Berner
Fachhochschule

Applied Statistics

Data Engineering

Jasmin Wandel
Version October 14, 2024

Berner Fachhochschule
Engineering and Computer Science
Abteilung MNG

Contents

About the lecture notes

Definitions, theorems, examples and code are presented in boxes. Each category is indicated by a specific color.

Definitions

Orange labeled boxes contain definitions. Mathematical definitions are there to make sure that we speak about the same thing.

Theorems

Red labeled boxes contain theorems and other mathematical facts that can be deduced from the definitions. A theorem requires a proof (how else would we know it’s true?). Because of the lack of time, the proofs of the discussed theorems are often omitted in this course.

Examples

Blue labeled boxes contain examples.

Code

Green boxes contain code (usually Python).

1 Defining the Data

1.1 Vocabulary

Before we start with statistics in depth, we need to define a few terms. This is the only way we can guarantee that we are talking about the same thing.

Population and Sample

The population is the entire group that you want to draw conclusions about and this group must be accurately defined. On the other hand, the sample is the subset of the population that is actually observed.

Ideally, the sample is a random subset of the population which represents the most important characteristics of the population. Consequently, the statistician only needs to deal with the sampling variation and not with a bias introduced via inappropriate sampling. Sometimes in practice more sophisticated sampling strategies are necessary, e.g. stratified sampling can be advantageous, if relevant characteristics are unequally balanced in the population.

Population

A population is just an accurately defined group and can be of any sample size, e.g.

- all data engineer students at the BFH
- all data engineer students in Switzerland
- all data engineer students in the world
- all data engineer students in the world at present and in the future

Classically, we think of very large groups when we think about populations. However, a population can be of any size. It naturally does not make sense to draw a sample from a very small population. A complete survey is possible for a small population and there is no need to take a random sample. In this situation, the population can be described completely and it is often not necessary to perform so-called inferential statistics (see later).

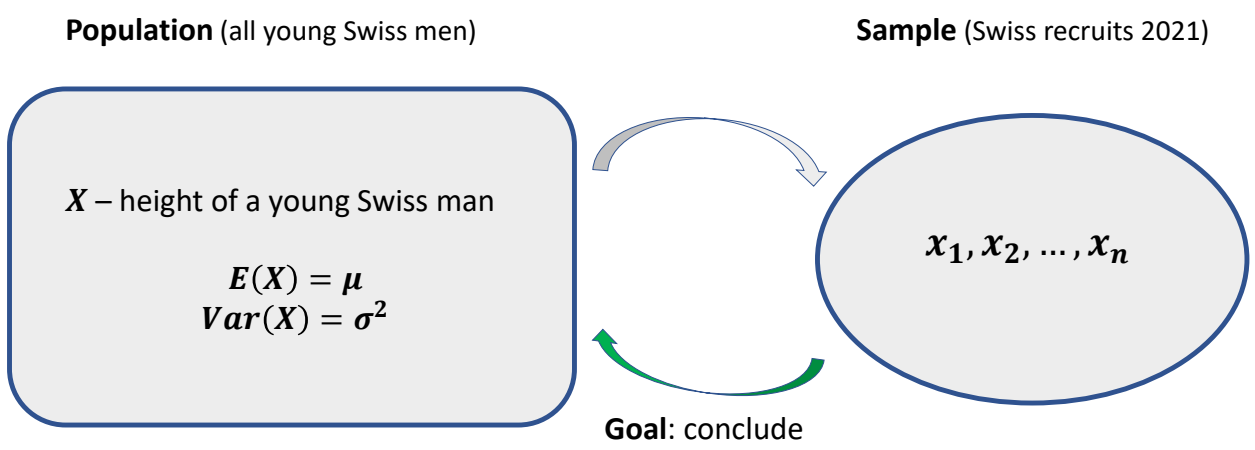
1.2 Link to probability theory

There is a strong link between probability theory and statistics. In probability theory you learned to model phenomena with uncertainty, e.g. the outcome of the random experiment of rolling a die. On the other hand, statistics aims to perform inference for probabilistic models. In that sense, the population can be seen as the „generating random variable X“ and the sample as the realizations from this probabilistic model, i.e. the concrete realizations $x_1, \ldots, x_n$.

Probability vs. statistics

Let us assume you want to make inference about the height of young Swiss men. The height of a person is a priori a random variable X and in that sense defines the probabilistic model –

the population – we want to make conclusions on. On the other hand, the data from the Swiss recruits is a sample form the defined population of interest (young Swiss men) and consists of the given realizations $x_1, \ldots, x_n$ of X, namely the measured heights of the recruits. These recruits are now representatives of young Swiss men and give us important information about the complete population.



1.3 Types of variable

Variable

A variable is an aspect of an individual in the sample that is measured or recorded.

A first step in choosing how best to display and analyze data, is to classify the measured variables into their different types – different types of variables ask for different methods!

Types of Variable

We differentiate the following types of variable:

- **Numerical (quantitative) variables**: Numerical variables take on a numerical value with an objective meaning. Classic examples are weight, height and age. Here we differentiate between **discrete** numerical variables and **continuous** numerical variables. Whether a numerical variable is discrete or continuous is defined via the probabilistic model assumed that generates the observations. If the generating random variable is continuous, so is the observed variable. The same is true for discrete numerical variables. As a side note: Of course, realizations are always on a discrete scale, as we cannot measure with arbitrary precision.
- **Categorical (qualitative) variables**: Categorical variables can take on a finite number of values in an arbitrary range. In particular, categorical variables, even if they take on numerical values, have no numerical meaning.
  - **Binary variable**: Categorical variables with only two groups, e.g. sex.

- **Nominal variables**: Categorical variables without natural order. Classic examples of nominal variables are gender or place of birth.
- **Ordinal variables**: Categorical variables whose values (=possible values) are in a natural order. Consider, for example, the BFH evaluation questionnaires. Here you can typically tick one of the following answers for each question: „very good“, „good“, „bad“, „very bad“. Sometimes ordinal variables are created from numerical variables by dividing their value range into a finite number of intervals, i.e. by categorization.

When we speak ov variables in statistical modeling, we further differentiate between the so-called **outcome** variable and the **explanatory** variables.

Outcome Variable

The outcome variable is the focus of attention when it comes to statistical modeling. Its the variable whose variation and occurrence we are seeking to understand and the variable we try to model. It is the outcome variable's type which defines the appropriate statistical method to be used.

Equivalent terms: response variable, dependent variable, y-variable

Explanatory Variables

With the explanatory variables we aim to explain the outcome variable. The statistical model tries to quantify their influence on the outcome variable.

Equivalent terms: exposure variable, independent variable, x-variable

Note that classical statistical models assume that the measurements of explanatory variables have no uncertainty – only the uncertainty of the outcome is modeled. This uncertainty represents a priori the model uncertainty. Measurement uncertainties are often ignored in statistics. However, very often they are negligible in comparison to the model uncertainty.

2 Basic Concepts of Probability Theory

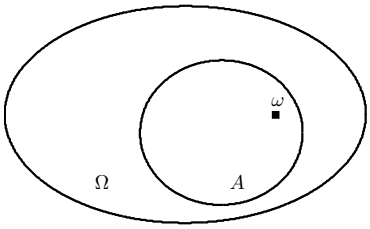
As mentioned in the previous chapter, there is a strong link between probability theory and statistics. Without some basic knowledge about probability theory, it is not possible to perform inferential statistics as statistical models always depend on probabilistic models. Even descriptive statistics heavily depend on probability theory, although not quite as obvious. One can only understand the plausibility of estimation, if one understand the data generating process as a random process.

2.1 Random events

Probability theory and statistics are used to investigate *random* events and their laws. On the other hand, there are so-called *deterministic* processes in which the result of an experiment can be predicted exactly under precisely known conditions. Deterministic processes are not influenced by chance. A random experiment, on the other hand, is characterized by the fact that its outcome is uncertain, i.e. random, within the scope of various possibilities. It should also be repeatable (at least mentally) any number of times under the same conditions.

Elementary event, event, sample space, σ-algebra

- An **elementary event** ω is a possible outcome of an experiment.
- The **sample space** Ω is the set of all elementary events.
- An **event** A is a subset of the sample space ($A \subset \Omega$).
- The **σ-algebra** $\mathcal{F}$ is the collection of all events considered.



Dice roll

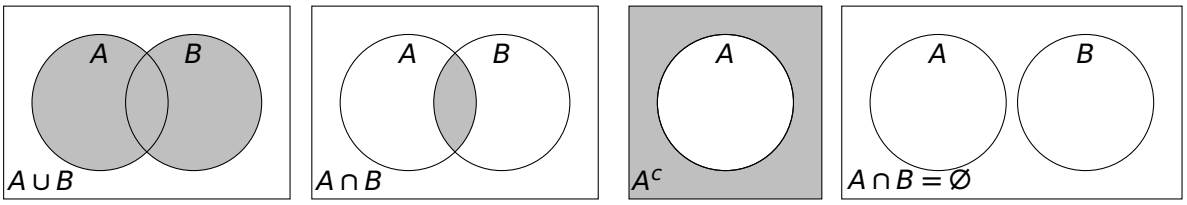
For a random dice roll, $\Omega = \{1, 2, 3, 4, 5, 6\}$. An event $A \subseteq \Omega$ is understood as a subset of Ω; it occurs if the random elementary event ω is in the set A. (The event „The dice roll is even“ is thus understood as the subset $\{2, 4, 6\}$ of Ω). The set of all such events („The dice roll is even“, „The dice roll is odd“, „The dice roll is 2“, „The dice roll is less than 5“, . . .) refers to the σ-algebra $\mathcal{F}$.

2.2 Linking of events

From single random events we can form additional more complicated events with the help of simple relations:

1. **The „or“ link** $A \cup B$: This event occurs when A *or* B arrives.
2. **The „and“ link** $A \cap B$: This event occurs when A *and* B arrive.
3. **The complementary event** $A^c = \Omega \setminus A$: This event occurs exactly when A does not arrive.

It is said that the events A and B are **disjoint** if $A \cap B = \emptyset$, i.e. the simultaneous occurrence of A and B is impossible.



Note: In the diagrams, the rectangular frames each represent the sample space Ω.

2.3 Probability measure

The outcome of a random experiment is uncertain. With a die, for example example, we know that it will show one of the sides „1“ to „6“; the result space of a dice roll is therefore $\Omega = \{1, 2, 3, 4, 5, 6\}$. With a fair die, we expect each result to be „equally likely“, i.e. that each number occurs equally often on average. However, we may also be dealing with a unfair die and find that the 6 appears on average as often as all the other numbers put together.

The *probability* of an event tells us how often the event occurs on average over many repetitions of a random experiment. The probability of an event $A \subseteq \Omega$ is a number between 0 and 1 and is denoted by $\mathrm{P}(A) \in [0, 1]$. A probability of 0 means that an event never occurs, a probability of 1 means that it always occurs; a probability of $\frac{1}{3}$ means that the event occurs on average one third of the time.

In the case of the fair die, the following applies

$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \ldots = \mathrm{P}(\{6\}) = \frac{1}{6};$$

For an unfair die, assume that the following applies:

$$\mathrm{P}(\{6\}) = \mathrm{P}(\{1, 2, 3, 4, 5\})$$

(The number 6 is as likely as all other numbers together.)

Probability measure

Let Ω be a sample space and $\mathcal{F}$ be a σ-algebra. A probability measure is a function $\mathrm{P} : \mathcal{F} \to [0, 1]$ that assigns a value between 0 and 1 to an event $A \subset \Omega$: $\mathrm{P}(A) \in [0, 1]$.
It obeys the following properties (axioms of Kolmogorov):

a) $0 \leq \mathrm{P}(A) \leq 1$ for every event $A \subset \Omega$

b) $\mathrm{P}(\Omega) = 1$

c) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ for *disjoint* (mutually exclusive) events A and B

(The fact that the probability of the entire event space Ω is equal to 1 also makes intuitive sense: after all, in 100% of cases we get an elementary event from Ω.)

These three properties can also be used to derive some additional useful properties:

4. *Complementary event*: $P(A^c) = 1 - P(A)$ for any event $A \subseteq \Omega$. In particular, $P(\emptyset) = P(\Omega^c) = 0$.
5. *Union*: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events $A, B \subseteq \Omega$, and therefore in particular $P(A \cup B) \leq P(A) + P(B)$.

Probability of unions

We saw as a natural consequence of the axioms of Kolmogorov that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

More generally: Let $A_1, A_2, \ldots, A_n$ be events. Then,

$$\begin{aligned} P(A_1 \cup A_2 \cup \ldots \cup A_n) &= \sum_{i_1=1}^{n} P(A_{i_1}) - \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} P(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \ldots \end{aligned}$$

Laplace experiment

In a Laplace experiment every elementary event is equally probable, i.e. $P(\{\omega_i\}) = 1/|\Omega|$. As a consequence, the probability of an event $A \subset \Omega$ in a Laplace experiment can be calculated as follows:

$$P(A) = \frac{\#\text{ favourable cases}}{\#\text{ possible cases}} = \frac{|A|}{|\Omega|}$$

Many real phenomena can at least be approximated as Laplace experiments; in the case of these phenomena, it can be assumed that each elementary event is equally probable. However, we will often also deal with situations where this assumption is not permissible. You should therefore ask yourself for each probability experiment whether the prerequisite for a Laplace experiment is actually fulfilled.

2.4 Binomial coefficient

As we have seen in the previous section, probabilities can be determined in a Laplace probability space by counting the „favorable“ and the „possible“ elementary events. In simple cases, this counting can be done „by hand“; in many cases, however, you cannot avoid calculating the numbers you are looking for in a suitable way.

Combinatorics deals with the investigation of such counts for cases where an explicit enumeration of all possible cases is no longer practicable. Hereby, *binomial coefficients* are an important tool in combinatorics. We will introduce these directly using an example.

Coin flip

We flip a coin a total of 50 times and are interested in the event B =„the coin shows heads“ exactly 17 times. It is not difficult to see that there are now $|\Omega| = 2^{50} \approx 1.126 \cdot 10^{15}$ different (equally probable) elementary events – but we probably don’t want to list them explicitly.

In how many of these possible outcomes does the event B occur, i.e. does it appear exactly 17 times? The answer to this question can be calculated using a *binomial coefficient*. In general,

the expression

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k \cdot (k-1) \cdot \ldots \cdot 2 \cdot 1}$$

describes is the number of possibilities to select exactly k elements from n elements (whereby the order of selection should not play a role). In other words, the binomial coefficient $\binom{n}{k}$ is the number of k-elementary subsets of any set of n elements.

In this example, we are interested in the number of possibilities to select exactly 17 out of 50 throws in which „head“ should appear. We can imagine that we distribute 17 „head“ events one after the other over the 50 throws. There are $50 \cdot 49 \cdot \ldots \cdot 35 \cdot 34$ possibilities. With this counting method, however, we have counted each result (= sequence of 50 throws with exactly 17 heads) several times, as the same result can be obtained in many different ways. More precisely: We have counted each result exactly 17! times, as 17 individual „head“ events can be distributed in $17 \cdot 16 \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = 17!$ different orders over 17 given throws (all of which lead to the same result). We must therefore divide our first count by this factor; there are therefore exactly

$$|B| = \frac{50 \cdot 49 \cdot \ldots \cdot 35 \cdot 34}{17!} = \binom{50}{17} \approx 9.847 \cdot 10^{12}$$

possibilities of rolling heads exactly 17 times in 50 throws. For the probability you are looking for, you therefore get

$$\mathrm{P}(B) = \frac{|B|}{|\Omega|} = \frac{\binom{50}{17}}{2^{50}} \approx 0.0087.$$

Binomial coefficient

The Binomial coefficient counts number of possibilities to choose k out of n elements without replacement:

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k \cdot (k-1) \cdot \ldots \cdot 2 \cdot 1}$$

The following applies to $n \in \mathbb{N}$:

- $\binom{n}{k} = \binom{n}{n-k} \quad (0 \leq k \leq n)$
- $\binom{n}{0} = \binom{n}{n} = 1$
- $\binom{n}{1} = \binom{n}{n-1} = n$

Binomial coefficient

The binomial coefficient can be calculated via `binom(n,k)` from `scipy.special`, or alternatively via `math.comb(n,k)` from the standard library.

2.5 Independence

Independent events

Two events A and B are called **independent**, if $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$.

Dice

If we throw two (fair or unfair) dice separately, it is intuitively clear that they do not influence each other. We then say that, for example, the events A =„die 1 shows 6“ and B =„die 2 shows 6“ are *independent*, and may *multiply* the corresponding probabilities:

$$\mathrm{P}(A \cap B) = \mathrm{P}(\text{„both dice show 6“}) = \mathrm{P}(A) \cdot \mathrm{P}(B) = \mathrm{P}(\text{„die 1 shows 6“}) \cdot \mathrm{P}(\text{„die 2 shows 6“}).$$

For two fair dice, the probability sought is therefore $1/6 \cdot 1/6 = 1/36$, as we calculated earlier as Laplace probability. For two unfair dice („6“ has probability 0.5), the probability we are looking for is $1/2 \cdot 1/2 = 1/4$. We have not been able to calculate this yet!

Note: In this sense, the events A =„Dice roll is even“ and B =„Dice roll is 1 or 2“ are also independent for a simple roll of a fair die, since

$$\mathrm{P}(A \cap B) = \mathrm{P}(\text{„Dice roll is 2“}) = 1/6 = 1/2 \cdot 1/3 = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

2.6 Conditional probabilities

Conditional probability

Let A and B be events (with $\mathrm{P}(B) > 0$). The **conditional probability of** A **given** B is defined as

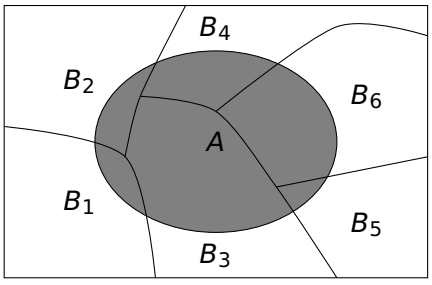
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \ .$$

You can also use conditional probabilities to calculate unconditional probabilities by applying conditional probabilities to several different cases.

Law ot total probability

Assume $B_1, B_2, \ldots, B_k$ are disjoint events with $B_1 \cup B_2 \cup \ldots \cup B_k = \Omega$. Then, the probability of any event A is

$$\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{k} \mathrm{P}(A|B_i)\,\mathrm{P}(B_i) \ .$$



The Bayes' theorem establishes a connection between the conditional probability $\mathrm{P}(A|B)$ and the inverse conditional probability $\mathrm{P}(B|A)$. This is useful because you can often estimate/measure the conditional probability in one direction but are actually interested in the other direction.

Bayes' theorem

Let A and B be events with $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$. Then, we have

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

In the setting of the law of total probability, we have

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A \cap B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}{\sum_{j=1}^{k} \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)} \,.$$

Dice

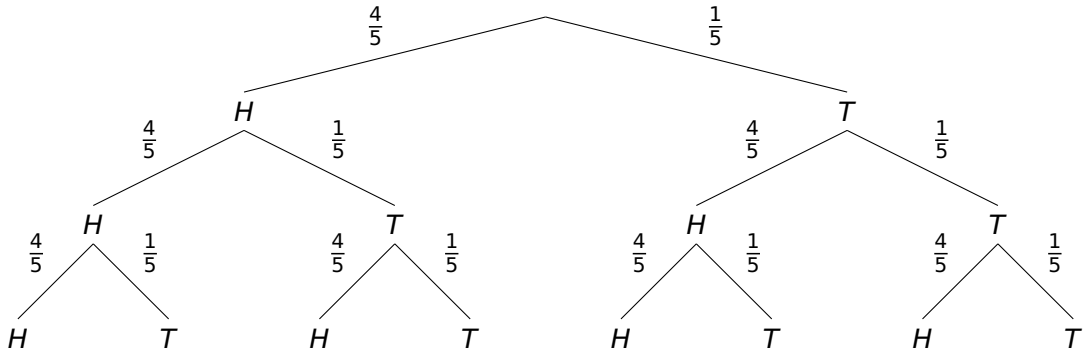
Conditioning on an event A can be used to refer to only a part of the event space. As an example, we again consider the simple dice roll and the events $\mathrm{P}(A)$ =„dice roll is even“ and $\mathrm{P}(B)$ =„dice roll is 4, 5 or 6“. For a fair die we get

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(\text{„Dice roll is 4 or 6“})}{\mathrm{P}(\text{„Dice roll is 4, 5 or 6“})} = \frac{1/3}{1/2} = \frac{2}{3},$$

which is intuitively correct, as exactly 2 of the 3 numbers 4, 5 and 6 are even.

2.7 Probability trees

When determining probabilities for multi-stage random experiments, so-called *event trees* are sometimes helpful. For example, the following event tree shows three consecutive unfair coin tosses. H represents the elementary event „head“ and T „tail“.



If we specify the probabilities at each level of the tree (as is the case in the tree above), we can, for example, calculate the probability of the elementary event (H, T, H) by multiplying the individual probabilities of the path: $\mathrm{P}((H, T, H)) = 4/5 \cdot 1/5 \cdot 4/5 = 0.128$.

In general, this results in the **1st path rule**: *The probability of an elementary event in a multi-stage random experiment is equal to the product of the (conditional) probabilities along the path corresponding to this elementary event in the event tree*.

How likely is the event A =„that no side falls twice in a row“ in the same random experiment? In the tree diagram, only the two paths (H, T, H) and (T, H, T) correspond to this event. Due to the additivity, we thus obtain

$$\mathrm{P}(A) = \mathrm{P}\big(\{(H, T, H)\} \cup \{(T, H, T)\}\big) = 4/5 \cdot 1/5 \cdot 4/5 + 1/5 \cdot 4/5 \cdot 1/5 = 0.16 \,.$$

In general, this results in the **2nd path rule**: *The probability of an event is equal to the sum of the probabilities of the paths that form this event in the event tree*.

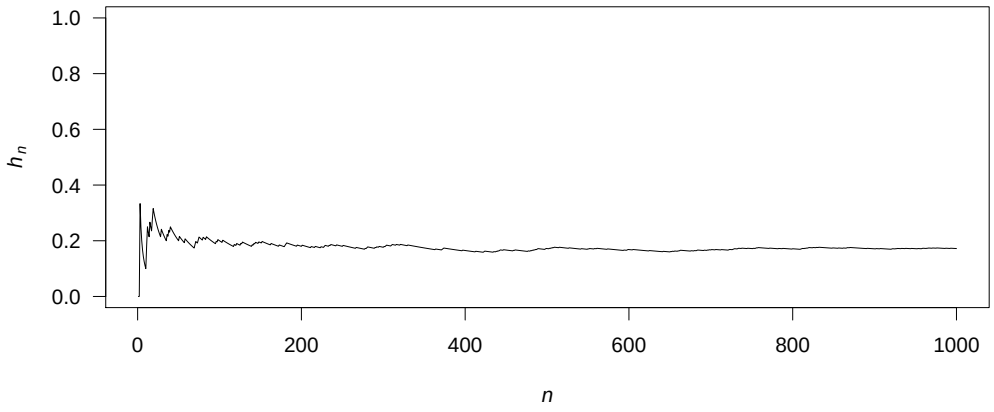
These path rules are particularly helpful for multi-stage random experiments in which the individual stages *not* are independent of each other. The individual branches of the tree must then be labeled with the corresponding *conditional* probabilities.

2.8 Statistical/Empirical probability

Until now, we have always assumed that we knew exactly the probability of a given event occurring. However, there are certainly many situations in which we do not know the probability of a random event. Imagine the following situation: You invite all your friends over for a barbecue on August 21. You are faced with a random experiment with two possible outcomes: Either there will be such bad weather on August 21, that you will have to cancel the garden party, or there will be enough good weather for the planned barbecue. Surely you have good reason to believe that one of these cases will happen more likely. You know from long personal experience that good weather is more likely in late summer, therefore, good weather is to be expected. „Equal probability“ is not at all plausible here based on this experience, i.e. it is certainly not a Laplace experiment.

In such situations, we must rely on experience somehow. We naturally gain experience by repeating the random experiment. In our specific example, we are interested in how likely it is that the weather will be bad on a late summer day. From the Meteo-Suisse database we know that of the 400 late summer days in the last 40 years years, there have only been 44 days with bad weather. We can therefore use the *relative frequency* of a bad weather situation in late summer late summer 44/400 = 0.11 as an (approximate) probability, that you will have to cancel your garden party due to bad weather.

Let us look at the dice experiment again. You have the suspicion that your opponent’s die is unfair and that the six in particular is more likely to come up than the other numbers. How can you check this suspicion? Of course, we could roll the die a few times and keep statistics on the occurrence of a six. We assume that the relative frequency of the event „number of eyes is six“ after repeating the random experiment many times, is a good approximation of the probability for this event. The following figure shows an example of such an experiment.



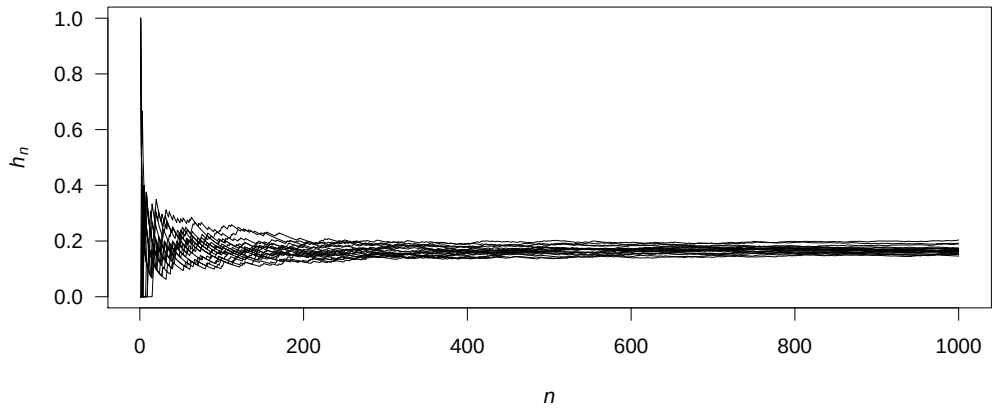
The relative frequency $h_n(A)$ of an event A with n repetitions is calculated as the quotient of the absolute frequency $H_n(A)$ of this event (i.e. the number of repetitions in which A occurs) and the total number of repetitions:

$$h_n(A) = \frac{H_n(A)}{n}.$$

It is intuitively clear that the larger n is, i.e. the more trials we observe, the smaller the deviation of the relative frequency from the probability of a random event. In this respect, the probability P(A) could be regarded as a „limit“, against which the relative frequencies $h_n(A)$ converge with a high probability for $n \to \infty$. We call this stabilization of the relative frequencies *empirical law of large numbers*.

The figure below, shows the course of the relative frequencies for the event frequencies for the event „number of eyes is six“ for 20 of such test series. It is noticeable that the progression curves change with increasing n and approach the desired probability.

This principle can be used in many cases to approximately determine probabilities using so-called *Monte Carlo simulations*.



Relative frequencies have the same properties as probability masses:

1. $0 \leq h_n(A) \leq 1$ for all events A
2. Normalization: $h_n(\Omega) = 1$, because the certain event Ω occurs with every repetition of the experiment
3. Additivity: $h_n(A \cup B) = h_n(A) + h_n(B)$ if A and B are disjoint and the two events cannot occur simultaneously.

3 Random Variables

A random variable is a variable that takes numerical values which depend on the outcome of a random experiment. It is an assignment of an event to a real number.

Random variable

A random variable X is a function mapping a sample space Ω to $\mathbb{R}$ (or a subset or $\mathbb{R}$):

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

A random variable X induces a probability measure on $\mathbb{R}$.

Note: Random variables are denoted by capital letters (e.g. X), whereas a realized value of a random variable is denoted by a lower case letter (e.g. x):

- Capital letter X: description of an experiment (e.g., „measurement of the length of a tibia“)
- Lower case letter x: outcome of the experiment, i.e. concrete realization of X (e.g., 38.5 cm)

We differentiate between two main types of random variables:

- A **discrete random variable** is a random variable with finite (or countable) image (i.e. set of possible values):
  $X : \Omega \to \{x_1, x_2, \dots\}$
- A **continuous random variable** is random variable whose image contains an interval, or $\mathbb{R}$.

Two types of random variables

a) **Family size**: The random selection of a family from a population with X being be the number of children of the selected family. Therefore, possible values for X are $X \in \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$. If e.g. 23% of the families have 2 children, then $P(X = 2) = 0.23$; if 72% of the families have *at most* 2 children, then $P(X \in \{0, 1, 2\}) = P(X \leq 2) = 0.72$. In this example X describes a discrete random variable.

b) **Length of fish**: Measurement of the length of a fish randomly sampled from a population with X being the length in cm. If e.g. 64% of the fish have a length between 11.5 cm and 16.2 cm, then $P(X \in [11.5, 16.2]) = P(11.5 \leq X \leq 16.2) = 0.64$. In this example X describes a continuous random variable.

Cumulative distribution function (cdf)

The cumulative distribution function (cdf) of a random variable X is defined as

$$F_X(x) := P(X \leq x).$$

Properties of a cdf F_X:

- F_X is monotonically increasing
- $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$
- $P(a < X \leq b) = F_X(b) - F_X(a)$

Quantile function

Let X be a random variable with distribution function F_X and let $\alpha \in (0,1)$. The **α-quantile** of X fulfills

$$\mathrm{P}(X \leq q) \geq \alpha \quad \text{and} \quad \mathrm{P}(X \geq q) \geq 1-\alpha.$$

Roughly speaking, an α-quantile q is a value where the graph of the cdf crosses (or jumps over) α. There is an inverse relation between the quantiles and the values of the cdf. In fact, the quantile function is given by the (generalized) inverse function of the cdf.

3.1 Discrete random variables

A discrete random variable X has a finite (or countable) image (i.e. set of possible values):

$$X : \Omega \to \{x_1, x_2, \ldots\}$$

The distribution of a discrete random variable is mainly characterized by its probability mass function. Of course, the cdf describes the distribution of X as well. In fact, there is a one-to-one correspondence between the cumulative distribution function and the probability mass function.

Probability mass function (pmf)

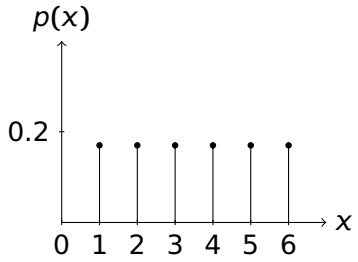
The probability mass function $p(x_k) := \mathrm{P}(X = x_k)$ of a a (discrete) random variable has the following properties:

- For each set $A \subset \{x_1, x_2, \ldots\}$, we have
$$\mathrm{P}(X \in A) = \sum_{k: x_k \in A} \mathrm{P}(X = x_k)$$
- Normalization: $\sum_k \mathrm{P}(X = x_k) = 1$
- Connection to cdf: $F_X(x) = \mathrm{P}(X \leq x) = \sum_{k: x_k \leq x} \mathrm{P}(X = x_k)$

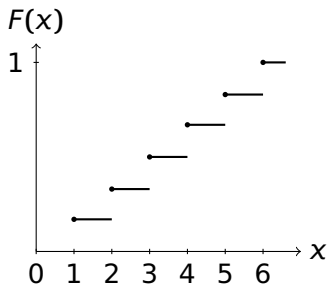
Fair die

A die can take values in $\{1, 2, \ldots, 6\}$; if it is fair, it takes all values with the same probability. Its probability mass function and cumulative distribution function look as follows:

Probability mass function (pmf)



Cumulative distribution function (cdf)



The expected value of a random variable describes which value we can expect on average from this random variable, if we could repeat the experiment infinitely many times. Naturally, for a fair die we have the given pmf:

x_k	1	2	3	4	5	6
$P(X = x_k)$	1/6	1/6	1/6	1/6	1/6	1/6

On average, we would expect 3.5 spots (even though this specific value is no realization of X).

We can also do the same thought experiment for a non-fair die. Let us assume the following pmf:

x_k	1	2	3	4	5	6
$P(X = x_k)$	1/10	1/10	1/10	1/10	1/10	1/2

Clearly, the expected number of spots is in this situation corresponds to the *weighted* mean number of spots, i.e. $0.1 \cdot (1 + 2 + 3 + 4 + 5) + 0.5 \cdot 6 = 4.5$.

Generally, the expected value $E(X)$ of a discrete random variable X corresponds to the weighted mean of all possible values of X. The weights are determined by the corresponding probability mass function. This leads to the following definition.

Expected value

The expected value of a discrete random variable X is defined as

$$E(X) := \sum_k x_k \cdot P(X = x_k) \, .$$

As direct consequences of the given definition, we get the following properties of the expected value:

- The expected value is **linear**. That means that $E(X)$ fulfills the following two properties.
  a) $E(a \cdot X) = a \cdot E(X), \; a \in \mathbb{R}$
  b) $E(X + Y) = E(X) + E(Y)$ for any two random variables X and Y
  (This property also applies in particular to random variables that are not independent. We will look at what this exactly means in more detail later.)
- Consequently, for any $a, b \in \mathbb{R}$: $E(a \cdot X + b) = a \cdot E(X) + b$.
  (The expected value of the constant b is b.)

Besides the expected value the variance of a random variable is as well an important characteristic of the distribution of a random variable. While the expected value is a measure for „the center“ of the distribution, the variance of a random variable describes how much the values vary around the expected value. It is therefore a measure of the dispersion of the random variable.

Variance

The variance of a discrete random variable X is defined as

$$\mathrm{Var}(X) := E\left((X - E(X))^2\right) = \sum_k (x_k - E(X))^2 \cdot P(X = x_k) \, .$$

As a measure of variation of a random variable, the standard deviation $\sigma(X)$ is often used, which is defined as the (positive) root of the variance:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

The standard deviation has the advantage that it (unlike the variance) has the same unit as the random variable X.

In contrast to the expected value, variance and standard deviation are not linear. The following rules of calculation apply to variance:

- a) $\mathrm{Var}(a \cdot X) = a^2 \cdot \mathrm{Var}(X),\ a \in \mathbb{R}$

  b) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ *only if* X and Y are independent (see later)
- As a consequence, the following applies to any $a, b \in \mathbb{R}$: $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$. (The variance of the constant b is 0.)

For the standard deviation, this results in:

- a) $\sigma(a \cdot X) = a \cdot \sigma(X),\ a \in \mathbb{R}$

  b) $\sigma(X + Y) = \sqrt{(\sigma(X))^2 + (\sigma(Y))^2}$ *only if* X and Y are independent.
- Consequently, for any $a, b \in \mathbb{R}$: $\sigma(a \cdot X + b) = a \cdot \sigma(X)$.

Not that the terms defined here, namely expected value and variance, are not to be confused with the associated quantities of descriptive statistics! The quantities $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ are based on the theoretical probabilities with which the random variable X takes its values. In empirical studies, formulas similar to those used here are used to calculate the empirical mean $\bar{x}$, the empirical variance s^2 and the empirical standard deviation s. These are important characteristics of a data set and can be regarded as estimates for $\mathrm{E}(X)$, $\mathrm{Var}(X)$ and $\sigma(X)$ of a random variable X underlying the data set. This makes sense if we remember that we understand data as realizations of a probabilistic model. The random variable X represents the probabilistic model with realizations $x_1, x_2, \ldots$ that correspond to the sample actually observed. We will discuss this in more detail later on.

3.2 Special discrete distributions

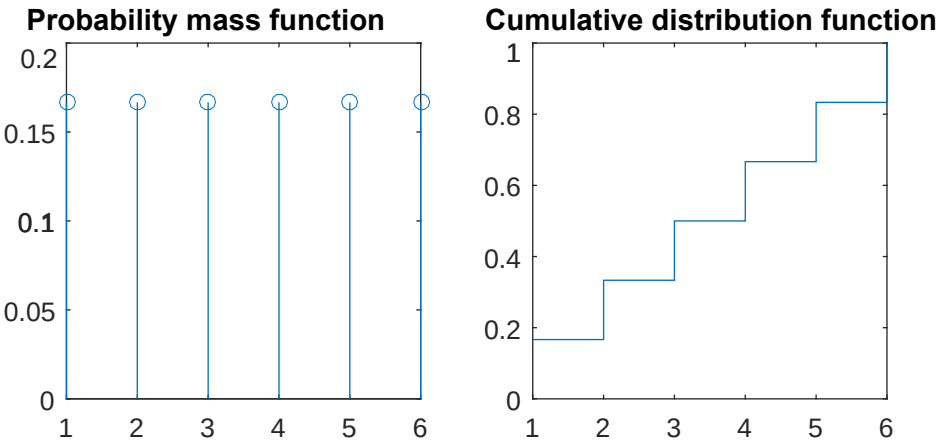
So far, we have talked about discrete random variables in very general terms. However, there are some special distributions that prove to be very helpful in application. We will get to know the most important discrete distributions in this section.

Discrete uniform distribution

If a discrete random variable X assumes all possible values $x_1, x_2, \ldots, x_n$ with the same probability, it is said to be uniformly distributed.

Fair die

The number of dots X when throwing a fair dice is obviously equally distributed. The values $1, \ldots, 6$ are each assumed to have the probability $\mathrm{P}(X = x_k) = \frac{1}{6}$ (for all $k = 1, \ldots, 6$).



Discrete uniform distribution

In general, the probability distribution of a discrete uniformly distributed random variable X is

$$\mathrm{P}(X = x_k) = \frac{1}{n}, \quad k = 1, 2, \ldots, n.$$

For the expected value and the variance this means

$$\mathrm{E}(X) = \frac{1}{n} \sum_{i=k}^{n} x_k$$

$$\mathrm{Var}(X) = \frac{1}{n} \left(\sum_{k=1}^{n} x_k^2 - \frac{1}{n} \left(\sum_{k=1}^{n} x_k \right)^2 \right).$$

With the example of a fair die, we have already considered the special case $x_1 = 1, x_2 = 2, \ldots x_n = n$. In this situation the expected value and variance simplify to

$$\mathrm{E}(X) = \frac{n+1}{2}$$

$$\mathrm{Var}(X) = \frac{n^2 - 1}{12}.$$

Bernoulli / Binomial distribution

The Bernoulli distribution is the simplest non-trivial discrete distribution function.

Bernoulli distribution

A discrete random variable X that can only take the values 0 and 1 is said to have Bernoulli distribution. The distribution is specified by the probability $\pi := \mathrm{P}(X = 1)$. We write $X \sim$ **Bernoulli**(π).

Binomial distribution

The binomial distribution deals with repeated random experiments in which only two outcomes are possible or of interest, such as the coin toss (heads or tails). Classically, we call the two

possible outcomes *success* and *failure* (whereby it is of course arbitrary a priori whether one wants to call „heads“ a success or a failure in the case of a coin toss).

In general, a binomially distributed random variable X with parameters n and π counts the number of successes if such an experiment is repeated n times and the probability of success in each attempt is π. It is assumed that the individual trials are successful independently of each other or not. X can therefore take the values $0, 1, 2, \ldots n$. The probability distribution of a binomially distributed random variable $X \sim \text{Bin}(n, \pi)$ is

$$\text{P}(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

The binomial coefficient describes the number of possibilities that there are to have exactly k successes with a total of n attempts. Each of these possibilities occurs with probability $\pi^k(1-\pi)^{n-k}$.

For the expected value and the variance one can show that

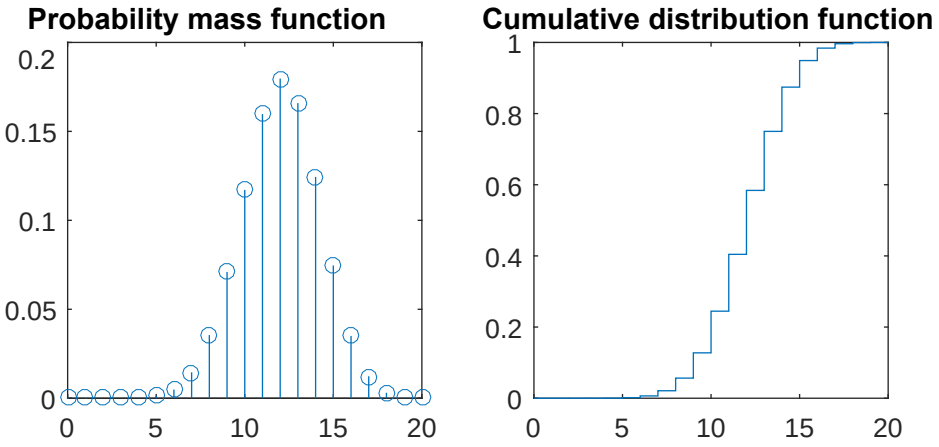
$$\text{E}(X) = n\pi$$
$$\text{Var}(X) = n\pi(1-\pi).$$

Note: The sum of n independent identically Bernoulli-distributed random variables is exactly a binomially distributed random variable.

Penalty

The number of goals X in the penalty shootout if the probability of scoring is constant. For example, in n attempts with a probability of scoring $\pi = 0.6$, the probability of scoring exactly 12 times is

$$\text{P}(X = 6) = \binom{20}{12} 0.6^{12} 0.4^8 = 0.1797.$$



Poisson distribution

Poisson process

In a Poisson process, events occur independently of each other at random points in time at a constant rate λ (expected number of events per time unit). (Here we consider *not* individual discrete „trials“, but a continuous time axis!) Typical events whose occurrence can be modeled

approximately by a Poisson process: Failure of a component of a machine; new customer joins the back of the queue; mutations on a chromosome; accidents on a particular stretch of road; etc.

If we now count the number of events during a certain period of time d in such a Poisson process, the resulting random variable is *poisson distributed* with parameter $\lambda = \overline{\lambda} \cdot d$ (expected number of events).

Road accidents

On average, 0.4 accidents per month occur on a stretch of road. This can be modeled by a Poisson process with a constant rate $\overline{\lambda} = 0.4$ [accidents per month]. The number of accidents in a month is then Poisson distributed with parameter $\lambda = 0.4$ [accidents]; the expected number of accidents in a year is Poisson distributed with parameter $\lambda = 12 \cdot 0.4 = 4.8$ [accidents].

Poisson distribution

A discrete random variable $X \in \mathbb{N}_0$ has Poisson distribution with parameter $\lambda > 0$, if

$$P(X = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0,$$

with expected value and variance

$$\begin{aligned} \mathrm{E}(X) &= \lambda \\ \mathrm{Var}(X) &= \lambda. \end{aligned}$$

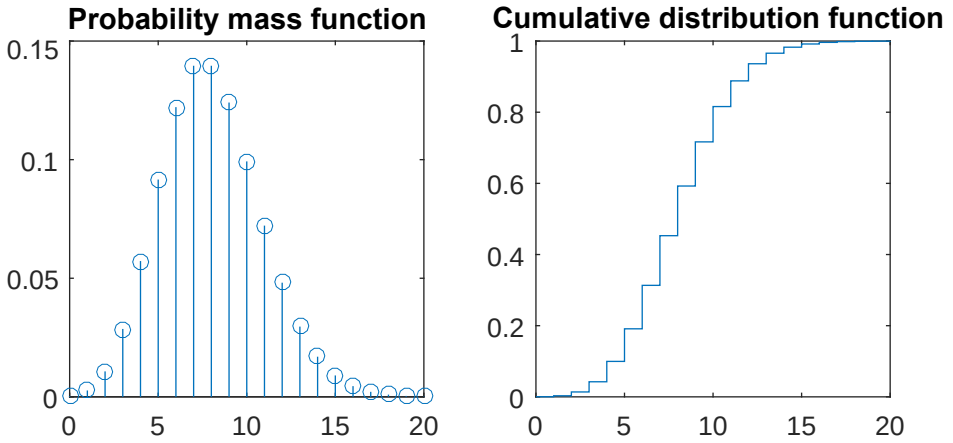
The normalization factor $\mathrm{e}^{-\lambda}$ ensures that all probabilities add up to 1, i.e. $\sum_{k=0}^{\infty} P(X = k) = 1$ applies. (As we know, $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \mathrm{e}^{\lambda}$).

We write $X \sim \mathrm{Po}(\lambda)$.

Call center

A call center receives an average of 2 calls per quarter of an hour. The calls can be modeled using a Poisson process with a constant rate $\overline{\lambda} = 2$ [calls per quarter hour]. Consequently, the number of calls per hour is $X \sim \mathrm{Po}(8)$ and the probability that at least 2 calls will come in during the next hour is

$$\mathrm{P}(X \geq 2) = 1 - \mathrm{P}(X < 2) = 1 - \big(\mathrm{P}(X = 0) + \mathrm{P}(X = 1)\big) = 1 - \left(\mathrm{e}^{-8} + 8 \cdot \mathrm{e}^{-8}\right) = 0.9970.$$



The Poisson distribution can also be regarded as an approximation of the binomial distribution. To do this, we let the number of trials n tend towards infinity for a binomially distributed random variable X and at the same time assume that $\mathrm{E}(X) = n\pi = \lambda$ remains constant. Then X is a Poisson-distributed random variable in the limit with possible values $0, 1, 2, \ldots$.

The requirement that $n\pi = \lambda$ remains constant in the limiting process implies that the probability π of a „success“ tends to 0. This is the formal justification for using the Poisson distribution to model the number of *rare* events in a fixed time window or space. As a rule of thumb, the Poisson distribution can be used as an approximation of the binomial distribution for $n \geq 50$ and $\pi \leq 0.05$ (but this is really only a rule of thumb and not a strict condition). This approximation is practical because the distribution function of a binomial distribution quickly becomes computationally expensive as n increases. The approximation is particularly useful when we do not have a computer available to calculate probabilities.

Printing errors

The number of printing errors X per page in a book can be regarded as a Poisson-distributed random variable. We can assume that each letter on a page has a small probability π of being misplaced. Since the number of letters per page n is typically large, the Poisson approximation with parameter $\lambda = n\pi$ applies.

Python functions

In the library `scipy.stats` many useful functions are implemented.

```
#---------------------------------------
# import libraries
from scipy.stats import binom, poisson
#---------------------------------------

# Binomial distribution Bin(200,0.001)
>>> n, p = 200, 0.001
>>> binom.pmf(1,n,p)
0.16389365955527033
>>> 1-(binom.pmf(0,n,p)+binom.pmf(1,n,p)+binom.pmf(2,n,p))
0.0011337680974732312
>>> 1-binom.cdf(2,n,p)
0.001133768097462684

# Poisson distribution Po(3.5)
```

```
>>> poisson.pmf(6,3.5)
0.07709834987526801
>>> poisson.pmf(5,3.5)
0.13216859978617376
>>> 1-poisson.cdf(5,3.5)
0.14238644690422175
```

3.3 Continuous random variables

So far, we have considered discrete random variables X. We were able to list the possible values that X can take, and the distribution of X was fully characterized by the probabilities $P(X = x)$ for all possible values x.

A random variable X that *all* values in $\mathbb{R}$ (or at least in one interval $[c, d] \subset \mathbb{R}$) is called a **continuous random variable**. The probability distribution can now no longer be specified as a „list“; the probability of *exactly* hitting a point $x \in \mathbb{R}$ is 0:

$$P(X = x) = 0 \quad \text{for all } x \in \mathbb{R}.$$

You can only make statements about the probability with which the value of X lies in a certain subinterval $[a, b] \subseteq \mathbb{R}$, i.e. about probabilities of the form $P(a \leq X \leq b)$.

The distribution of a continuous random variable X can be described by a so-called *density function* f_X. Probabilities can be calculated using the formula $P(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$.

Probability density function (pdf)

a) A function f_X on $\mathbb{R}$ with values in $[0, \infty[$ is called (probability) density function if

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1.$$

Each density function f_X has a distribution function $F_X(x) = P(X \leq x)$, namely

$$F_X(x) := \int_{-\infty}^{x} f_X(t)\mathrm{d}t.$$

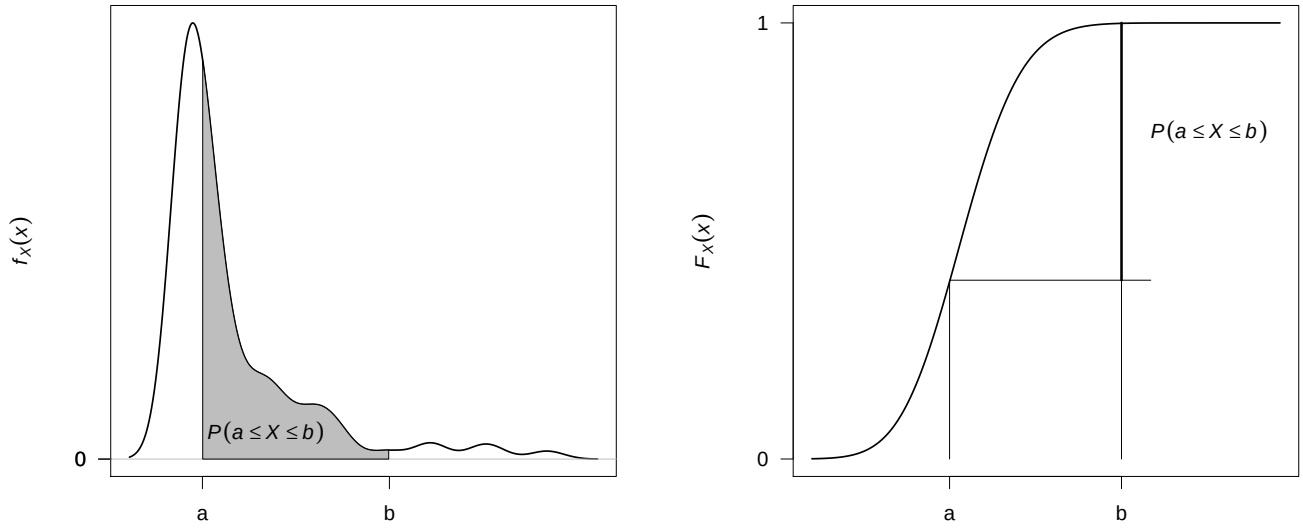
b) A random variable X is distributed according to the density function f_X and the associated distribution function F_X if the following applies for any interval $[a, b]$

$$P(X \in [a, b]) = P(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x = F_X(b) - F_X(a).$$

Remarks

- The distribution function F_X is therefore a special primitive of the density function f_X, namely the one for which $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.
- The interval $[a, b]$ does not necessarily have to be closed. The statements just made apply to any interval, in particular also to open or infinitely large intervals, such as $]a, b[$ or $[a, \infty[$.

The following figure illustrates the relationship between the density function and the distribution function of a continuous random variable.



For random variables with a discrete distribution, we define the expected value $\mathrm{E}(X)$ as the sum $\sum_{i=1}^{n} x_i \cdot \mathrm{P}(X = x_i)$. This does not work here, as all probabilities $\mathrm{P}(X = x_i)$ are equal to zero.

Expected value

The expected value for continuous random variables is defined as follows.

$$\mathrm{E}(X) := \int_{-\infty}^{\infty} x \cdot f_X(x) \mathrm{d}x$$

This seems plausible insofar as $f_X(x) \cdot \mathrm{d}x$ corresponds to the probability that X assumes a value in the infinitesimal interval $[x, x + \mathrm{d}x]$ of length $\mathrm{d}x$.

Variance and standard deviation

Recall that we have defined the variance by the following two equivalent expressions:

$$\mathrm{Var}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \tag{3.1}$$

The variance of a continuous random variable can therefore be calculated by integral as follows

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} \left(x - \mathrm{E}(X)\right)^2 \cdot f_X(x) \mathrm{d}x = \int_{-\infty}^{\infty} x^2 \cdot f_X(x) \mathrm{d}x - \mathrm{E}(X)^2$$

The standard deviation is still defined as

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Median and quantiles

The median of a distribution is the value below (and therefore also above) which exactly half of the total probability lies. More precisely: For the median $q_{0.5}$ of the continuous random variable X the following applies $\mathrm{P}(X \le q_{0.5}) = \mathrm{P}(X \ge q_{0.5}) = 0.5$. Since the first probability can also be written as $F_X(q_{0.5})$ using the distribution function, the median is therefore obtained as $q_{0.5} = (F_X)^{-1}(0.5)$.

Remarks

- For symmetric distributions, the median is obviously equal to the expected value. For asymmetric distributions, however, this is usually not the case.
- The median is also used for discrete distributions, but it is not always clear how to define it precisely: For the discrete uniform distribution on $\{1, 2, 3, 4, 5\}$, it is clear that 3 is the unique median of the distribution. With the fair dice roll, however, i.e. with the discrete uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, both $\mathrm{P}(X \geq 3.5) = \mathrm{P}(X \leq 3.5) = 0.5$ and (e.g.) $\mathrm{P}(X \geq 3.7) = \mathrm{P}(X \leq 3.7) = 0.5$ apply. In this sense, every value between 3 and 4 is a median of the distribution. For reasons of symmetry, 3.5 is by far the most obvious value here, but the same ambiguity also occurs with asymmetric distributions. In practice, however, this ambiguity is rarely relevant.

As a generalization of the median, the β-quantile q_β of a distribution is the value below which a β fraction of the total probability lies. Often β is also given as a percentage here. Analogous to above, $q_\beta = (F_X)^{-1}(\beta)$ applies.

„Median“ is therefore simply another word for „0.5-quantile“ or „50%-quantile“.

3.4 Transformations of random variables

We will occasionally have to deal with transformed random variables, i.e. random variables that we obtain as functions of other random variables. We have already seen discrete examples of this without actively realizing it – for example, the random variable $Y := X^2$ appears in the second calculation formula for the variance. For example, if X describes the number of points of a fair dice roll, it is obvious that $Y = X^2$ is uniformly distributed with values in $\{1, 4, 9, 16, 25, 36\}$.

In order to be able to carry out such transformations for continuous random variables, you normally have to calculate using the distribution functions $F_Y(y)$ and $F_X(x)$. The relationship between these can usually be easily written down (see examples below), and the probability density f_Y is then obtained by deriving F_Y.

Side length of a square

Let the side length X of a square be uniformly distributed by chance in the interval [5,7]. The area $Y = X^2$ is therefore between $5^2 = 25$ and $7^2 = 49$, but what is its exact distribution?

The transformed random variable Y is *not* uniformly distributed; the distribution results instead from the following calculation: For $y \in [25, 49]$,

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X^2 \leq y) = \mathrm{P}(X \leq \sqrt{y}) = F_X(\sqrt{y}).$$

Together with $F_X(x) = \frac{x-5}{2}$ we thus obtain $F_Y(y) = \frac{\sqrt{y}-5}{2}$. We now obtain the probability density by derivation: $f_Y(y) = F_Y'(y) = \frac{1}{4\sqrt{y}}$. In total,

$$f_Y(y) = \begin{cases} \frac{1}{4\sqrt{y}}, & \text{for } y \in [25, 49] \\ 0, & \text{other} \end{cases}$$

(In fact, $\int_{25}^{49} \frac{1}{4\sqrt{y}} \mathrm{d}y = \left[\frac{1}{2}\sqrt{y}\right]_{y=25}^{49} = 1$, as must be the case for a density function!)

Linear transformation

For any two numbers $a > 0$ and $b \in \mathbb{R}$ we consider the transformed random variable

$$Y := a \cdot X + b$$

with associated distribution function $F_Y(y) = \mathrm{P}(Y \le y)$. The relationship between F_X and F_Y is

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(a \cdot X + b \le y) = \mathrm{P}\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

We now obtain the density function of Y by deriving it from y:

$$\begin{aligned} f_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_Y(y) &= \tfrac{\mathrm{d}}{\mathrm{d}y}\left(F_X\left(\frac{y-b}{a}\right)\right) \\ &= F_X'\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} \\ &= f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} \end{aligned}$$

The factor $\frac{1}{a}$ arises here as an inner derivative (chain rule!).

Remark:
The distribution and density function just calculated can also be obtained more directly with graphical considerations alone: To obtain F_Y and f_Y, the function graphs of f_X and F_X must be stretched horizontally by the factor a and shifted to the right by b. This corresponds to inserting $\frac{x-b}{a}$ into these functions. The graph of the density function f_Y must also be compressed by the factor a in the vertical direction so that the area under the curve (= overall probability) is 1 again. So

$$F_Y(x) = F_X\left(\frac{x-b}{a}\right), \qquad f_Y(x) = \frac{1}{a} \cdot f_X\left(\frac{x-b}{a}\right),$$

although instead of the variable x you can of course also write y as above.

Logarithmic transformation

If we consider the logarithmic random variable

$$Y := \ln(X),$$

we can determine the distribution function of Y as follows:

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(\ln(X) \le y) = \mathrm{P}(X \le \mathrm{e}^y) = F_X(\mathrm{e}^y)$$

We obtain the density function analogous to the linear transformation by derivation:

$$\begin{aligned} f_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_Y(y) &= \tfrac{\mathrm{d}}{\mathrm{d}y}\left(F_X(\mathrm{e}^y)\right) \\ &= F_X'(\mathrm{e}^y) \cdot \mathrm{e}^y \\ &= f_X(\mathrm{e}^y) \cdot \mathrm{e}^y. \end{aligned} \tag{3.2}$$

For the concrete case that X is uniformly distributed on $[1, 4]$,

$$f_X(x) = \begin{cases} 0, & \text{for } x < 1 \\ \frac{1}{3}, & \text{for } x \in [1, 4] \\ 0, & \text{for } x > 4 \end{cases}$$

and thus for the transformed variable $Y = \ln(X)$:

$$f_Y(y) = \begin{cases} 0 & \text{for } e^y < 1, \text{ thus } y < 0 \\ \frac{1}{3}e^y & \text{for } e^y \in [1, 4], \text{ thus } y \in [0, \ln(4)] \\ 0, & \text{for } e^y > 4, \text{ therefore } y > \ln(4) \end{cases}$$

3.5 Special continuous distributions

So far, we have talked about continuous random variables in very general terms. As in the discrete case, there are some special distributions that prove to be very helpful in application. We will get to know the most important discrete distributions in this section.

Uniform distribution