moments if there is no unique solution for $\hat{\theta}_1, \ldots, \hat{\theta}_K$ with K equations (for instance, because some equations are linearly dependent).

Example 58: Fitting the mean of a Poisson distribution

Revisiting Example 57 we have the equation $\mu = m_1$. We know that the parameter λ is equal to the theoretical mean μ. Hence, we choose $\hat{\lambda} = \bar{x} = m_1$ as an estimator for the parameter λ.

Example 59: Fitting the parameters of a normal distribution

Assume that our data points $x_1, \ldots, x_n$ are IID realizations of a random variable X that follows a normal distribution. The normal distribution has parameters θ_1 (equal to the theoretical mean μ) and θ_2 (equal to the standard deviation σ). Also, $\sqrt{\bar{\mu}_2} = \theta_2$ since $\bar{\mu}_2 = \sigma^2$.

We use the two equations $\mu = m_1$ and $\bar{\mu}_2 = \bar{m}_2$. Thus, we directly get as a solution that $\theta_1 = m_1 = \bar{x}$ (from the first equation) and $\theta_2 = \sqrt{\bar{m}_2}$ (from the second equation).
As a variant, assume that it is known that X has mean $\mu = 0$. Then, for the single remaining parameter σ we do not set $K = 1$ and consider only the first equation $\mu_1(\mu, \sigma) = m_1$ since the first moment $\mu_1(\mu, \sigma)$ is equal to the first parameter μ and thus does not give any constraints on σ. Instead we consider the second equation $\bar{\mu}_2(\mu, \sigma) = \bar{m}_2$ since $\bar{\mu}_2(\mu, \sigma) = \sigma^2$ and thus $\sigma = \sqrt{\bar{m}_2}$.

The method of moments has a number of disadvantages compared to the estimation methods considered in the sequel. The estimators may be biased and thus may have a systematic estimation error. An example is the bias of the sample variance

Example 60: Bias of the sample variance

As already discussed above, according to the method of moments, we would estimate the variance of the true distribution by

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

instead of using the unbiased estimator

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

Intuitively, $\tilde{s}^2$ underestimates the variance since in its definition we use an estimator for the mean instead of the true mean and this estimator must lie in the center of the samples (while the true mean might not) and therefore the deviation from it is smaller than the deviations from the true mean. Consider for instance a normal distribution with mean 0 and variance 100. For a small sample size of, say, $n = 3$ one might get the data points $x_1 = -1.9412$, $x_2 = -21.3836$, $x_3 = -8.3959$ using the following Python command `x = np.random.normal(0,100,3)`*. The sample mean $\bar{x} = -10.5736$ is far away from the true mean. Obviously, the deviations*

from the true mean 0 *are much larger than the deviations from* $\bar{x}$ *and thus* $\tilde{s}^2$ *is only* `np.var(x)` $= 65.3717$ *while* `np.var(x,ddof=1)` $= S^2 = 98.0576$ *is closer to the true standard deviation (note that the Python command* `np.var(x)` *uses the biased estimator!).*
If we could consider the deviations from the (usually unknown) true mean 0 *the estimator with the factor* $\frac{1}{n}$ *becomes unbiased:*

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2)\right) &= \frac{1}{n}\sum_{i=1}^{n} E((X_i - E(X))^2) \\ &= \frac{1}{n} n \cdot VAR(X_i) = VAR(X) \end{aligned}$$

The method of moments has the drawback that it does not take into account information about more moments than needed to get a unique solution for the equation system, i.e. if we have a single parameter, we fully rely on only the mean or only the variance to estimate it. For instance, in Example 57 we only use $\bar{x}$ and do not additionally take into account the variation of the data, i.e. s^2 or $\tilde{s}^2$.
The generalized method of moments *does* take into account this information by considering cost functions instead of equating the population moments of the distribution and the sample moments.

7.2 Maximum Likelihood Estimation

Again, we assume that n independent observations $X_1, \ldots, X_n$ are given as well as the family of the distribution. The maximum likelihood method is the most common way of estimating parameters of a distribution. Recall that an estimator is a function of the data. Here, we consider maximum likelihood estimators.
Let θ be the unknown parameter and let $P_\theta(\cdot)$ be the corresponding discrete probability mass function, i.e. the X_i are all independent and identically distributed with $P(X_i = x) = P_\theta(x)$, $x \in \mathbb{R}$. Therefore, the probability of observing $X_1 = x_1, \ldots, X_n = x_n$, assuming that θ is the true value, is given by

$$L_\theta(x_1, \ldots, x_n) = P_\theta(x_1)P_\theta(x_2)...P_\theta(x_n).$$

$L_\theta(x_1, \ldots, x_n)$ is also called the likelihood of the data and we say that $\hat{\theta}$ is a maximum likelihood estimator (MLE) of θ if for all possible θ

$$L_\theta(x_1, \ldots, x_n) \leq L_{\hat{\theta}}(x_1, \ldots, x_n).$$

Note that the MLE of θ is, in general, not unique. Intuitively, the MLE is the parameter value that "best explains" the observed data.

To maximize the likelihood, we can consider its derivatives w.r.t. θ and find parameter values where $\frac{\partial}{\partial\theta}L_\theta(x_1,\ldots,x_n)$ equals 0 (in some cases such points do not exist and we find $\hat{\theta}$ at the boundary of the set of possible values for θ).
Often the maximum of the log-likelihood

$$\ln L_\theta(x_1,\ldots,x_n) = \ln P_\theta(x_1) + \ln P_\theta(x_2) + ... + \ln P_\theta(x_n),$$

which is equal to that of the likelihood (as the logarithm is strictly monotonically increasing), is easier to compute. More concretely, $\hat{\theta}$ maximizes L_θ if and only if $\hat{\theta}$ maximizes $\ln L_\theta$.

Example 61: MLE OF THE GEOMETRIC DISTRIBUTION

Assume we have collected data that represents inter-arrival times of customers (in minutes). Since the inter-arrival times are approximately geometrically distributed (we hypothesize this after looking at certain summary statistics such as the range of the data, the mean, the skewness, the variance or a histogram plot of the data), we wish to fit the data to a geometric distribution with parameter θ, i.e. we want to find a θ that best explains the data. The likelihood of the data $x_1,\ldots,x_n$ is in this case given by

$$\begin{aligned} L_\theta(x_1,\ldots,x_n) &= P_\theta(x_1)P_\theta(x_2)...P_\theta(x_n) \\ &= \theta(1-\theta)^{x_1}\cdot\theta(1-\theta)^{x_2}\cdot\ldots\cdot\theta(1-\theta)^{x_n} \\ &= \theta^n(1-\theta)^{\sum_{i=1}^{n} x_i} \end{aligned}$$

since the probability to observe x_i is $P_\theta(x_i) = \theta\cdot(1-\theta)^{x_i}$. Note, that we use the alternative definition of the geometric distribution here, where we only count the number of unsuccessful trials until the first success. *Since $L_\theta(x_1,\ldots,x_n)$ is a function of θ, we shortly write $L(\theta)$ and omit the dependence on the data. Considering the log-likelihood yields*

$$\ln L(\theta) = n\cdot\ln\theta + \sum_{i-1}^{n} x_i \ln(1-\theta)$$

We maximize $\ln L(\theta)$ by setting the derivative to zero:

$$\frac{d\ln L(\theta)}{d\theta} = \frac{n}{\theta} + \frac{(-1)}{1-\theta}\sum_i x_i = \frac{n}{\theta} - \frac{1}{1-\theta}\sum_i x_i$$

$$0 \stackrel{!}{=} \frac{n}{\theta} - \frac{1}{1-\theta}\sum_i x_i$$

$$\Leftrightarrow \sum_i x_i = \frac{n(1-\theta)}{\theta} \Leftrightarrow \bar{x} = \frac{1}{\theta} - 1 \Leftrightarrow \hat{\theta} = \frac{1}{\bar{x}+1}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. *Next we find that* $\hat{\theta}$ *is really a maximizer:*

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = -\frac{n}{\theta^2} - \frac{1}{(1-\theta)^2}\sum_i x_i < 0$$

Hence, the MLE of θ *is* $\hat{\theta} = \frac{1}{\bar{x}+1}$. *If we use the alternative definition of the geometric distribution here, where we count the number of unsuccessful trials* and *the first success, how would the MLE change? Hint: this means that we transform* X *to* $X + 1$.

Note that in the above example, since the mean of the geometric distribution is $(1-\theta)/\theta$ we matched the mean $\bar{x}$ of the data and the mean of the distribution. Note that this is not always the case for a maximum likelihood estimator, i.e. the estimator of the method of moments is not always the same as the MLE.
Assume now that we have hypothesized a continuous distribution for our data. In that case, the *likelihood function* is

$$L_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n)$$

where f_θ is the density of the distribution that depends on the parameter θ. Again, since $L_\theta(x_1, \ldots, x_n)$ is a function of θ, we shortly write $L(\theta)$ and omit the dependence on the data. An MLE $\hat{\theta}$ of θ is defined as a value that maximizes $L(\theta)$ (over all permissible values of θ).
The reason why the likelihood is a density in the continuous case is that the probability of a single outcome is zero in the continuous case. Hence one could consider

$$P(x_i - h < X < x_i + h) = \int_{x_i-h}^{x_i+h} f_\theta(x)dx$$

for the outcome x_i and some very small h. But this quantity is approximately equal to $2hf_\theta(x_i)$ and thus proportional to the factors of the likelihood defined above. Hence, this approach would result in the same MLE.

Example 62: MLE OF THE EXPONENTIAL DISTRIBUTION

Assume that data were collected on the inter-arrival times for cars in a drive-up banking facility. Since histograms of the data show exponential curve, we hypothesize an exponential distribution. Note that some distributions generalize the exponential distribution but have more parameters

(e.g. gamma distribution). They would provide a fit which is at least as good as *the fit of the exponential distribution.*
For the exponential distribution the parameter is $\theta = \lambda, \lambda > 0$ *and the density is* $f_\lambda(x) = \lambda e^{-\lambda x}$*. Hence, the likelihood of the samples* $x_1, \ldots, x_n$ *is*

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \cdot e^{-\lambda \sum_i x_i}$$

and the log-likelihood is

$$\ln L(\lambda) = n \cdot \ln \lambda + (-\lambda \sum_i x_i)$$

and its derivative w.r.t. λ *is*

$$\frac{d}{d\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_i x_i$$

Setting the derivative to zero yields

$$0 \stackrel{!}{=} \frac{n}{\lambda} - \sum_i x_i \;\Leftrightarrow\; \frac{1}{\lambda} = \frac{1}{n} \sum_i x_i = \bar{x}$$

and thus $\lambda = 1/\bar{X}$*. We find that* λ *is a maximizer:*

$$\frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{n}{\lambda^2} < 0$$

and finally get $\hat{\lambda} = 1/\bar{x}$*.*

We remark that for some distributions, the log-likelihood function may not be useful and also, finding a maximum by setting the derivative to zero is not always possible. In general, if an analytic solution is not possible, global optimization methods have to be applied in order to determine, which of several local optima of the likelihood has the highest value.
We list some important properties of Maximum Likelihood Estimators (some of the properties require mild "regularity" assumptions, e.g. likelihood function must be differentiable, support of distribution does not depend on θ.)

- Uniqueness: For most common distributions, the MLE is unique; that is, $L(\hat{\theta})$ is strictly greater than $L(\theta)$ for any other value of θ.

- Asymptotic unbiasedness: $\hat{\theta}$ is a function of the samples $x_1, ..., x_n$ and thus a random variable if $X_i = x_i$ is not given, $i \in \{1, 2, \ldots, n\}$. Assume now that $X_1, ..., X_n$ have distribution P_θ and $\hat{\theta}_n$ is the MLE

of θ based on the observations $X_1, ..., X_n$. Then

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta.$$

Therefore, we say that $\hat{\theta}_n$ is *asymptotically unbiased*.

- Invariance: If $\hat{\theta}$ is an MLE of θ and if g is a one-to-one function[1], then $g(\hat{\theta})$ is an MLE of $g(\theta)$. To see that this holds, we note that

$$L(\theta) = \overbrace{L(g^{-1}}^{\tilde{L}}(g(\theta))$$

are both maximized by $\hat{\theta}$, so $\widehat{g(\theta)} = g(\hat{\theta})$ (where $\widehat{g(\theta)}$ is the MLE of $g(\theta)$ interpreted as the unknown parameter).

Example 63: INVARIANCE

Assume that $x_1, x_2, ..., x_n$ are samples from a Bernoulli distribution with parameter $\theta = p$. Then the likelihood is

$$\prod_{i=1}^{n} [\theta x_i + (1-\theta)\chi_{=0}(x_i)] = \theta^{\sum_i x_i} \cdot (1-\theta)^{n-\sum_i x_i}$$

where $\chi_{=0}(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$

Hence the log-likelihood is

$$\ln L(\theta) = \sum_i x_i \ln(\theta) + (n - \sum_i x_i)\ln(1-\theta)$$

and setting its derivative to zero yields

$$\frac{\partial}{\partial\theta} L(\theta) = \sum_i x_i/\theta - (n - \sum_i x_i)/(1-\theta) \stackrel{!}{=} 0$$

$$\Rightarrow \ (1-\hat{\theta})\sum_i x_i = \hat{\theta}(n - \sum_i x_i) \Leftrightarrow \sum_i x_i = n \cdot \hat{\theta} \Leftrightarrow \hat{\theta} = \sum_i x_i/n$$

Next we find that $\hat{\theta} = \sum_i x_i/n =: \bar{x}$ is a maximizer:

$$\frac{\partial^2}{\partial\theta^2} L(\theta) = -\sum_i x_i/\theta^2 - (n - \sum_i x_i)/(1-\theta)^2 < 0 \text{ for any } \theta \in [0,1]$$

[1] A function is one-to-one if every element of the range of the function corresponds to exactly one element of the domain.

Thus, $\hat{\theta} = \bar{x}$.
Next, we assume that the variance $g(\theta) = \theta(1-\theta)$ *is our parameter (note that in the permissible range of* θ, g *is one-to-one). Define* $\tilde{L}$ *such that* $\tilde{L}(g(\theta)) = L(\theta)$, *i.e., we consider* the same values *for the likelihood but the likelihood function is different as it is a function in* $g(\theta)$. *Setting the derivative to zero yields*

$$\frac{\partial}{\partial g}\tilde{L}(g) = \frac{\partial}{\partial \theta}\underbrace{\tilde{L}(g)}_{=L(\theta)} \cdot \frac{\partial \theta}{\partial g} \stackrel{!}{=} 0.$$

Since $\frac{\partial}{\partial \theta} L(g(\theta))\,|_{\theta=\hat{\theta}} = 0$ *it holds that* $\frac{\partial}{\partial g}\tilde{L}(g) = 0$ *if we choose* $\hat{g} = g(\hat{\theta}) = \bar{x}(1-\bar{x})$.

- Asymptotically normally distributed: For $n \to \infty$, $\hat{\theta}_n$ converges in distribution to $N(\theta, \delta(\theta))$ (normal distribution with mean θ and variance $\delta(\theta)$), where $\hat{\theta}_n$ is our estimation based on n observations, θ is "true" parameter of the distribution of the observations. Moreover,

$$\delta(\theta) = -\left(E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right]\right)^{-1}$$

  One can even show that for any other estimator $\tilde{\theta}$, which converges in distribution to $N(\theta, \sigma^2)$, we have $\delta(\theta) \leq \sigma^2$ (variance is greater or equal). Therefore, MLEs are called *best asymptotically normal*. For large n, we can estimate the probability that $\hat{\theta}$ deviates from the true value by more than ϵ (see confidence intervals; later).

- strongly consistent: MLEs are strongly consistent, i.e,

$$P(\lim_{n\to\infty} \hat{\theta}_n = \theta) = 1.$$

In the case of more than one parameter the MLE is found in a very similar way, i.e. we find the vector $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ that maximizes $L(\theta)$ or $\ln L(\theta)$. For instance, for two parameters α and β we try to solve $\frac{\partial}{\partial \alpha} \ln L(\theta) = 0$ and $\frac{\partial}{\partial \beta} \ln L(\theta) = 0$ simultaneously for α and β where $\theta = (\alpha, \beta)$. (See exercises for an example.)

7.2.1 Variance of MLE (optional content)

Before we consider the variance of the maximum likelihood estimator, we will discuss the variance of estimators in general.

It is very important to analyze how good the estimated value given by some estimator $\hat{\theta}$ is. If $VAR(\hat{\theta})$ is large, then our estimation is of bad quality and might lead to wrong conclusions about the real system. Typically, when results of parameter estimations are reported, the (estimated) standard deviations $\sqrt{VAR(\theta)}$ are given as well.

Example 64: Poisson distribution

In the previous sections we found that both the method of moments and the maximum likelihood approach gives the estimator

$$\hat{\theta} = \bar{X}$$

for the Poisson distribution, i.e. the best value for the unknown mean of the Poisson distribution is the sample mean. In Section **??** *we already found that the variance of $\bar{X}$ is σ^2/n where σ^2 is the variance of the distribution of the X_i and thus for the Poisson distribution we get*

$$VAR(\hat{\theta}) = VAR(\bar{X}) = \theta^*/n.$$

Here, θ^* is the 'true' (unknown) value that we estimate with $\hat{\theta} = \bar{X}$. Thus, we can estimate $VAR(\hat{\theta})$ as $\bar{X}/n$.
Alternatively, we can estimate σ^2 by computing the sample variance S^2 (as explained in Example 60) and estimate $VAR(\hat{\theta})$ as S^2/n.

Often, analytic formulas for $VAR(\hat{\theta})$ cannot be derived. In this case, one can either use the bootstrap method or estimate $VAR(\hat{\theta})$ based on the properties of the estimator.

Maximum Likelihood Method. To estimate the variance of a maximum likelihood estimator, we exploit the fact that MLEs are asymptotically normally distributed. Thus, for large sample sizes and a single parameter $\theta \in \mathbb{R}$, we have approximately a variance of

$$\delta(\theta^*) = -\left(E\left[\frac{d^2 \ln L(\theta^*, X_1, \ldots, X_n)}{d\theta^2} \right] \right)^{-1}.$$

In the above expression we wrote $L(\theta, X_1, \ldots, X_n)$ for the likelihood to emphasize that L is a function of θ and $X_1, \ldots, X_n$. Moreover, here θ^* is the true value of θ. For an estimator based on the data we simply replace $X_1, \ldots, X_n$ by their corresponding realizations $x_1, \ldots, x_n$ and the (unknown) true value of θ by the estimate $\hat{\theta}$, i.e.

$$-\left(\frac{\partial^2 \ln L(\hat{\theta}, x_1, \ldots, x_n)}{\partial \theta^2} \right)^{-1}.$$

Intuitively, the second derivative tells us the curvature of the likelihood and if it is 'flat' at $\hat{\theta}$ then our estimated value might not be very accurate and the variance (negative inverse) is large. Then either we do not have enough samples or the parameter is difficult to identify (see also the example after next below). We are not very confident about our estimated value since perturbing $\hat{\theta}$ slightly also gives us a similar likelihood.
For several parameters, the same approach is used, i.e. the negative diagonal entries of the inverse of the Hessian matrix give estimates for the variances of the parameters.

Example 65: Parameters of the normal distribution

Assume that we have n i.i.d. samples $X_1, \ldots, X_n$ that follow a normal distribution with (unkown) mean μ and (unkown) variance σ^2. Hence, $\theta = (\mu, \sigma^2)$.
The likelihood of the data is

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right)$$

. Thus, log-likelihood of the data is

$$\ln L(X_1, \ldots, X_n; \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Next we compute the derivatives w.r.t. μ and σ^2:

$$\begin{aligned} \tfrac{\partial}{\partial\mu} \ln L(X_1, \ldots, X_n; \theta) &= \tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^{n}(X_i - \mu) \\ \tfrac{\partial}{\partial\sigma^2} \ln L(X_1, \ldots, X_n; \theta) &= -\tfrac{n}{2}(\sigma^2)^{-1} + \tfrac{1}{2}(\sigma^2)^{-2}\textstyle\sum_{i=1}^{n}(X_i - \mu)^2. \end{aligned}$$

Setting the two derivatives to zero yields from the first equation

$$\sum_{i=1}^{n}(X_i - \mu) = 0 \implies \sum_{i=1}^{n} X_i - n\mu = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Inserting this into the second equation and multiplying by $2\sigma^4$ gives

$$-n\sigma^2 + \sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Thus, the MLE of σ is a biased estimator.
To prove that $\hat{\mu}$ and $\hat{\sigma}^2$ yield a maximum we have to consider the Hessian matrix.

$$H(\mu, \sigma) = \begin{bmatrix} \frac{\partial^2}{\partial\mu\partial\mu} \ln L & \frac{\partial^2}{\partial\mu\partial\sigma^2} \ln L \\ \frac{\partial^2}{\partial\sigma^2\partial\mu} \ln L & \frac{\partial^2}{\partial\sigma^2\partial\sigma^2} \ln L \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{n}{\sigma^2} & -(\sigma^2)^{-2}\sum_{i=1}^{n}(X_i-\mu) \\ -(\sigma^2)^{-2}\sum_{i=1}^{n}(X_i-\mu) & \frac{n}{2}(\sigma^2)^{-2} - (\sigma^2)^{-3}\sum_{i=1}^{n}(X_i-\mu)^2 \end{bmatrix}$$

We have a local maximum at $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ *if* $\frac{\partial^2}{\partial\mu\partial\mu}\ln L < 0$ *and if the determinant*

$$\det(H(\hat{\mu},\hat{\sigma}^2)) = \frac{\partial^2}{\partial\mu\partial\mu}\ln L \cdot \frac{\partial^2}{\partial\sigma^2\partial\sigma^2}\ln L - \frac{\partial^2}{\partial\mu\partial\sigma^2}\ln L \cdot \frac{\partial^2}{\partial\sigma^2\partial\mu}\ln L$$

is positive at $\hat{\theta}$*. Next we insert concrete numbers to simplify the computation. We generate 50 random numbers that are normally distributed with mean zero and variance one. We get the estimates*

$$\hat{\mu} = -0.1229 \text{ and } \hat{\sigma}^2 = 0.9903$$

and the Hessian is (approximately)

$$\begin{bmatrix} -50.4906 & -0.0000 \\ -0.0000 & -25.4930 \end{bmatrix}.$$

(Note that its determinant is indeed positive.) The inverse of the Hessian is

$$\begin{bmatrix} -0.0198 & 0.0000 \\ -0.0000 & -0.0392 \end{bmatrix}.$$

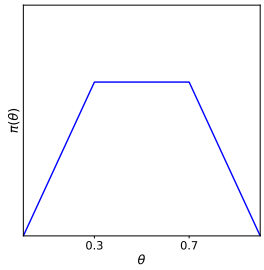
Thus, the estimated variances are 0.0198 *for* $\hat{\mu}$ *and* 0.0392 *for* $\hat{\sigma}^2$*, yielding standard errors of* $\epsilon_1 = 0.1407$ *and* $\epsilon_2 = 0.1981$*. Note that the true mean and variance are elements of the intervals* $[\hat{\mu}-\epsilon_1, \hat{\mu}+\epsilon_1]$ *and* $[\hat{\sigma}^2-\epsilon_2, \hat{\sigma}^2+\epsilon_2]$.

If the standard errors are extremely high, this shows that the log- likelihood function is 'very flat' around the minimum and thus the estimated values may not be close to the true values of the parameters. We then have the problem of parameters that are not identifiable. However, sometimes at least ratios of parameters are estimated very accurately.

7.3 Bayesian Inference

In the previous sections, we followed the so called frequentist approach when estimating population parameters. We worked with distributions, expectations, and variances of the given random samples (population data) as well as of estimators and other statistics. The likelihood $L_\theta(x_1, \ldots, x_n)$ of the data given the parameter θ is a central function in the frequentist approach.

In this section, we will follow the Bayesian approach, which differs from the frequentist approach in that θ is treated as a random variable[2] and has a certain probability distribution. Hence, not only the data is a source of uncertainty but also θ. Its distribution $\pi(\theta)$, called *prior distribution*, reflects our ideas, beliefs, and past experiences about θ before we make use of the data $x_1, \ldots, x_n$, i.e., before we perform inference based on the data.



For example, our prior belief could be that it is likely that the true value of $\theta \in [0, 1]$ lies in a certain interval $[a, b] \subset [0, 1]$. However, we cannot completely exclude the cases $\theta < a$ and $\theta > b$. Hence, we could assume a prior distribution $\pi(\theta)$ as depicted in the figure on the left for $[a, b] = [0.3, 0.7]$.

The information given by the population data may lead to a change of our prior beliefs, i.e., after the Bayesian inference we obtain a *posterior distribution* which tells us how likely different values of θ are when we know the data. The main advantage of this approach is that it gives also meaningful results in the case of only few samples since in this setting we do not rely on asymptotic estimator properties (which require a large number of samples). On the other hand, the posterior distribution does not give a fixed value or confidence interval for θ but only a distribution. In the above example, we could, for instance, compute the posterior probability of the interval $[0.3, 0.7]$ and compare it to the probability of the prior to see whether the data lead to an increase of the probability that $\theta \in [0.3, 0.7]$.
Hence, to perform Bayesian inference, we need besides the observed data a prior distribution $\pi(\theta)$ for θ. Assume that the data has likelihood $L_\theta(x_1, \ldots, x_n) = L(x_1, \ldots, x_n \mid \theta)$ (it is helpful to make the condition on θ explicit at this point by writing '$\mid \theta$'). Note that this condition means that our random parameter – let us call it Θ – takes the concrete value θ. Since $\pi(\theta)$ is the probability that $\Theta = \theta$, we get the posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$ as

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{L(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{f(x_1, \ldots, x_n)} \tag{7.1}$$

according to Bayes' Theorem. The denominator $f(x_1, \ldots, x_n)$ is the marginal probability of the data, i.e., independent of a concrete choice of $\theta = \Theta$. It can be computed using the law of total probability.

$$f(x_1, \ldots, x_n) = \sum_{\theta} L(x_1, \ldots, x_n \mid \theta)\pi(\theta) \tag{7.2}$$

[2]In the previous sections we used upper case letters to denote random variables. In this section, we use θ although it is a random variable as this is the standard notation in the literature.

If the prior is a continuous distribution, we compute the marginal as

$$f(x_1, \ldots, x_n) = \int_\theta L(x_1, \ldots, x_n \mid \theta)\pi(\theta)\ d\theta. \tag{7.3}$$

Example 66: Quality inspection[3]

A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on θ, the proportion of defective parts. Before we see the real data, let?s assign a 50-50 chance to both suggested values of θ, i.e.,

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones.
We calculate the posterior distribution of θ as follows: First, we find that the probability of X defective parts in a sample of n parts is binomially distributed with parameters n and θ. Hence,

$$L(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and therefore

$$L(3 \mid 0.05) \approx 0.0596 \text{ and } L(3 \mid 0.10) \approx 0.1901.$$

For the marginal distribution we compute

$$f(x) \quad = \quad 0.5 \cdot L(x \mid \Theta = 0.05) \ + \ 0.5 \cdot L(x \mid \Theta = 0.10) \approx 0.12485.$$

Finally, the posterior probabilities are

$$\begin{aligned} \pi(0.05 \mid X = 3) &= \tfrac{L(X=3|0.05)\pi(0.05)}{f(X=3)} \approx 0.2387, \\ \pi(0.10 \mid X = 3) &= \tfrac{L(X=3|0.10)\pi(0.10)}{f(X=3)} \approx 0.7613, \end{aligned}$$

which indicate a threefold higher chance that the proportion of defective parts is 10% compared to the 5% claimed by the manufacturer.

7.3.1 Conjugate families of distributions

It is often an important advantage of Bayesian approaches if we can determine the posterior distribution analytically, since otherwise we may not be

[3] Taken from [1].

able to efficiently compute the posterior distribution. We say that a family of prior distributions is conjugate to a model if the posterior distribution belongs to the same family as the prior.

Example 67: Gamma prior is conjugate to a Poisson model

We consider realizations $x_1, \ldots, x_n$ of a Poisson distribution with parameter θ as our model. Hence,

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}.$$

Next, we assume for θ a Gamma(α, λ) distribution as a prior, i.e.,

$$\pi(\theta) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta} \sim \theta^{\alpha-1}e^{-\lambda\theta},$$

where $\sim$ means "proportional to". In the sequel, we will determine the posterior only up to a constant coefficient to simplify the derivation. Therefore, we will also drop the factor $\prod_{i=1}^{n} \frac{1}{x_i!}$ of the density f of the data, which is constant in θ, i.e., for $\vec{x} = (x_1, \ldots, x_n)$

$$f(\vec{x} \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \sim \prod_{i=1}^{n} e^{-\theta}\theta^{x_i} = e^{-n\theta}\theta^{\sum x_i}.$$

Now, the posterior is given by

$$\begin{aligned} \pi(\theta \mid \vec{x}) &\sim f(\vec{x} \mid \theta)\pi(\theta) \\ &\sim \left(e^{-n\theta}\theta^{\sum x_i}\right)\left(\theta^{\alpha-1}e^{-\lambda\theta}\right) \\ &= \theta^{\alpha+\sum x_i-1}e^{-(\lambda+n)\theta}. \end{aligned}$$

Thus, the posterior is a Gamma$(\alpha + \sum x_i,\ \lambda + n)$ distribution. Note that the missing constant factor of the distribution can be uniquely determined as $\pi(\theta \mid \vec{x})$ is a proper density.
It is interesting to observe how the mean and the variance of the Gamma distribution is adjusted, when we take the data into account: For the prior we have $E[\Theta] = \frac{\alpha}{\lambda}$ and $VAR[\Theta] = \frac{\alpha}{\lambda^2}$, while the posterior has $E[\Theta \mid \vec{x}] = \frac{\alpha+\sum x_i}{\lambda+n}$ and $VAR[\Theta \mid \vec{x}] = \frac{\alpha+\sum x_i}{(\lambda+n)^2}$.

A list of further conjugate prior distributions can be found on Wikipedia.

7.3.2 Bayesian point-estimators

From the posterior distribution different estimators can be determined. The most common estimator is the *posterior mean*, which gives the average value of θ conditioned on the data, i.e.,

$$\hat{\theta}_M = E[\Theta \mid \vec{x}] = \sum_{\theta} \theta \, \pi(\theta \mid \vec{x})$$

if the posterior/prior is a discrete distribution. In the continuous case, the sum is replaced by an integral over θ.
For this estimator, the variance based on the posterior distribution is computed as

$$VAR[\Theta \mid \vec{x}] = E[(\hat{\theta}_M - \Theta)^2 \mid \vec{x}].$$

Other popular point estimators are the *posterior median* and the *maximum a posteriori (MAP)*,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi(\theta \mid \vec{x})$$

which is given by the maximum of the posterior distribution. Note that the MAP estimator is not a good choice if the maximum of the posterior distribution is not representative for the distribution, e.g. if the posterior distribution is bimodal and $\hat{\theta}_{MAP}$ is determined by a narrow peak, while other very different values of θ are also quite likely.

Example 68: Quality inspection (continued)

Recall the posterior distribution of Example 66,

$$\begin{aligned} \pi(0.05 \mid x) &\approx 0.2387, \\ \pi(0.10 \mid x) &\approx 0.7613. \end{aligned}$$

The posterior mean is

$$\hat{\theta}_M = E[\Theta \mid x] = \sum_{\theta} \theta \, \pi(\theta \mid x) \approx 0.05 \cdot 0.2387 + 0.1 \cdot 0.7613 \approx 0.0880,$$

which is much closer to the inspector's estimate. The corresponding variance is

$$\begin{aligned} VAR[\theta \mid x] &= E[(\hat{\theta}_M - \Theta)^2 \mid \vec{x}] \\ &\approx (0.088 - 0.05)^2 \cdot 0.2387 + (0.088 - 0.1)^2 \cdot 0.7613 \\ &\approx 0.0004, \end{aligned}$$

which means that the standard deviation is about 0.02. Obviously, the MAP estimator would be $\theta_{MAP} = 0.1$ as the posterior probability at 0.1 is higher than that at 0.05.

Chapter 8

Statistical Testing

Carrying out statistical test that are related to the observations of the real system and/or the model is very important when we make claims or statements about the system. A popular class of tests are hypothesis tests that are used to verify statistical hypotheses.
In a statistical hypothesis test, we usually first formulate a *(null) hypothesis* H_0 and an *alternative hypothesis* H_A. The two statements H_0 and H_A must be mutually exclusive and the test can either accept or reject H_0 (in favor of H_A). The null hypothesis is either

- an equality
- the absence of an effect or some relation

Note that this leads to null hypotheses that are in many cases not the same as the statement that we want to verify. The reason for the above constraint is that intuitively, we need an equality (or absence of an effect or some relation) to fix the distribution that we consider during the test.

Chapter learning objectives

- understand the concept of statistical tests
- apply Z-tests and T-tests to observed data and know the corresponding prerequisites
- derive correct interpretations from statistical test
- understand the concept of p-values

We begin with a motivating example.

Example 69: Defective Products

Assume that a manufacturer claims that at most 3% of his products are defective. We want to verify this statement to decide whether we accept the shipment of the products or not. We define

$$H_0\text{: fraction of defective products is equal to 3\%}$$
$$H_A\text{: fraction of defective products is greater than 3\%}$$

where for H_A we used the right-tail alternative. If we have evidence that H_0 is rejected in favor of H_A then we reject the shipment.
Note that H_A: 'fraction of defective products is smaller than 3%' is not useful since then we will always accept the shipment - no matter if we have evidence for H_0 or for H_A.

Example 70: Concurrent Users

Assume that we want to verify the statement that the average number of concurrent users of an online PC gaming platform increased by 2000 this year. We define

$$H_0\text{: } \mu_2\text{-}\mu_1 = 2000$$
$$H_A\text{: } \mu_2\text{-}\mu_1 \neq 2000$$

where μ_1 is the average number of concurrent users of the last year and μ_2 the average number of this year. Here, H_A is called a two-sided alternative since it covers both cases $\mu_2\text{-}\mu_1 > 2000$ and $\mu_2\text{-}\mu_1 < 2000$.

From the two examples above we see that there can be two-sided alternatives, one-sided, left-tail alternatives (H_A is $\mu < \mu_0$) and one-sided, right-tail alternatives (H_A is $\mu > \mu_0$), where H_0 is $\mu = \mu_0$.

	Result of the test	
	Reject H_0	**Accept** H_0
H_0 **is true**	*Type I error*	correct
H_0 **is false**	correct	*Type II error*

The outcome of our test depends on a finite random sample and thus we may always take a wrong decision. The four situations depicted on the left are possible. Our goal is to keep each of the two errors small. Thus, a good test only results in a wrong decision if the sample is not very representative (i.e. extreme). Often the type I error is seen as more dangerous since it corresponds to 'convicting an innocent defendant' or 'sending a healthy patient to a surgery'. Therefore, we fix the probability α of a type I error which is also called the *significance level* of

the test.

$$\alpha = P\{\text{reject } H_0 \mid H_0 \text{ is true}\}$$

The probability of rejecting a false hypothesis (avoid a type II error) is the *power of the test* and a function of the parameter θ about which we make our hypothesis:

$$p(\theta) = P\{\text{reject } H_0 \mid \theta; H_A \text{ is true}\}$$

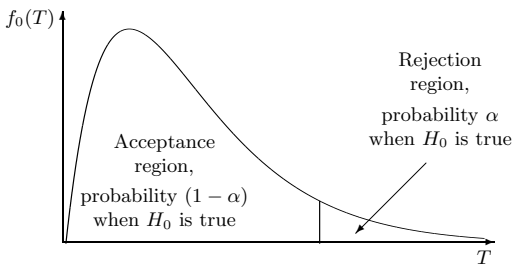
Typically, α is chosen very small, e.g. $\alpha \in \{0.01, 0.05, 0.10\}$ such that the type I error is kept small and we only reject H_0 with a lot of confidence.

8.1 Level α tests: general approach

To test H_0 against H_A we perform the following steps:

1. Compute a *test statistic* T which is a function of the sample (or an estimator) and thus a random number. The distribution of T, given H_0 is true, is known.

2. Consider the *null distribution* F_0 of T, given H_0 is true, and find the portion that corresponds to α, i.e. the part of the area below the density curve that gives α and is thus the region where we reject H_0. The remaining part (of area 1-α) is the *acceptance region*.



In the illustration on the left, T is expected to be large if H_A is true. Therefore, the rejection region is at the right tail of the distribution. In a two-sided test, we consider portions of $\alpha/2$ at both tails of the distribution.
We always have that

$$P\{T \in \text{ acceptance region} \mid H_0 \text{ is true}\} = 1 - \alpha$$
and
$$P\{T \in \text{ rejection region} \mid H_0 \text{ is true}\} = \alpha.$$

3. Accept H_0 if T belongs to the acceptance region and reject it otherwise. It is important to mention that if we accept H_0 we cannot say that 'with probability $1-\alpha$ the hypothesis H_0 is true'. The reason is that H_0 is not random, i.e. it either holds with probability one or it does not hold with probability one. The only correct interpretation is that if we reject H_0 then the data provides sufficient evidence against H_0 and in favor of H_A. Either H_0 is really not true or our data is not representative - which happens with probability α. If we accept H_0

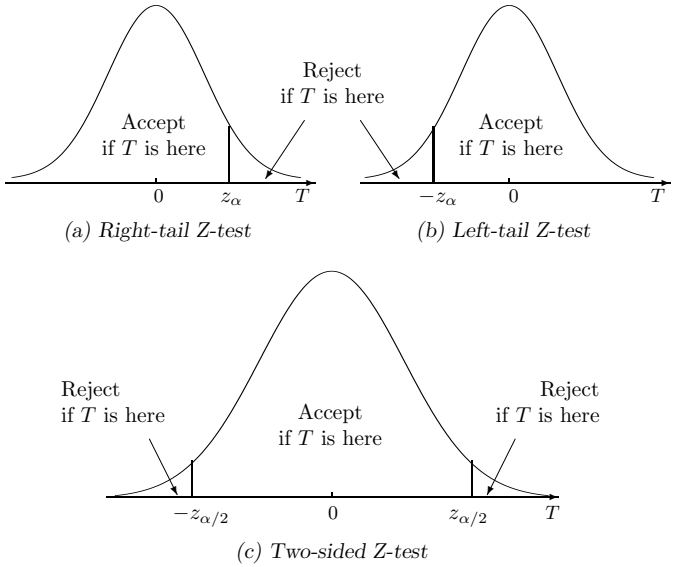


Figure 8.1: Acceptance and rejection regions for a normally distributed test statistic. a) one-sided right-tail alternative; b) one-sided left-tail alternative; c) two-sided alternative.

then the data does not provide sufficient evidence to reject H_0. In the absence of sufficient evidence, by default, we accept H_0.

Let us now in detail reason about finding the rejection region with area α since there are many regions with area α below the density curve. The best choice is an area that ensures that the type II error is small. Thus, we choose the rejection region such that it is likely that T falls into this region if H_A is true. This will maximize the power of the test, i.e. the probability of rejecting H_0 given H_A is true. Often, this results in the a choice of the rejection region as illustrated in Figure 8.1 for a normally distributed test statistic. Usually, the test statistic is defined such that

- the right-tail alternative forces T to be large,
- the left-tail alternative forces T to be small,
- the two-sided alternative forces T to be either large or small.

8.2 Standard Normal Null Distribution (Z-test)

For a large number of applications, the null distribution of T (the distribution of T given H_0 is true) is standard normal. Then the test is called a Z-test. Usually, one the following cases applies:

- we consider sample means of normally distributed data,
- we consider sample means of arbitrarily distributed data where the number of samples is large,
- we consider sample proportions of arbitrarily distributed data where the number of samples is large,
- we consider differences of sample means or sample proportions where the number of samples is large.

In all of these cases a Z-test can be used.
Let z_α be the number such that $P(Z > z_\alpha) = \Phi(z_\alpha) = \alpha$ if Z is a standard normally distributed random variable. Then we reject H_0 if

- $Z \geq z_\alpha$ for a test with right-tail alternative,
- $Z \leq -z_\alpha$ for a test with left-tail alternative,
- $|Z| \geq z_{\alpha/2}$ for a test with two-sided alternative.

Note that each time we have

$$P\{\text{reject } H_0 \mid H_0 \text{ is true}\} = 1 - \Phi(z_\alpha) = \alpha.$$

Example 71: A TEST FOR THE MEAN

Assume that we have used $\bar{X}$ to estimate the unknown mean μ_0 of a normal distribution based on $n = 100$ independent samples. Since the samples are normally distributed, $\bar{X}$ is normally distributed too and we know that $E[\bar{X}] = \mu_0$ (see previous chapter) and that $VAR[\bar{X}] = \sigma^2/n$. Assume further that $\sigma = 800$ is known and that the data is such that $\bar{X} = 5200$. We would like to verify (with a significance of 5%, $\alpha = 0.05$) whether the true mean is greater than 5000, i.e. H_0: $\mu_0 = 5000$ and H_A: $\mu_0 > 5000$ (right-tail alternative).

1. *We first compute the test statistic*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{5200 - 5000}{800/\sqrt{100}} = 2.5.$$

2. *The critical value is $z_\alpha = 1.645$ (from the table of the standard normal distribution). Thus, we should reject H_0 if $Z \geq 1.645$ and accept it otherwise.*

3. *Since Z falls into the rejection region, we have enough evidence to reject H_0 and support the alternative hypothesis H_A: $\mu_0 > 5000$.*

Note that if, for instance, $\bar{X} = 5100$, we would get $Z = 1.25$ and not have enough evidence to reject H_0 and believe in the alternative hypothesis that $\mu_0 > 5000$.

Example 72: Two-Sample Z-test of proportions

A quality inspector finds 10 defective parts in a sample of $n = 500$ parts received from manufacturer A. Out of $m = 400$ parts from manufacturer B, she finds 12 defective ones. A computer-making company uses these parts in their computers and claims that the quality of parts produced by A and B is the same. At the 5% level of significance, do we have enough evidence to disprove this claim?

We test H_0: $p_A = p_B$, or H_0: $p_A - p_B = 0$, against H_A: $p_A \neq p_B$ where p_A (p_B) is the portion of defective parts from manufacturer A (B), respectively.
This is a two-sided test because no direction of the alternative has been indicated.

1. *We first compute the values of the estimated portions*

$$\hat{p}_A = \frac{10}{500} = 0.02 \text{ and } \hat{p}_B = \frac{12}{400} = 0.03.$$

   These estimators are asymptotically normally distributed (sum of n and m Bernoulli variables divided by n and m, respectively) where the means are the true portions p_A and p_B and the variances are $p_A(1-p_A)/n$ and $p_B(1-p_B)/m$. Obviously, $\hat{p}_A - \hat{p}_B$ is also asymptotically normally distributed with mean zero and variance $p_A(1-p_A)/n + p_B(1-p_B)/m$. Thus, inserting the estimated values we standardize and get

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p_A(1-p_A)/n + p_B(1-p_B)/m}} = -0.945.$$

2. *The critical value is $z_{\alpha/2} = 1.96$ (from the table of the standard normal distribution). Since this is a two-sided test we should reject H_0 if $|Z| \geq 1.96$ and accept it otherwise.*

3. *Since Z falls into the acceptance region, we do not have enough evidence to reject H_0. Although the sample proportions of defective parts are unequal, the difference between them appears too small to claim that population proportions are different.*

Alternatively, we can consider a single estimator $\hat{p}$ for the overall proportion of defective products since we assume that H_0 ($p_A = p_B$) holds. But with $p := p_A = p_B$ this implies that

$$VAR[X_A] = VAR[X_B] = p(1-p)$$

if X_A and X_B are Bernoulli distributed with parameter p. Hence, if we estimate the common portion of defective parts as

$$\hat{p} = \frac{number\ of\ defective\ parts}{total\ number\ of\ parts} = \frac{n\hat{p}_A + m\hat{p}_B}{n+m} = 0.0244$$

then we can use it to replace the unknown probability p in the variance estimators $VAR[\hat{p}_A] = \frac{p(1-p)}{n}$ and $VAR[\hat{p}_B] = \frac{p(1-p)}{m}$. Hence,

$$VAR[\hat{p}_A - \hat{p}_B] = \frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m}$$

and get

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m}}} = -0.966.$$

Again, Z falls into the acceptance region and we do not have enough evidence to reject H_0.

8.3 T-tests for Unknown σ

In the previous section we used an estimator for the unknown true variance σ^2. In the special case that our data $X_1, \ldots, X_n$ is normally distributed with mean μ and variance σ^2 and we estimate the unknown mean with

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

we can make use of a result from statistics that tells us the following:
We know that $E[\bar{X}_n] = \mu$ and thus

$$Z = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

must follow a standard normal distribution (we standardized it by subtracting the mean and dividing by the standard deviation!). However, since σ^2 is unknown we estimate it using

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

It can be shown that

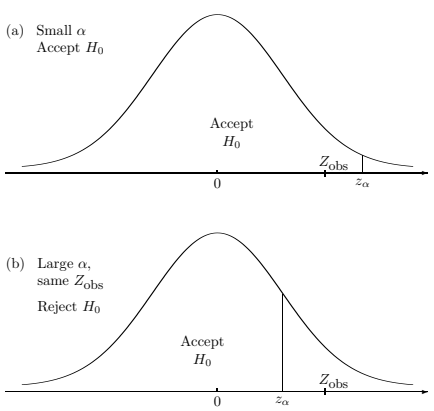
$$T = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}}$$

follows a Student's t-distribution with $n-1$ degrees of freedom. Thus, once μ is fixed (because of our hypothesis H_0) and the data $X_1, \ldots, X_n$ is given we can compute T and check whether it falls into the acceptance or rejection region by looking at the values t_α, $-t_\alpha$ or $t_{\alpha/2}$ depending on the type of alternative. Note that t_α is the value such that

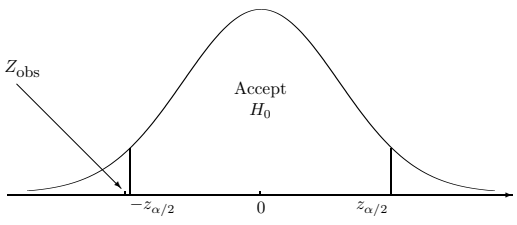
$$P(T > t_\alpha) = \alpha$$

and it can be found in the table of the Student's t-distribution (in the same way we find $t_{\alpha/2}$). Note that the degrees of freedom $n-1$ is a parameter of the Student's t-distribution since the distribution of T changes when the samples size n changes.

8.4 p-Value



So far, we were testing hypotheses by means of acceptance and rejection regions where we need to know the significance level α in order to conduct a test. However, there is no systematic way of choosing α, the probability of making type I error. Of course, when it seems too dangerous to reject a true H_0, we choose a low significance level. But how low? Should we choose $\alpha = 0.01$? Perhaps, 0.001? Or even 0.0001? If we choose α small enough we can always make sure that H_0 is accepted as illustrated on the right.



Also, if our observed test statistic Z_{obs} belongs to a rejection region but it is "too close to call" as illustrated on the left, then how do we report the result?

Formally, we should reject the null hypothesis, but practically, we realize that a slightly different significance level α could have expanded the acceptance region just enough to cover Z_{obs} and force us to accept H_0. Is there a statistical measure to quantify how far away we are from a "too close to call"?

The idea is to try to test a hypothesis using all levels of significance. Then we have two cases:
1) Very small values of α make it very unlikely to reject the hypothesis because they yield very small rejection regions. 2) High significance levels α will make it likely to reject H_0 and corresponds to a large rejection region. We will be forced to reject H_0. The P-value is the boundary value between the accept case 1) and reject case 2). Thus,

the p-value is the lowest significance level α that forces rejection of H_0 and also the highest significance level α that forces acceptance of H_0.

Usually $\alpha \in [0.01, 0.1]$ (although there are exceptions). Then, a P-value greater than 0.1 exceeds all natural significance levels, and the null hypothesis should be accepted. Conversely, if a P-value is less than 0.01, then it is smaller than all natural significance levels, and the null hypothesis should be rejected. Only if the P-value happens to fall between 0.01 and 0.1, we really have to think about the level of significance. This is the "too close to call". A good decision is to collect more data until a more definitive answer can be obtained.
We compute p by fixing $Z_{obs} = z_\alpha$ and selecting $p = \alpha$ such that for a one-sided right-tail alternative we have

$$p = \alpha = P(Z \geq z_\alpha) = P(Z \geq Z_{obs}) = 1 - \Phi(Z_{obs})$$

where Z is standard normally distributed and Z_{obs} is the test statistic. The computation of p is similar for the one-sided left-tail and the two-sided case as well as for T-tests.

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\}$	$1 - \Phi(Z_{\text{obs}})$
	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{Z \leq Z_{\text{obs}}\}$	$\Phi(Z_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\lvert Z\rvert \geq \lvert Z_{\text{obs}}\rvert\}$	$2(1 - \Phi(\lvert Z_{\text{obs}}\rvert))$

We summarize the computation of the p-value for Z-test on the left where we distinguish the three different cases for the alternative hypothesis H_A.

From the definition of the p-value, it is also clear that

$$p = P(\text{observing test statistic } T \text{ that is at least as extreme as } T_{obs} \mid H_0).$$

In Figure 8.2 we illustrate this interpretation (the green shaded area is the p-value here and T_{obs} is the observed data point). Thus, it is wrong to say that the p-value tells us something about the probability that H_0 is true (given the observation)! A high p-value tells us that the observed or even more extreme values of Z_{obs} is not so unlikely (given H_0), and therefore, we see no contradiction with H_0 and do not reject it. Conversely, a low p-value

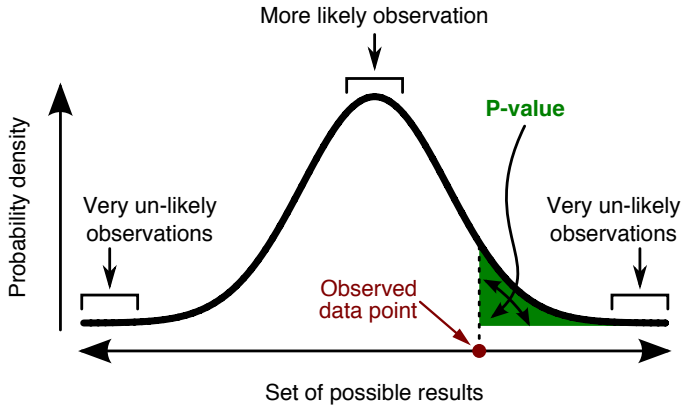


Figure 8.2: Interpretation of the p-value as the probability that we observe a test statistic T that is at least as extreme as T_{obs} given that H_0 is true.

signals that such an extreme test statistic is unlikely if H_0 is true. Since we really observed it, our data are not consistent with H_0 and we reject it.

Example 73: Two-Sample Z-Test of Proportions (revisited)

In Example 72 we computed a test statistic of $Z_{obs} = -0.945$ for a two-sided test which compares the quality of parts produced by two different manufactures. We compute p as

$$p = P(|Z| \geq |-0.945|) = 2(1 - \Phi(0.945)) = 0.3472.$$

This p-value is quite high and indicates that the null hypothesis should not be rejected. If H_0 is true then the chance of observing a value for Z that is as extreme or more extreme than Z_{obs} is 34%. This is no contradiction with the assumption that H_0 is true.

Bibliography

[1] Michael Baron. *Probability and statistics for computer scientists*. CRC Press, 2013.

[2] Radford M. Neal. Sta 247 - week 2 lecture summary. https://towardsdatascience.com/skewed-data-a-problem-to-your-statistical-model-9a6b5bb74e37, Last accessed on 2020-03-22.

[3] Radford M. Neal. Sta 247 - week 2 lecture summary. `http://www.utstat.utoronto.ca/~radford/sta247.F11/lec2.html`, Last accessed on 2020-02-11.