Lecture Notes for Math 447 - Probability

Student edition with proofs

Michael Fochler
Department of Mathematics
Binghamton University

Last update: February 20, 2025

Contents

History of Updates:

Date	Topic
2020-12-23	Created.

1 Some Preliminaries

1.1 About This Document

These lecture notes are supporting material to the required text of this Math 447 course on probability theory. This text is [13] Wackerly, D. and Mendenhall, W. and Scheaffer, R.L.: Mathematical Statistics with Applications, 7th edition.

At this point in time (July, 2023) it focuses quite a bit on some of the foundations of probability theory which cannot be found at a sufficient level of generality in that text. Examples are preimages and σ–algebras. It has not been determined at this point in time what further topics will be included at some future time.

Note the uses of the symbol $\boxed{\star}$ for material that will not appear on exams, quizzes and other graded assignments. Unless you see this symbol in a footnote, please note that I will utilize such material and build on it in my lectures. Thus, you should understand this material well enough to follow my lectures, even though you will not be directly tested on it.

Also we use colored boxes according to the following. Generally speaking,

These boxes contain important definitions or parts thereof.

These boxes contain important theorems and propositions or parts thereof.

These boxes contain other kinds of important items that are worth while to know.

There are definitions and theorems that contain two or even three small boxes rather than a big one. There is a technical reason: such boxes do not span pages and will needlessly inflate the page count of the document.

1.2 A First Look at Probability

"All models are wrong, but some are useful".

Attributed to the statistician George E. P. Box (1919–2013)

This quote certainly applies to the probabilistic models and the role they play in answering statistical questions such as

- How do I collect data to predict next month's average unemployment rate?
- What is the risk that this prediction will be off by more than 0.5 percent?

You probably agree that we also could have formulated the second question as follows.

- What is the probability that this prediction will be off by more than 0.5 percent?

It is not easy to find a satisfactory answer to that question and it will depend on the assumptions that go into your model. We will consider probability in much simpler settings.

Example 1.1 (Empirical probability)**.** The concept of probability serves as a model for quantifying how likely an event will happen that depends on chance. When we say that the probability of obtaining an even number when rolling a die equals 0.5, then we mean the following.

Assume that

- X_1 denotes the action of rolling that die for the first time.
- X_2 denotes the action of rolling that die for the second time.
- $\ldots X_k$ denotes the action of rolling that die for the kth time.

Under those assumptions we expect the following:

In the long run (for large k), close to half of $X_1, X_2, \ldots, X_k$ result in an even outcome.

- In the long run (for large k), close to half of $X_1, X_2, \ldots, X_k$ should result in an even outcome.

We formulate this in the language of mathematics as follows:

- We write P for probability.
- We write $\{2,4,6\}$ for the event that rolling the die results in a 2 or a 4 or a 6, i.e., in an even outcome. So we write this event as a set that contains the outcomes $2, 4$,and 6 as its elements.
- We write n_k for the number of outcomes during those k rolls that result in a 2 or a 4 or a 6. We define $P\{2,4,6\} = \lim_{k\to\infty} \frac{n_k}{k}$ and call this limit the probability of the event $\{2,4,6\}$. [1]

We expect this particular limit to be 0.5.

- We write Ω (the Greek capital letter Omega) [2] for the set of all potential outcomes. It is customary to drop the word "potential" and refer to the elements of Ω simply as **outcomes**.
- We call the subsets of Ω **events**. Thus, an event A is a set A that satisfies $A \subseteq \Omega$, [3] i.e., each element of A also belongs to Ω, i.e., A is a collection of outcomes.
- It is expedient to also call the empty set $\emptyset$ (the set that contains no elements) an event.

For the roll of a die the list of all outcomes is $1, 2, \ldots, 6$. Thus, $\Omega = \{1,2,3,4,5,6\}$. An event is any set that consists of zero or more integers between 1 and 6.

We can apply the steps we used to determine $P\{2,4,6\}$ to ANY event $A \subseteq \Omega$. Now, n_k denotes the number of outcomes during the first k rolls that result in a number that is listed in A. We define

$$P(A) = \lim_{k\to\infty} \frac{n_k}{k}\,. \tag{1.1}$$

To be precise, this formula denotes the **empirical probability** of the event A.

Observe that the assignment $A \mapsto P(A)$ of (1.1) satisfies the following for all subsets A of Ω:

- $0 \le P(A) \le 1$.
- $P(\emptyset) = 0$, since $n_k = 0$ for all k. (Recall that $\emptyset$ is empty set which cont ains no elements.)
- $P(\Omega) = 1$, since $n_k = k$ for all k.
- If the subsets A, B of Ω have no elements in common (we speak of **mutually disjoint** sets), then the union $P(A \cup B)$ satisfies

$$P(A \cup B) = P(A) + P(B)\,. \tag{1.2}$$

To see the validity of (1.2), let $n_k(A)$ be the number of times an outcome in A is observed during k trials, and let and $n_k(B)$ be defined likewise for B. Since an outcome ω is in $A \cup B$ if and only if ω either belongs to A or to B, we have $n_k(A \cup B) = n_k(A) + n_k(B)$, hence,

$$P(A \cup B) = \lim_{k\to\infty} \frac{n_k(A \cup B)}{k} = \lim_{k\to\infty} \frac{n_k(A)}{k} + \lim_{k\to\infty} \frac{n_k(B)}{k} = P(A) + P(B)\,.$$

Note the following about the nature of the formula $P(A) = \lim_{k\to\infty} \frac{n_k}{k}$ for subsets A of Ω.

[1] In general we write $P(A)$ or $P[A]$ for the probability of an event A. Accordingly, we could also have written $P(\{2,4,6\})$. However, if the event is of the form $\{\ldots\}$, we are permitted to omit the parenteses/square brackets, since they obscure readability.

[2] For a list of all Greek letters see Section 15.1 (Greek Letters) on page 342.

[3] See Definition 2.3 (Subsets and supersets) on p.19 on page 342.

- ⊡ It is a function $A \mapsto P(A) = \lim_{k\to\infty} \frac{n_k}{k}$ the same way $x \mapsto f(x) = x^2 + 4$ is a function.
- ⊡ We are familiar with the latter: It assigns to each argument x (which happens to be a real number) the function value $f(x)$, also a real number. For example, $f(3) = 3^2 + 4 = 13$.
- ⊡ The function $A \mapsto P(A)$ is harder to deal with only because its arguments A are not numbers or vectors of such numbers. Rather, those arguments are events, i.e., sets. □

You are strongly encouraged to take a first look at Section 2.4 (Functions and Sequences). It is **very important that you understand** the following:

- The assignment $A \mapsto P(A)$ discussed at the end of Example 1.1 constitutes a function

$$P : \{ \text{ all subsets of } \Omega \} \longrightarrow [0,1] \qquad ([0,1] = \{ \text{ numbers } x : 0 \le x \le 1 \})$$

in the sense of Definition 2.17 on p.33, with domain = { all subsets of Ω } and codomain = $[0,1]$.

Remark 1.1. There are some issues with (1.1) as a definition of $P(A)$.

What if the limit $\lim_{k\to\infty} n_k/k$ does not exist? For example, the following is very unlikely but not impossible.

Let ω_k denote the outcome of the kth roll of the die. Assume that we obtain the following sequence of outcomes (draw a picture!):

- $\omega_1 = 1$.
- From now on, only the number 6 appears until $n_k/k > 5$. We write $K(1)$ for that index k.
- From now on, only the number 1 appears until $n_k/k < 2$. We write $K(2)$ for that index k.
- From now on, only the number 6 appears until $n_k/k > 5$. We write $K(3)$ for that index k.
- From now on, only the number 1 appears until $n_k/k < 2$. We write $K(4)$ for that index k.
- ..... and so on .....

The resulting sequence $K(1) < K(2) < K(3) < \cdots$ satisfies the following: [4]

- There are infinitely many indices $k = K(1), K(3), K(5), \ldots$ such that $\frac{n_k}{k} > 5$.
- There are infinitely many indices $k = K(2), K(4), K(6), \ldots$ such that $\frac{n_k}{k} < 2$.

Accordingly, $\lim_{k\to\infty} \frac{n_k}{k}$ does not exist, and we are not able to determine $P(A)$.

But there are issues even if that limit exists. Consider again the event $A = \{2, 4, 6\}$. Let us assume that, by some freak of nature, all outcomes ω_k are 4. [5] Accordingly, we declare that $P\{2, 4, 6\} = 1$. The teamleader has doubts about this result and asks for a repetition of the experiment. This time all outcomes ω_k are either 3 or 5.

What to do? Should we decide that $P\{2, 4, 6\} = 0$? Should the experiment repeated once more? How about settling on the average, $P\{2, 4, 6\} = (1 + 0)/2 = 1/2$?

You may decide that this is a completely ficticious example without any bearing on reality, and this author agrees. That being said, consider the following:

[4] A strict proof can be obtained by using the fact that the limit of a sequence does not depend its first k members, no matter how big k may be chosen.

[5] You will learn the following: If each $j_1, j_2, \ldots$ is a given potential outcome (an integer between 1 and 6), then $P\{\omega_1\} = j_1, P\{\omega_2\} = j_2, \ldots, P\{\omega_k\} = j_k\} = (1/6)^k$. That number becomes very small for large k, since the sequence $(1/6)^k$ converges to zero. Nevertheless, $(1/6)^k > 0$ for each fixed k, so it is not impossible to obtain $\omega_k = 4$ for all k. (This is the case where $j_1 = j_2 = \cdots = j_k = 4$ for all k).

- The infinite repetition of an action such as rolling a die is in itself an abstraction that serves to model reality, and so is the limit of a (infinite) sequence.
- In the real world the determination of probabilities $P(A)$ often is based on (1.1) as follows: It is decided to conduct an experiment of k trials. The larger this number k is chosen, the more confidence we will have that $P(A)$ is a good enough APPROXIMATION of the likelihood that the event A happens.

Unfortunately there are factors to consider that will limit the size of k.

- The more repetitions, the longer it will take to obtain the result. If A is the event that the Old Faithful geyser in Yellowstone National Park erupts to a height of at least 150 feet and it is not possible for some reason to use the previously obtained records, then we must base the determination of $P(A)$ on a very small number of observations.
- Money is another limiting factor. The more repetitions, the more it will cost to obtain the result. □

Example 1.2 (Single roll of a die)**.** To avoid the issues concerning the use of formula (1.1) (empirical probability) on p.7, we also could have employed a model from physics or geometry, that of a fair die. A fair die is a model of reality obtained from geometry or physics. Such a die is assumed to be perfectly symmetrical and this symmetry implies that each of the outcomes $1, 2, \ldots, 6$ is equally likely. Consequently, each outcome must have the same likelihood (probability) of $1/6$.

We consider again the probability of rolling an even number The even outcomes are $2, 4, 6$. Thus,

$$P\{\text{ even outcome }\} = P\{2,4,6\} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5\,. \tag{1.3}$$

Note that fair dice do not exist in the real world. Matter of fact, if we had a sample of $1,000$ dice and we were able to determine with infinite precision the probability that a throw of die $\#_k$ comes up even, chances are that we would obtain several different answers, due to imperfections in the manufacturing process. However, chances are that we work in an environment where the error we commit when assuming that the die is fair does not matter, so let us make that assumption.

We model the **random action** of rolling such a fair die just once as follows.

- As in Example 1.1 (Empirical probability) on p.6, the set Ω of all (potential) outcomes is $\{1, 2, 3, 4, 5, 6\}$.
- We associate with each outcome $\omega \in \Omega$ the probability $P(\{\omega\}) = 1/6$.
- For each outcome $\omega \in \Omega$ there is a corresponding event $\{\omega\} \subseteq \Omega$. [6] It is a common abuse of language to also refer to such "atomar" events as outcomes.
- We generalize (1.3) and associate with each event $A \subseteq \Omega$ the probability

$$P(A) = \sum_{\omega \in A} P(\{\omega\})\,. \tag{1.4}$$

  Here, $\sum_{\omega \in A} P(\{\omega\})$ means that we sum up all those expressions $P(\{\omega\})$ that satisfy $\omega \in A$.
- For example, let $A = \{2, 4, 6\}$ and $B = \{\omega \in \Omega : \omega > 4\}$. Thus. A is the event of rolling an even outcome and B is that of rolling a 5 or 6. Then,

[6] Such sets of size 1 are often called **singleton sets** or simply **singletons**.

$$P(A) = \sum_{\omega\in A} P(\{\omega\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$
$$P\big(B\big) = P\big(\{5,6\}\big) = P(\{5\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

It is customary to write $P\{...\}$ for $P(\{...\})$. Thus, the last equation can also be written as

$$P\big(B\big) = P\{5,6\} = P\{5\} + P\{6\} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

- The assignment $A \mapsto P(A)$ satisfies for $A \subseteq \Omega$ the following:

$\boxdot\ 0 \le P(A) \le 1$ $\boxdot\ P(\emptyset) = 0$ $\boxdot\ P(\Omega) = 1$ $\boxdot\ P(A \cup B) = P(A) + P(B)$, ($A, B$ disjoint)

Note that $A \mapsto P(A)$ of Example 1.1 (Empirical probability) obeys the same rules. □

Example 1.3 (Two rolls of a die)**.** Consider what happens when two fair dice are rolled or, equivalently, when one fair die is rolled twice in a row. The set of outcomes is

$$\Omega = \{1,2,\dots,6\}^2 = \{1,2,\dots,6\} \times \{1,2,\dots,6\} = \{\omega : \omega = (i,j) \text{ and } i,j = 1,2,\dots,6\}.$$

- We make a willful decision to consider the outcomes (i,j) and (j,i) different for $i \neq j$. For example, if die #1 is red and #2 is white, we distinguish between the outcome of a red 2 and a white 5 and that of a red 5 and a white 2. Then Ω consists of 36 outcomes

$$(1,1),\ (1,2),\dots,(1,6),\ (2,1),\ (2,2),\dots,(2,6),\dots,(6,1),\ (6,2),\dots,(6,6)$$

  and symmetry considerations show that each outcome $\omega \in \Omega$ has probability $P\{\omega\} = 1/36$.

- We are faced with the same situation as in Example 1.2. The probabilities $P\{\omega\}$ of the outcomes determine the probability of any event $A \in \Omega$ just as we saw in (1.4):

$$P(A) = \sum_{\omega\in A} P(\{\omega\}). \tag{1.5}$$

- For example, if $A = \{$ die #1 shows a $4\} = \{(4,j) : j = 1,2,\dots,6\}$ then

$$P(A) = \sum_{\omega\in A} P(\{\omega\}) = \sum_{(i,j)\in A} P(\{(i,j)\})$$
$$= P\{(4,1)\} + P\{(4,2)\} + \cdots + P\{(4,6)\} = 6\left(\frac{1}{36}\right) = \frac{1}{6}.$$

- As in examples 1.1 (Empirical probability) and 1.2 (Single roll of a die), there is again a assignment $A \mapsto P(A)$ of probabilities that satisfies the familiar rules

$\boxdot\ 0 \le P(A) \le 1$ $\boxdot\ P(\emptyset) = 0$ $\boxdot\ P(\Omega) = 1$ $\boxdot\ P(A \cup B) = P(A) + P(B)$, ($A, B$ disjoint) □

Example 1.4 (Sum of two die rolls)**.** Consider what happens if two fair dice are rolled and we are interested in the sum of points obtained that way. For example,

- the outcome 8 is obtained when either of the following are rolled:
  ⊡ a 2 and a 6 ⊡ a 3 and a 5 ⊡ a 4 and a 4 ⊡ a 5 and a 3 ⊡ a 6 and a 2.
- the outcome 5 is obtained when either of the following are rolled:
  ⊡ a 1 and a 4 ⊡ a 2 and a 3 ⊡ a 3 and a 2. ⊡ a 4 and a 1.
- The set of outcomes is

$$\Omega \;=\; \{2, 3, \ldots, 11, 12\}\,.$$

Since a roll of two dice has 36 outcomes $(1,1), \ldots, (6,6)$ and each of those has probability $1/36$ (see Example 1.3), it follows for the outcomes 8 and 5 that

- $$P(\{8\}) \;=\; \frac{5}{36}\,; \qquad P(\{5\}) \;=\; \frac{4}{36}\,.$$

Here is the complete list of outcome probabilities $P(\{\omega\})$:

$$\begin{aligned} P(\{2\}) = P(\{12\}) &= \frac{1}{36}\,;\;\; P(\{3\}) \;=\; P(\{11\}) \;=\; \frac{2}{36}\,;\;\; P(\{4\}) \;=\; P(\{10\}) \;=\; \frac{3}{36}\,; \\ P(\{5\}) = P(\{9\}) &= \frac{4}{36}\,;\;\; P(\{6\}) \;=\; P(\{8\}) \;=\; \frac{5}{36}\,;\;\; P(\{7\}) \;=\; \frac{6}{36}\,. \end{aligned} \tag{1.6}$$

- In the previous two examples there was **equiprobability**: Each outcome had the same probability. Clearly, there is no equiprobability for the sum of points obtained when rolling two dice.
- Nevertheless, the probability of any event $A \in \Omega$ is obtained again by the formula

$$P(A) \;=\; \sum_{\omega \in A} P(\{\omega\})\,. \tag{1.7}$$

- For example, if $A = \{$ the sum is between 8 and 11$\}$, then

$$\begin{aligned} P(A) \;&=\; \sum_{\omega \in A} P(\{\omega\}) \;=\; \sum_{\omega=8}^{11} P(\{\omega\}) \\ &= P\{8\} \,+\, P\{9\} \,+\, P\{10\} \,+\, P\{11\} \;=\; (5+4+3+2)\left(\frac{1}{36}\right) \;=\; \frac{7}{18}\,. \end{aligned}$$

- As in examples 1.1 (Empirical probability) and 1.2 (Single roll of a die), there is again a assignment $A \mapsto P(A)$ of probabilities that satisfies the familiar rules

⊡ $0 \le P(A) \le 1$ ⊡ $P(\emptyset) = 0$ ⊡ $P(\Omega) = 1$ ⊡ $P(A \cup B) \;=\; P(A) + P(B)$, ($A, B$ disjoint) □

Let us examine what the examples we have studied so far have in common.

Remark 1.2. In the examples given so far a probability $P(A)$ was assigned to each event $A \subseteq \Omega$. In each case thhis assignment $A \mapsto P(A)$ satisfies the following.

(1.8) $0 \leq P(A) \leq 1$.

(1.9) $P(\emptyset) = 0$. Here $\emptyset$ is the empty set which contains no outcomes.

(1.10) $P(\Omega) = 1$. Here Ω is the set which contains all potential outcomes.

If the events A, B have no outcomes in common, the union $A \cup B$ satisfies

(1.11) $$P(A \cup B) = P(A) + P(B).$$

- The probabilist likes to speak of the **probability space** Ω, since it comes with a **probability measure** (WMS: probability function), $A \mapsto P(A)$, which assigns to the events A of Ω, the probability $P(A)$ that this event might "occur" or "happen".
- Statisticians tend to call Ω a **sample space**. An element ω of Ω still is referred to as an **outcome** but some, like WMS, write S instead of Ω (that's S as in **S**ample). They also call an element s of S a **sample point** of S.

We translate some of the examples already encountered into the language of sample spaces and sample points.

- In example 1.2 (Single roll of a die) on p.9, $S = \{1, 2, \ldots, 6\}$ is the sample space. Its outcomes or sample points are $1, 2, 3, 4, 5, 6$. They can be considered sample of size $n = 1$. Further, all events that result from the single roll of a die are formed from those sample points.
- In example 1.3 (Two rolls of a die) on p.10, the sample points $(1, 1)$, $(1, 2)$, $\ldots$, $(6, 5)$, $(6, 6)$ constitute the sample space $S = \{1, 2, \ldots, 6\}^2$.
- In example 1.4 (Sum of two die rolls) on p.11, $S = \{2, 3, \ldots, 12\}$ is the sample space. The sample points from which all relevant events are formed, are the numbers $2, 3, \ldots, 12$. □

Example 1.5. This example needs more computational skills than the ones we have encountered so far.

- To understand whether a traffic light works as expected, the following experiment is conducted. 200 cars are observed and a record is made for each one of those cars whether it reached the intersection on red, green or yellow.
- This "**sampling action**" of observing those 200 cars results in ONE sample point of size 200. Its actual outcome depends on chance
- Once the experiment is completed, the result will be a **realization** of this sampling action (the SPECIFIC sample point that was obtained). If we write r for red, g for green, y for yellow, this realization might be, e.g., $\{r, r, y, g, g, g, r, y, \ldots, r\}$.
- Once that realization has been obtained, the sampling action has lost its random character.
- It is customary among statisticians to use the term **sample** for both the process of obtaining a sample (the sampling action) and a realization of this action. We will in general follow this convention.
- The sample space S of all (potential) sample points for this experiment is huge: It contains 3^{200} sample points. This will be discussed in Chapter 7 (Combinatorial Analysis)
- Each event $A \subseteq S$ comes with a probability $P(A)$ and one can show that the assignment $A \mapsto P(A)$ satisfies the formulas (1.8) – (1.11) of Remark 1.2 on p.12. □

Here is a formal definition of probability. It is based on the formulas (1.8) – (1.11) of Remark 1.2 on p.12. **This definition is PRELIMINARY and will be amended!**

This definition uses the concept of an abstract function. Such functions, which assign the arguments of an arbitrary set X (the domain) to the elements (the function values) of another arbitrary set Y (the codomain) are discussed in Section 2.1 (Sets, Numbers, Sequences and Functions) on page 18. We suggest that you look at it **now!**

Definition 1.1 (Probability measure - Preliminary Definition, version I)**.**

A **probability measure** P on a set Ω is a function which assigns to each subset A of Ω a real number $P(A)$ between 0 and 1 as follows.

(a) $P(\emptyset) = 0$ and $P(\Omega) = 1$. Here $\emptyset$ denotes the empty set which contains no elements.

(b) If the subsets A, B of Ω have no elements in common, then probability is **additive**:

$$P(A \cup B) \;=\; P(A) \,+\, P(B)\,.$$

This last formula makes disjoint unions so important that we have reserved the special symbol "$\uplus$" as a visual aid. Henceforth, we usually write $U \uplus V$ for $U \cup V$ if we know that $U \cap V = \emptyset$:

$$P(A \uplus B) \;=\; P(A) \,+\, P(B)\,. \;\square$$

Remark 1.3. The additivity condition also holds for three disjoint sets $A, B, C \in 2^\Omega$ since,

$$P\,(\,A \uplus B \uplus C\,) \;= P\,[\,(A \uplus B) \uplus C\,] \;=\; P\,(\,A \uplus B) + P(C) \;=\; P(A) + P(B) + P(C)\,.$$

From this equation one obtains additivity for four disjoint $A, B, C, D \in 2^\Omega$ as follows:

$$\begin{aligned} P\,(\,A \uplus B \uplus C \uplus D) \;&= P\,[\,(A \uplus B \uplus C) \uplus D\,] \\ &= P\,(\,A \uplus B \uplus C) + P(D) \;=\; P[A] + P[B] + P[C] + P[D]\,. \end{aligned}$$

In a similar fashion one obtains additivity for five, then for six, ..., for any finite number of disjoint subsets $A_1, \ldots, A_n$ of Ω. However, we want more than

$$\textbf{additivity:} \qquad P\,(A_1 \uplus A_2 \uplus \cdots \uplus A_n) \;=\; P(A_1) + P(A_2) + \cdots + P(A_n)$$

for only any finite number n of events, since it has proven extremely fruitful to extend additivity to infinite sequences of disjoint events and replace it with

$$\sigma\textbf{–additivity:}^{7} \qquad P\,(A_1 \uplus A_2 \uplus A_3 \uplus \cdots) \;=\; P(A_1) + P(A_2) + P(A_3) + \cdots \;\square$$

Definition 1.2 (Probability measure - Preliminary Definition, version II)**.**

A **probability measure** P on a set Ω is a function which assigns to each subset A of Ω a real number $P(A)$ between 0 and 1 as follows.

(a) $P(\emptyset) = 0$ and $P(\Omega) = 1$.

(b) If the subsets $A_1, A_2 \ldots$ of Ω are mutually disjoint, then probability is **σ–additive**:

$$P(A_1 \uplus A_2 \uplus \cdots) \;=\; P(A_1) \,+\, P(A_2) \,+\cdots \;=\; \sum_{j=1}^{\infty} P(A_j)\,. \tag{1.12}$$

[7] σ ("sigma") is a greek letter. See the appendices for a complete list.

We include the following informal definition from earlier parts of this section into this definition:

- The combination (Ω, P) is called a **probability space** aka **sample space**.
- An element ω of Ω is called an **outcome** aka **sample point**
- A subset of Ω is called an **event**. □

Remark 1.4. Generally speaking, adding requirements to a model restricts the scenarios for which the model is useful. So what are the disadvantages of replacing additivity for probability measures with σ–additivity? The consensus is that there are none to be concerned about. [8] On the other hand, σ–additivity greatly enriches the tool kit for solving problems in the area of probability and statistics and their real–world applications. □

Remark 1.5.

- Note that Definition 1.2 makes no mention about how one should interpret the number $P(A)$. It may or may not reflect what happens in the real world!

For example, one could take a fair coin and define $P\{H\} := 0.1$. Here, $H =$ Heads and $T =$ Tails. This uniquely defines a probability measure $A \mapsto P(A)$ on the sample space $S := \{H, T\}$, since the missing probabilities for the events $\emptyset, \{H\}, S$ can be determined as follows:

$$P(\emptyset) = 0 \quad \text{and} \quad P(S) = 1\,, \qquad \text{by Definition 1.1\textbf{(a)}}\,.$$
$$P\{T\} + P\{H\} = P(\{T\} \uplus \{H\}) = P(S) \;\Rightarrow\; P\{T\} = 1 - 0.1 = 0.9\,,$$
$$\text{by Definition 1.1\textbf{(b)}, since } S \text{ is the disjoint union of } \{H\} \text{ and } \{T\}.\ \square$$

Remark 1.6. If the last example strikes you as nonsensical, here is a model used by Wall Street that uses a probability measure in which the probability of an event is different from the chance that this event will happen.

The so called binomial asset model is a probabilistic model to determine today's price of a stock option which will be exercised at some future point in time. [9] In this model, trading of a specific stock (e.g., IBM or Amazon), happens at times $0, 1, 2, \ldots$. There are only two possible ways that stock price can change and there are two "real world" probabilities, one for each possibility:

- $p_u := P\{$ the price of a share of stock changes by the factor u.
- $p_d := P\{$ the price of a share of stock changes by the factor $d < u = 1 - p_u$.

These two numbers p_u and p_d are sufficient to determine a probability space Ω and probability measure P for trading in that stock.

[8] It would be more accurate to say that there are no issues as far as building models of reality is concerned. We will discuss at length in Chapter 5 (The Probability Model) that there is a cost: One may not be able to assign a probability $P(A)$ to all subsets A of Ω. Rather, one must require $A \in \mathfrak{F}$, where

$$\mathfrak{F} \subseteq \{\text{ all subsets of } \Omega\ \}\,.$$

So $\mathfrak{F}$ is a set which contains sets as its elements(!) However, $\mathfrak{F}$ can be chosen so big that it includes all sets that matter for the applications of probability and statistics.

[9] Since this is not a course on probabilistic finance, we must refer you to the literature for details. Some references are [10] Shreve, Steve: Stochastic Calculus for Finance I: The Binomial Asset Pricing Model, [2] Björk, Thomas: Arbitrage Theory in Continuous Time and this author's Math 454 lecture notes (Spring 2023).

Strangely enough, p_u and p_d are replaced by the so-called risk–neutral probabilities $\tilde{p}_u$ and $\tilde{p}_d$, which are sufficient to determine an altgernate probability measure $\widetilde{P}$ on that same probability space Ω.

Even stranger, the real world probability measure P has no bearing on the determination of $\widetilde{P}$, i.e., of $\tilde{p}_u$ and $\tilde{p}_d$. [10] And yet, even though $\tilde{p}_u$ and $\tilde{p}_d$ do not reflect the actual probabilities that govern the stock price, they are used to set today's price of an option on that stock that can be redeemed only, say, 90 days from today. □

Next, we combine Example 1.3 and Example 1.4.

Example 1.6. When computing the outcome probabilities of the sum of points obtained by rolling two dice, we argued with a result obtained in Example 1.3. There, the probability of an outcome (i, j) was $1/36$ for all $i, j = 1, 2, \ldots, 6$. It should not be surprising that there is a connection between the probability models of those examples. Both had a set of outcomes which we denoted Ω and a function $P : A \mapsto P(A)$ which associated a probability $P(A)$ with each event $A \subseteq \Omega$. Since this example deals with both outcome sets and both probability measures, we must change our notation. We proceed as follows.

- We keep the notation (Ω, P) for the probability space of Example 1.3 and define

$$\begin{aligned} \Omega &:= \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} = \{\omega : \omega = (i, j) \text{ and } i, j = 1, 2, \ldots, 6\}, \\ P\{(i, j)\} &:= \frac{1}{36} \qquad \text{for } i, j = 1, 2, \ldots, 6. \end{aligned}$$

- For the outcome set and probability measure of Example 1.4, we write

$$\begin{aligned} \Omega' &:= \{2, 3, \ldots, 11, 12\}, \\ P'\{2\} &:= P'\{12\} := \frac{1}{36},\ P'\{3\} := P'\{11\} := \frac{2}{36}, \ldots \end{aligned} \tag{1.13}$$

See (1.6) on p.11.

Note that $P'\{k\}$ equals the probability that the sum of the two die rolls equals k, since the first probability is given by (1.13), the second by (1.6) on p.11, [11] and both formulas match.

Let $(i, j) \in \Omega$, i.e., i is the outcome of rolling die #1 and j is that of rolling die #2. The assignment

$$(i, j) \mapsto Y(i, j) := i + j$$

associates with this outcome an integer between 2 and 12, i.e., an outcome in Ω'. Think of Y as a function which assigns to each argument $(i, j) \in \Omega$ the function value $Y(i, j) = i + j \in \Omega'$. [12]

We assign to each $B \subseteq \Omega'$ the probability

$$P_Y(B) := P\{(i, j) \in \Omega : i + j \in B\}. \tag{1.14}$$

- Observe that $P_Y(B)$ has been defined by means of the probability measure P (not P'), defined on Ω (not on Ω')

Since $i + j = Y(i, j)$, (1.14) can also be written the following two ways:

$$P_Y(B) = P\{(i, j) \in \Omega : Y(i, j) \in B\} = P\{\omega \in \Omega : Y(\omega) \in B\}. \tag{1.15}$$

[10] Rather, the interest earned by depositing money in a bank plays a major role.

[11] After all, $P\{k\}$ of (1.6) was determined by computing the likelyhood that the sum of two rolls equals k.

[12] Looking ahead, Definition 2.17 on p.33 will refer to Y as a function $Y : \Omega \longrightarrow \Omega'$.

We spend most of the remainder of this example to prove that

$$P_Y(B) \;=\; P'(B), \quad \text{for all} \quad B \subseteq \Omega'\,. \tag{1.16}$$

Step 1: We show (1.16) for singletons B of Ω': We assume $B = \{k\}$ for some $k \in \Omega'$. Let

$$A_k \;:=\; \big\{\,(i,j) \in \Omega \,:\, Y(i,j) = k\,\big\} \;=\; \big\{\,(i,j) \in \Omega \,:\, Y(i,j) \in \{k\}\,\big\}\,. \tag{1.17}$$

Then,

$$P(A_k) \overset{(1.17)}{=} P\big\{\,(i,j) \in \Omega \,:\, Y(i,j) \in \{k\}\,\big\} \overset{(1.15)}{=} P_Y\{k\}\,. \tag{1.18}$$

If we can show that $P'\{k\} = P(A_k)$, then (1.18) yields (1.16) for $B = \{k\}$. We see this as follows.

$$\begin{aligned} P'\{k\} \;&= \text{probability that sum of points of both rolls equals } k \\ &= \left(\frac{1}{36}\right) \times \text{(the number of elements in } A_k) \\ &= \sum_{\omega \in A_k} \frac{1}{36} \;=\; \sum_{(i,j)\in A_k} P\{(i,j)\} \;=\; P(A_k)\,. \end{aligned}$$

To summarize, we have shown that

$$P_Y\{k\} \;=\; P'\{k\}\,, \quad \text{for all} \quad k \in \Omega'\,. \tag{1.19}$$

Step 2: We extend (1.16) to arbitrary events of Ω'.
We start with the observation that any set B is the disjoint union $\biguplus_{b\in B} \{b\}$ of the singletons $\{b\}$ such that $b \in B$. For example, the set $\{2,4,6,8,10,12\}$ of the even members of Ω' can be written as

$$\{2,4,6,8,10,12\} \;=\; \{2\} \uplus \{4\} \uplus \{6\} \uplus \{8\} \uplus \{10\} \uplus \{12\}\,.$$

Let $B \subseteq \Omega'$. For brevity, we write $\{Y \in B\}$ for the set of all $\omega \in \Omega$ such that $Y(\omega) \in B$:

$$\{Y \in B\} \;:=\; \{\omega \in \Omega : Y(\omega) \in B\} \;=\; \{(i,j) \in \Omega : i+j \in B\}\,. \tag{1.20}$$

Even simpler, for singleton sets $B = \{k\}$ where $k \in \Omega'$, we write $\{Y = k\}$ for $\big\{\,Y \in \{k\}\,\big\}$:

$$\{Y = k\} \;:=\; \big\{\,Y \in \{k\}\,\big\} \;=\; \{\omega \in \Omega : Y(\omega) = k\} \;=\; \{(i,j) \in \Omega : i+j = k\}\,. \tag{1.21}$$

We suggest that you examine (1.17) and verify that $A_k = \{Y \in \{k\}\} = \{Y = k\}$.
Since $\Omega' = \{2,3,\ldots,12\}$ only contains 11 numbers and $B \subseteq \Omega'$, there is $n \le 11$ such that

$$B \;=\; \{k_1, k_2, \ldots k_n\} \qquad \text{and thus,} \qquad B \;=\; \{k_1\} \uplus \{k_2\} \uplus \cdots \uplus \{k_n\}\,. \tag{1.22}$$

You will learn in section 2.5 (Preimages) that $\{Y \in B\}$ is called the preimage of the set B under the function $Y : (i,j) \to i+j$ (Definition 2.27 on p.42), and that the preimage of a union (disjoint union) is the union (disjoint union) of the preimages. In particular,

$$\{Y \in B\} \overset{(1.20)}{=} \{Y \in \{k_1\} \uplus \{k_1\} \uplus \cdots \uplus \{k_n\}\} \overset{(1.21)}{=} \{Y = b_1\} \uplus \{Y = b_2\} \uplus \cdots \uplus \{Y = b_n\} \tag{1.23}$$

We apply the probability measure P to both sides of (1.23) [13] and apply (σ–)additivity of the probability measure P.

$$P_Y(B) \overset{(1.14)}{=} P\{Y \in B\} \overset{(1.23)}{=} P\big(\{Y \in \{k_1\} \uplus \{k_1\} \uplus \cdots \uplus \{k_n\}\}\big) = \sum_{j=1}^{n} P\{Y \in \{b_j\}\} \tag{1.24}$$

In likewise manner, we apply (σ–)additivity of the probability measure P'.

$$P'(B) \overset{(1.22)}{=} P'\big(\{k_1\} \uplus \{k_2\} \uplus \cdots \uplus \{k_n\}\big) = \sum_{j=1}^{n} P'\big(\{b_j\}\big)$$

Since $P_Y\{k\} = P'\{k\}$ by (1.19), (1.24) and (1.24) have matching right–hand sides. This shows that (1.16) is valid for general subsets $B \subseteq \Omega'$ and concludes **Step 2**. □

Remark 1.7. Example 1.6 is important because is illustrates a very general way of constructing probability measures from existing ones.

(1) Let (Ω, P) be any kind of probability space rather than $\Omega = \{1, \ldots, 6\}^2$ with equiprobability $P\{(i,j)\} = 1/36$.
(2) Let Ω' be any kind of nonempty set, not necessarily $\Omega' = \{2, \ldots, 12\}$.
(3) Let Y be any function $\omega \mapsto Y(\omega)$, which assigns arguments $\omega \in \Omega$ to function values $Y(\omega) \in \Omega'$, not necessarily $Y(i,j) = i + j$.

Then the formula which corresponds to (1.14) of Example 1.6:

$$P_Y(B) := P\{Y \in B\}, \quad \text{i.e.,} \quad P_Y(B) = P\{\omega \in \Omega : Y(\omega) \in B\}, \quad \text{for } B \subseteq \Omega', \tag{1.25}$$

"transports" the probability measure P on Ω to a probability measure P_Y on Ω'. Later we will call such a function Y that assigns elements of (Ω, P) to elements of Ω', a random element. Moreover, we will refer to the probability measure P_Y on Ω', given by (1.25) as the distribution of Y. □

[13] That's P and NOT P' or P_Y: All sets of (1.23) are subsets of Ω NOT of Ω'!

2 Sets, Numbers, Sequences and Functions

Introduction 2.1. □

The student should read this chapter carefully, with the expectation that it contains material that they are not familiar with, as much of it will be used in lecture without comment. Very likely candidates are power sets, a function $f : X \to Y$ where domain X and codomain Y are part of the definition.

2.1 Sets – The Basics

An entire book can be filled with a mathematically precise theory of sets. For our purposes the following "naive" definition suffices:

Definition 2.1 (Sets)**.**

- A **set** is a collection of stuff called **members** or **elements** which satisfies the following rules: The order in which you write the elements does not matter and if you list an element two or more times then **it only counts once**.
- We write $x_1 \in X$ to denote that an item x_1 is an element of the set X and $x_2 \notin X$ to denote that an item x_2 is not an element of the set X.
- Occasionally we are less formal and write x_1 **in** X for $x_1 \in X$ and x_2 **not in** X for $x_2 \notin X$.

We write a set by enclosing within curly braces the elements of the set. This can be done by listing all those elements or giving instructions that describe those elements. For example, to denote by X the set of all integer numbers between 18 and 24 we can write either of the following:

$$X := \{18, 19, 20, 21, 22, 23, 24\} \quad \text{or} \quad X := \{n : n \text{ is an integer and } 18 \leq n \leq 24\}$$

Both formulas clearly define the same collection of all integers between 18 and 24. On the left the elements of X are given by a complete list, on the right **setbuilder notation**, i.e., instructions that specify what belongs to the set, is used instead.

For the above example we have $20 \in X$, $27 - 6 \in X$, $38 \notin X$, 'Jimmy' $\notin X$.

It is customary to denote sets by capital letters and their elements by small letters We try to adhere to this convention as much as possible. □

Example 2.1. We looked in the introduction at the set $\Omega = \{1, 2, 3, 4, 5, 6\}$ of potential outcomes for the roll of a die. Then $3 \in \Omega$, $5 \in \Omega$, $-2 \notin \Omega$, $2.34 \notin \Omega$. □

Example 2.2 (No duplicates in sets)**.** The following collection of alphabetic letters is a set:

$$S_1 = \{a, e, i, o, u\}$$

and so is this one:

$$S_2 = \{a, e, e, i, i, i, o, o, o, o, u, u, u, u, u\}$$

Did you notice that those two sets are equal? □

Remark 2.1. The symbol n in the definition of $X = \{n : n \text{ is an integer and } 18 \leq n \leq 24\}$ is a **dummy variable** in the sense that it does not matter what symbol you use. The following sets all are equal to X:

$$\{x : x \text{ is an integer and } 18 \leq x \leq 24\},$$
$$\{\alpha : \alpha \text{ is an integer and } 18 \leq \alpha \leq 24\},$$
$$\{\mathfrak{Z} : \mathfrak{Z} \text{ is an integer and } 18 \leq \mathfrak{Z} \leq 24\} \ \square$$

Definition 2.2 (empty set)**.**

$\emptyset$ denotes the **empty set**. It is the set that does not contain any elements. $\square$

Definition 2.3 (subsets and supersets)**.**

- We say that a set A is a **subset** of the set B and we write $A \subseteq B$ if any element of A also belongs to B. Equivalently we say that B is a **superset** of the set A and we write $B \supseteq A$. We also say that B includes A or A is included by B. Note that $A \subseteq A$ and $\emptyset \subseteq A$ is true for any set A.
- If $A \subseteq B$ but $A \neq B$, i.e., there is at least one $x \in B$ such that $x \notin A$, then we say that A is a **strict subset** or a **proper subset** of B. We write "$A \subsetneq B$" Alternatively we say that B is a **strict superset** or a **proper superset** of A and we write "$B \supsetneq A$")

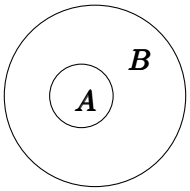


Figure 2.1: Set inclusion: $A \subseteq B,\ B \supseteq A$ $\square$

Remark 2.2. (a) We STRONGLY discourage the use of "$A \subset B$" in place of "$A \subsetneq B$" and of "$B \supset A$" in place of "$A \supsetneq B$". These are outdated symbols for $A \subseteq B$ and $A \supseteq B$

(b) Two sets A and B are equal means that they both contain the same elements. In other words, since $U \subseteq V$ means that the set V contains all elements of the set U,

$$A = B \ \Leftrightarrow \ \big[\, A \subseteq B \text{ and } B \subseteq A \,\big]\,. \tag{2.1}$$

In the above, "⇔" denotes the phrase "if and only if": The expression to the left ("$A = B$") means the same as the expression to the right ("$A \subseteq B$ and $B \subseteq A$"). The square brackets only serve to clarify that everything inbetween belongs to the scope of the right–hand side of "⇔". $\square$

Definition 2.4 (unions, intersections and disjoint unions of two sets)**.** Given are two sets A and B. No assumption is made that either one is contained in the other or that either one is not empty!

- The **union** $A \cup B$ (pronounced "A union B") is defined as the set of all elements which belong to at least one of A, B.
- The **intersection** $A \cap B$ (pronounced "A intersection B") is defined as the set of all elements which belong to both A and B.
- We call A and B **disjoint** , also **mutually disjoint** , if $A \cap B = \emptyset$. We then often write $A \uplus B$ (pronounced "A disjoint union B") rather than $A \cup B$.

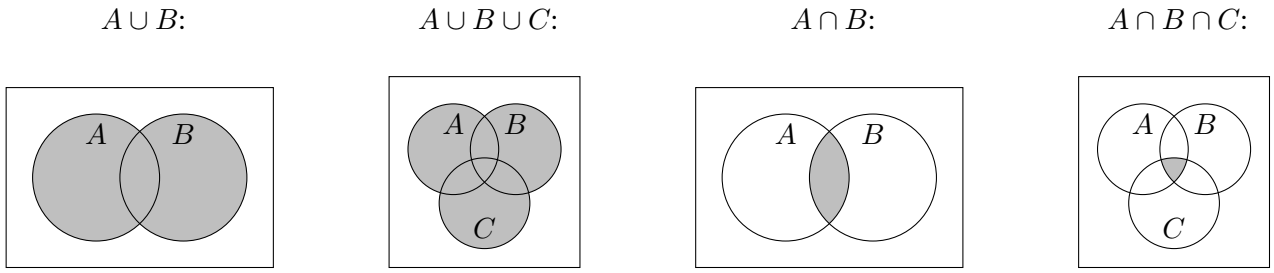


Figure 2.2: Union and intersection of sets □

A moment's reflection shows that we can characterize unions, intersections and disjoint unions to collections of more than two sets: 3 sets, 4 sets, 40 sets, $40 \cdot 10^{40}$ sets, even infinitely many sets. We do this in the next definition.

Definition 2.5 (Arbitrary unions, intersections and disjoint unions of sets)**.** Let J be an arbitrary, nonempty set. J may be finite or infinite. J may or may not be a set of numbers.

Assume that each $j \in J$ is associated with a set A_j. [14] For $J = \{\diamond, 3, \boldsymbol{\mathcal{X}}\}$, the sets are $A_\diamond, A_3, A_{\boldsymbol{\mathcal{X}}}$; and $J = \{1, 2, \dots\}$, yields the infinite sequence (of sets!) $A_1, A_2, \dots$.

- The **union** $\bigcup_{j \in J} A_j$ is defined as the set of all elements which belong to at least one A_j, where $j \in J$.
- The **intersection** $\bigcap_{j \in J} A_j$ is defined as the set of all elements which belong to each A_j, where $j \in J$.
- We call this collection of sets **disjoint** , also **mutually disjoint** , if $A_i \cap A_j = \emptyset$ whenever $i, j \in J$ and $i \neq j$. We then often write $\biguplus_{j \in J} A_j$ rather than $\bigcup_{j \in J} A_j$. □

[14] You might call this a **collection** of sets A_j which is **indexed by** the elements j of J and write $(A_j)_{j \in J}$ for this **indexed collection**. Later on, in Definition 2.25 on p.38, $(A_j)_{j \in J}$ will be called an indexed family of sets.

Remark 2.3. If $J = \{k_\star, k_\star + 1, k_\star + 2, \dots, k^\star - 1, k^\star\}$, we also write

$$\bigcup_{j=k_\star}^{k^\star} A_j, \quad \bigcap_{j=k_\star}^{k^\star} A_j, \quad \biguplus_{j=k_\star}^{k^\star} A_j, \qquad \text{for} \qquad \bigcup_{j\in J} A_j, \quad \bigcap_{j\in J} A_j, \quad \biguplus_{j\in J} A_j\,.$$

If $J = \{k_\star, k_\star + 1, k_\star + 2, \dots, \}$,in particular if $k_\star = 1$ (so $J = 1, 2, \dots$), we also write

$$\bigcup_{j=k_\star}^{\infty} A_j, \quad \bigcap_{j=k_\star}^{\infty} A_j, \quad \biguplus_{j=k_\star}^{\infty} A_j, \qquad \text{for} \qquad \bigcup_{j\in J} A_j, \quad \bigcap_{j\in J} A_j, \quad \biguplus_{j\in J} A_j\,. \ \square$$

Example 2.3. Some of the examples given here demonstrate that the index set need not be called J and its elements (they are dummy variables, just like t in $\int_a^b f(t)dt$ and k in $\sum_{k=5}^{25} x_k$). The third one also shows that the left to right order of the elements of the index set does not have to correspond to the order in which the unions or intersections are taken.

- If $I = \{1, 2\}$ and $A_1 \cap A_2 = \emptyset$, then $\biguplus_{\alpha\in I} A_\alpha = \biguplus_{\alpha=1}^{2} A_\alpha = A_1 \uplus A_2$.
- If $\boldsymbol{\mathcal{A}} = \{-1, 0, 1, 2\}$, then $\bigcap_{i\in\boldsymbol{\mathcal{A}}} A_i = \bigcap_{i=-1}^{2} A_i = A_{-1} \cap A_0 \cap A_1 \cap A_2$.
- If $J = \{\diamond, 9, \boldsymbol{\mathcal{X}}, F, 2\}$, then $\bigcap_{j\in J} \mathfrak{F}_j = \mathfrak{F}_{\boldsymbol{\mathcal{X}}} \cap \mathfrak{F}_9 \cap \mathfrak{F}_F \cap \mathfrak{F}_\diamond \cap \mathfrak{F}_2$.
- If $U = \{5, 6, 7, \dots\}$, then $\bigcup_{j\in U} C_j = \bigcup_{j=5}^{\infty} C_j = C_5 \cup C_6 \cup C_7 \cup \cdots$. $\square$

Remark 2.4. Convince yourself that for any sets A, B and C.

$$A \cap B \subseteq A \subseteq A \cup B, \tag{2.2}$$

$$A \subseteq B \Rightarrow A \cap B = A \text{ and } A \cup B = B, \tag{2.3}$$

$$A \subseteq B \Rightarrow A \cap C \subseteq B \cap C \text{ and } A \cup C \subseteq B \cup C. \tag{2.4}$$

The symbol $\Rightarrow$ stands for "allows us to conclude that". So $A \subseteq B \Rightarrow A \cap B = A$ means "From the truth of $A \subseteq B$ we can conclude that $A \cap B = A$ is true". Shorter: "From $A \subseteq B$ we can conclude that $A \cap B = A$". Shorter: "If $A \subseteq B$, then it follows that $A \cap B = A$". Shorter: "If $A \subseteq B$, then $A \cap B = A$". More technical: $A \subseteq B$ implies $A \cap B = A$. $\square$

Definition 2.6 (set differences and symmetric differences)**.** Given are two arbitrary sets A and B. No assumption is made that either one is contained in the other or contains any elements!

- The **difference set** or **set difference** $A \setminus B$ (pronounced "A minus B") is defined as the set of all elements which belong to A but not to B:

$$A \setminus B := \{x \in A : x \notin B\} \tag{2.5}$$

- The **symmetric difference** $A \triangle B$ (pronounced "A delta B") is defined as the set of all elements which belong to either A or B but not to both A and B:

$$A \triangle B := (A \cup B) \setminus (A \cap B) \quad \square \tag{2.6}$$

Definition 2.7 (Universal set)**.**

Usually there always is a big set Ω that contains everything we are interested in and we then deal with all kinds of subsets $A \subseteq \Omega$. Such a set is called a **"universal" set**. $\square$

Example 2.4.

(a) Often the context are the real numbers and their subsets. An appropriate universal set will then be $\mathbb{R}$. [15]

(b) We will discuss at length why the set $\{1, 2, 3, 4, 5, 6\}$ can be considered a universal set in the context of rolling a die. See Section 1.2 (A First Look at Probability). $\square$

If there is a universal set, it makes perfect sense to talk about the complement of a set:

Definition 2.8 (Complement of a set)**.** Let Ω be a universal set. The **complement** of a set $A \subseteq \Omega$ consists of all elements of Ω which do not belong to A. We write $A^{\complement}$. In other words:

$$A^{\complement} = \Omega \setminus A = \{\omega \in \Omega : x \notin A\} \quad \square \tag{2.7}$$

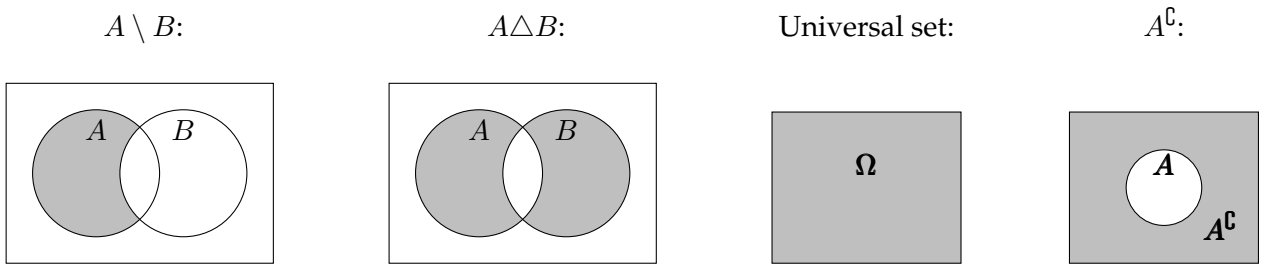


Figure 2.3: Difference, symmetric difference, universal set, complement

[15] $\mathbb{R}$ is the set of all real numbers, i.e., the kind of numbers that make up the x-axis and y-axis in a beginner's calculus course (see Section 2.3 (Numbers) on p.27).

Remark 2.5. Note that for any kind of universal set Ω it is true that

$$\Omega^{\complement} = \emptyset, \qquad \emptyset^{\complement} = \Omega. \ \square \tag{2.8}$$

Example 2.5 (Complement of a set relative to the unit interval)**.** Assume we are exclusively dealing with the unit interval, i.e., $\Omega = [0,1] = \{x \in \mathbb{R} : 0 \le x \le 1\}$. Let $a \in [0,1]$ and $\delta > 0$ and

$$A = \{x \in [0,1] : a - \delta < x < a + \delta\} \tag{2.9}$$

the "δ–neighborhood" [16] of a (with respect to $[0,1]$ because numbers outside the unit interval are not considered part of our universe). Then the complement of A is

$$A^{\complement} = \{x \in [0,1] : x \le a - \delta \text{ or } x \ge a + \delta\}. \ \square$$

Draw some Venn diagrams to visualize the following formulas. It is very important that you understand each one of them rather than simply trying to memorize them.

Proposition 2.1. *Let A, B, X be subsets of a universal set Ω and assume $A \subseteq X$. Then*

$$A \cup \emptyset = A; \qquad A \cap \emptyset = \emptyset \tag{2.10a}$$
$$A \cup \Omega = \Omega; \qquad A \cap \Omega = A \tag{2.10b}$$
$$A \cup A^{\complement} = \Omega; \qquad A \cap A^{\complement} = \emptyset \tag{2.10c}$$
$$A \triangle B = (A \setminus B) \uplus (B \setminus A) \tag{2.10d}$$
$$A \setminus A = \emptyset \tag{2.10e}$$
$$A \triangle \emptyset = A; \qquad A \triangle A = \emptyset \tag{2.10f}$$
$$X \triangle A = X \setminus A \tag{2.10g}$$
$$A \cup B = (A \triangle B) \uplus (A \cap B) \tag{2.10h}$$
$$A \cap B = (A \cup B) \setminus (A \triangle B) \tag{2.10i}$$
$$A \triangle B = \emptyset \text{ if and only if } B = A \tag{2.10j}$$

PROOF: The proof is left as exercise 2.2. See p.54. ■

Next we give a very detailed and rigorous proof of a simple formula for sets. You definitely want to remember the formulas, but it's perfectly OK to skip the proof.

Proposition 2.2 (Distributivity of unions and intersections for two sets)**.** *Let A, B, C be sets. Then*

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \tag{2.11}$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C). \tag{2.12}$$

PROOF: ⋆ We only prove (2.11). The proof of (2.12) is left as exercise 2.1.

[16] Draw a picture: The δ–neighborhood of a is the set of all points (in the universal set $[0,1]$) with distance less than δ from a.

PROOF of "$\subseteq$": Let $x \in (A \cup B) \cap C$. It follows from (2.2) on p.21 that $x \in (A \cup B)$, i.e., $x \in A$ or $x \in B$ (or both). It also follows from (2.2) that $x \in C$. We must show that $x \in (A \cap C) \cup (B \cap C)$ regardless of whether $x \in A$ or $x \in B$.

Case 1: $x \in A$. Since also $x \in C$, we obtain $x \in A \cap C$, hence, again by (2.2), $x \in (A \cap C) \cup (B \cap C)$, which is what we wanted to prove.

Case 2: $x \in B$. We switch the roles of A and B. This allows us to apply the result of case 1, and we again obtain $x \in (A \cap C) \cup (B \cap C)$.

PROOF of "$\supseteq$": Let $x \in (A \cap C) \cup (B \cap C)$, i.e., $x \in A \cap C$ or $x \in B \cap C$ (or both). We must show that $x \in (A \cup B) \cap C$ regardless of whether $x \in A \cap C$ or $x \in B \cap C$.

Case 1: $x \in A \cap C$. It follows from $A \subseteq A \cup B$ and (2.4) on p.21 that $x \in (A \cup B) \cap C$, and we are done in this case.

Case 2: $x \in B \cap C$. This time it follows from $A \subseteq A \cup B$ that $x \in (A \cup B) \cap C$. This finishes the proof of (2.11).

Epilogue: The proofs both of "$\subseteq$" and of "$\supseteq$" were **proofs by cases**, i.e., we divided the proof into several cases (to be exact, two for each of "$\subseteq$" and "$\supseteq$"), and we proved each case separately. For example we proved that $x \in (A \cup B) \cap C$ implies $x \in (A \cap C) \cup (B \cap C)$ separately for the cases $x \in A$ and $x \in B$. Since those two cases cover all possibilities for x the assertion "if $x \in (A \cup B) \cap C$ then $x \in (A \cap C) \cup (B \cap C)$" is proven. ■

Proposition 2.3 (De Morgan's Law for two sets)**.** *Let $A, B \subseteq \Omega$. Then the complement of the union is the intersection of the complements, and the complement of the intersection is the union of the complements:*

(2.13) ***a.*** $(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$ ***b.*** $(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$

PROOF:

1) First we prove that $(A \cup B)^{\complement} \subseteq A^{\complement} \cap B^{\complement}$:

Assume that $x \in (A \cup B)^{\complement}$. Then $x \notin A \cup B$, which is the same as saying that x does not belong to at least one of A and B. That in turn means that x belongs to all complements, i.e., to both $A^{\complement}$ and $B^{\complement}$ and hence, also to the intersection $A^{\complement} \cap B^{\complement}$.

2) Now we prove that $(A \cup B)^{\complement} \supseteq A^{\complement} \cap B^{\complement}$:

Let $x \in A^{\complement} \cap B^{\complement}$. Then x belongs to each one of $A^{\complement}, B^{\complement}$, hence to none of A, B, hence $x \notin A \cup B$. Therefore x belong to the complement of $A \cup B$. This completes the proof of formula **a**.

PROOF of **b**: The proof is very similar to that of formula **a** and left as an exercise. ■

Definition 2.9 (Power set)**.**

The **power set**

$$2^{\Omega} := \{A : A \subseteq \Omega\}$$

of a set Ω is the set of all its subsets. Note that many older texts also use the notation $\mathfrak{P}(\Omega)$ for the power set. □

Remark 2.6. Note that $\emptyset \in 2^{\Omega}$ for any set Ω, even if $\Omega = \emptyset$: $2^{\emptyset} = \{\emptyset\}$. It follows that the power set of the empty set is not empty. □

Definition 2.10 (Partition)**.** Let Ω be a set and $\mathfrak{A} \subseteq 2^{\Omega}$, i.e., the elements of $\mathfrak{A}$ are subsets of Ω.

We call $\mathfrak{A}$ a **partition** or a **partitioning** of Ω if
- **(a)** If $A, B \in \mathfrak{A}$ such that $A \neq B$ then $A \cap B = \emptyset$. In other words, $\mathfrak{A}$ consists of mutually disjoint subsets of Ω.
- **(b)** Each $x \in \Omega$ is an element of some $A \in \mathfrak{A}$. $\square$

Remark 2.7. Let Ω be a set and $\mathfrak{A} \subseteq 2^{\Omega}$. Then $\mathfrak{A}$ is a partition of Ω if and only if

$$\text{For each } x \in \Omega, \text{ there exists a UNIQUE } A \in \mathfrak{A} \text{ such that } x \in A\,. \ \square$$

Example 2.6.
- **a.** For $n \in \mathbb{Z}$ let $A_n := \{n\}$. Then $\mathfrak{A} := \{A_n : n \in \mathbb{Z}\}$ is a partition of $\mathbb{Z}$. $\mathfrak{A}$ is not a partition of $\mathbb{N}$ because not all its members are subsets of $\mathbb{N}$ and it is not a partition of $\mathbb{Q}$ or $\mathbb{R}$. The reason: $\frac{1}{2} \in \mathbb{Q}$ and hence $\frac{1}{2} \in \mathbb{R}$, but $\frac{1}{2} \notin A_n$ for any $n \in \mathbb{Z}$, hence condition **b** of def.2.10 is not satisfied.
- **b.** For $n \in \mathbb{N}$ let $B_n := [\,n^2, (n+1)^2[\, = \{x \in \mathbb{R} : n^2 \leq x < (n+1)^2\}$. Then $\mathfrak{B} := \{B_n : n \in \mathbb{N}\}$ is a partition of $[1, \infty[$. $\square$

Definition 2.11 (Size of a set)**.**
- **a.** Let X be a finite set, i.e., a set which only contains finitely many elements. We write $\big|X\big|$ for the number of its elements, and we call $\big|X\big|$ the **size** of the set X.
- **b.** For infinite, i.e., not finite sets Y, we define $\big|Y\big| := \infty$. $\square$

More will be said about sets later.

2.2 The Proper Use of Language in Mathematics: Any vs All, etc

Mathematics must be very precise in its formulations. Such precision is achieved not only by means of symbols and formulas, but also by its use of the English language. We will list some important points to consider early on in this document.

2.2.0.1 All vs. ANY

Assume for the following that X is a set of numbers. Do the following two statements mean the same?
- **(1)** It is true for ALL $x \in X$ that x is an integer.
- **(2)** It is true for ANY $x \in X$ that x is an integer.

You will hopefully agree that there is no difference and that one could rewrite them as follows:
- **(3)** ALL $x \in X$ are integers.
- **(4)** ANY $x \in X$ is an integer.
- **(5)** EVERY $x \in X$ is an integer.
- **(6)** EACH $x \in X$ is an integer.
- **(7)** IF $x \in X$ THEN x is an integer.

Is it then always true that ALL and ANY means the same? Consider

(8a) It is NOT true for ALL $x \in X$ that x is an integer.
(8b) It is NOT true for ANY $x \in X$ that x is an integer.
Completely different things have been said: Statement **(8)** asserts that as few as one item and as many as all items in X are not integers, whereas **(9)** states that no items, i.e., exactly zero items in X, are integers.

My suggestion: Express formulations like **(8b)** differently. You could have written instead

(8c) There is no $x \in X$ such that x is an integer.

2.2.0.2 AND vs. IF ... THEN

Some people abuse the connective AND to also mean IF ... THEN. However, mathematicians use the phrase "p AND q" exclusively to mean that something applies to both p and q. Contrast the use of AND in the following statements:

(9) "Jane is a student AND Joe likes baseball". This phrase means that both are true: Jane is indeed a student and Joe indeed likes baseball.
(10) "You hit me again AND you'll be sorry". **Never, ever use the word AND in this context!** A mathematician would express the above as "IF you hit me again THEN you'll be sorry".

2.2.0.3 OR vs. EITHER ... OR

The last topic we address is the proper use of "OR". In mathematics the phrase

(11) "p is true OR q is true"
is always to be understood as

(12) "p is true OR q is true OR BOTH are true", i.e., at least one of p, q is true.
This is in contrast to everyday language where "p is true OR q is true" often means that exactly one of p and q is true, but not not both.

When referring to a collection of items then the use of "OR" also is inclusive If the items $a, b, c, \ldots$ belong to a collection $\mathscr{C}$, e.g., if those items are elements of a set, then

(13) "a OR b OR c OR ..." means that we refer to at least one of $a, b, c, \ldots$.

> Note that "OR" in mathematics always is an **inclusive or**, i.e., "A OR B" means "A OR B OR BOTH". More generally, "A OR B OR ..." means "at least one of A, B, ...".
> To rule out that more than one of the choices is true you must use a phrase like "EXACTLY ONE OF A, B, C, ..." or "EITHER A OR B OR C OR ...". We refer to this as an **exclusive or**.

2.2.0.4 Some Convenient Shorthand Notation We have previously encountered the notation "$P \Rightarrow Q$" for "if P then Q", i.e., if P is true, then Q is true, and "$P \Leftrightarrow Q$" for "P iff Q", i.e., "P is true exactly when Q is true". We list them here again wich some additional convenient abbreviations.

- $\forall\, x \ldots$ For all x...
- $\exists\, x$ s.t. $\ldots$ There exists an x such that $\ldots$
- $\exists!\, x$ s.t. $\ldots$ There exists a UNIQUE x such that $\ldots$
- $P \Rightarrow Q$ If P then Q
- $P \Leftrightarrow Q$ P iff Q, i.e., P if and only if Q

It is important that you are clear about the difference between $\exists$ and $\exists!$.

$\exists x$: you can find at least one x but there might be more; potentially infinitely many!
$\exists! x$: you can find one and only one x; not zero, not two, not 200, ... $\square$

2.3 Numbers

We start with an informal classification of numbers.

Definition 2.12 (Types of numbers)**.** Here is a definition of the various kinds of numbers in a nutshell.

$\mathbb{N} := \{1, 2, 3, \ldots\}$ denotes the set of **natural numbers**.
$\mathbb{Z} := \{0, \pm 1, \pm 2, \pm 3, \ldots\}$ denotes the set of all **integers**.
$\mathbb{Q} := \{n/d : n \in \mathbb{Z}, d \in \mathbb{N}\}$ (fractions of integers) denotes the set of all **rational numbers**.
$\mathbb{R} :=$ {all integers or decimal numbers with finitely or infinitely many decimal digits} denotes the set of all **real numbers**.
$\mathbb{R} \setminus \mathbb{Q} =$ {all real numbers which cannot be written as fractions of integers} denotes the set of all **irrational numbers**. There is no special symbol for irrational numbers. Example: $\sqrt{2}$ and π are irrational. $\square$

Here are some customary abbreviations of some often referenced sets of numbers:

$\mathbb{N}_0 := \mathbb{Z}_+ := \mathbb{Z}_{\geq 0} := \{0, 1, 2, 3, \ldots\}$ denotes the set of nonnegative integers,
$\mathbb{R}_+ := \mathbb{R}_{\geq 0} := \{x \in \mathbb{R} : x \geq 0\}$ denotes the set of all nonnegative real numbers,
$\mathbb{R}^+ := \mathbb{R}_{>0} := \{x \in \mathbb{R} : x > 0\}$ denotes the set of all positive real numbers,
$\mathbb{R}_{\neq 0} := \{x \in \mathbb{R} : x \neq 0\}$. $\square$

Examples of rational numbers are

$$\tfrac{3}{4},\ -0.75,\ -\tfrac{1}{3},\ .\bar{3},\ \tfrac{7}{1},\ 16,\ \tfrac{13}{4},\ -5,\ 2.99\bar{9},\ -37\tfrac{2}{7}.$$

Note that a mathematician does not care whether a rational number is written as a fraction

$$\frac{\text{numerator}}{\text{denominator}}$$

or as a decimal numeral. The following all are representations of one third:

$$0.\bar{3} \;=\; .\bar{3} \;=\; 0.33333333333\ldots \;=\; \tfrac{1}{3} \;=\; \tfrac{-1}{-3} \;=\; \tfrac{2}{6}, \tag{2.14}$$

and here are several equivalent ways of expressing the number minus four:

$$-4 \;=\; -4.000 \;=\; -3.\bar{9} \;=\; -\tfrac{12}{3} \;=\; \tfrac{4}{-1} \;=\; \tfrac{-4}{1} \;=\; \tfrac{12}{-3} \;=\; -\tfrac{400}{100}. \tag{2.15}$$

Definition 2.13 (Intervals of Numbers)**.** For $a, b \in \mathbb{R}$ we have the following intervals.

- $[a,b] := \{x \in \mathbb{R} : a \le x \le b\}$ is the **closed interval** with endpoints a and b.
- $]a,b[:= \{x \in \mathbb{R} : a < x < b\}$ is the **open interval** with endpoints a and b.
- $[a,b[:= \{x \in \mathbb{R} : a \le x < b\}$ and $]a,b] := \{x \in \mathbb{R} : a < x \le b\}$ are **half-open intervals** with endpoints a and b.

The symbol "∞" stands for an object which itself is not a number but is larger than any (real) number, and the symbol "$-\infty$" stands for an object which itself is not a number but is smaller than any number. We thus have $-\infty < x < \infty$ for any number x. This allows us to define the following intervals of "infinite length":

$$\begin{aligned}]-\infty,a] :=& \{x \in \mathbb{R} : x \le a\}, \quad]-\infty,a[:= \{x \in \mathbb{R} : x < a\}, \\ ]a,\infty[:=& \{x \in \mathbb{R} : x > a\}, \quad [a,\infty[:= \{x \in \mathbb{R} : x \ge a\}, \quad]-\infty,\infty[:= \mathbb{R} \end{aligned} \tag{2.16}$$

You should always work with $a < b$. In case you don't, you get

- $[a,a] = \{a\}$; $[a,a[=]a,a[=]a,a] = \emptyset$
- $[a,b] = [a,b[=]a,b[=]a,b] = \emptyset$ for $a \ge b$ $\square$

Definition 2.14 (Extended real numbers)**.** ⋆

It is sometimes convenient to refer to the set

$$\overline{\mathbb{R}} := [-\infty,\infty] := \mathbb{R} \cup \{-\infty\} \cup \{\infty\} \tag{2.17}$$

as the **extended real numbers**. and to work with intervals such as

$$[-\infty,a] := \{-\infty\} \cup]-\infty,a\,], \quad]b,\infty] :=]b,\infty[\cup \{\infty\}\,, \ldots \ \square \tag{2.18}$$

Remark 2.8 (Extended real numbers arithmetic)**.** When working with extended real–valued functions we must be clear about the rules of arithmetic where $\pm\infty$ is involved. In the following assume that $c \in \mathbb{R}$ and $0 < p < \infty$.

Rules for Addition:

$$c \pm \infty = \infty \pm c = \infty, \tag{2.19}$$

$$c \pm (-\infty) = -\infty \pm c = -\infty, \tag{2.20}$$

$$\infty + \infty = \infty, \tag{2.21}$$

$$-\infty - \infty = -\infty, \tag{2.22}$$

$$(\pm\infty) \mp \infty = \textbf{UNDEFINED}. \tag{2.23}$$

Rules for Multiplication:

$$p \cdot (\pm\infty) = (\pm\infty) \cdot p = \pm\infty, \tag{2.24}$$
$$(-p) \cdot (\pm\infty) = (\pm\infty) \cdot (-p) = \mp\infty, \tag{2.25}$$
$$0 \cdot (\pm\infty) = (\pm\infty) \cdot 0 = \frac{0}{0} = 0, \quad \text{and} \quad \frac{1}{\infty} = 0, \tag{2.26}$$
$$(\pm\infty) \cdot (\pm\infty) = \infty, \tag{2.27}$$
$$(\pm\infty) \cdot (\mp\infty) = -\infty, \tag{2.28}$$

Be clear about the ramifications of those rules. Rule (2.23) implies that if we have two extended real–valued functions f, g defined on a domain A then $f + g$ is only defined on

$$A \setminus \{x \in A : \text{ either } [f(x) = \infty \text{ and } g(x) = -\infty] \text{ or } [f(x) = -\infty \text{ and } g(x) = \infty]\},$$

and $f - g$ is only defined on

$$A \setminus \{x \in A : \text{ either } [f(x) = g(x) = \infty] \text{ or } [f(x) = g(x) = -\infty]\}.$$

That is easy to understand and remember, but the real danger comes from rule (2.26) which you might not have expected:

$$0 \cdot \pm\infty = \pm\infty \cdot 0 = 0.$$

This convention is very convenient for integrals, but it comes at a price:
$a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$ no longer implies $\lim_{n\to\infty} a_n b_n = ab$.

A counterexample would be: $a_n = n, b_n = \frac{1}{n}$. □

Notation 2.1 (Notation Alert for intervals of integers or rational numbers)**.**

It is at times convenient to also use the notation $[\ldots],]\ldots[, [\ldots[,]\ldots]$, for intervals of integers or rational numbers. We will subscript them with $\mathbb{Z}$ or $\mathbb{Q}$. For example,

$$[\,3, n\,]_{\mathbb{Z}} = [\,3,\ n] \cap \mathbb{Z} = \{k \in \mathbb{Z} : 3 \le k \le n\},$$
$$]-\infty, 7\,]_{\mathbb{Z}} =]-\infty, 7\,] \cap \mathbb{Z} = \{k \in \mathbb{Z} : k \le 7\} = \mathbb{Z}_{\le 7},$$
$$]a, b[_{\mathbb{Q}} =]a, b[\cap \mathbb{Q} = \{q \in \mathbb{Q} : a < q < b\}.$$

An interval which is not subscripted always means an interval of real numbers, but we will occasionally write, e.g., $[a, b]_{\mathbb{R}}$ rather than $[a, b]$, if the focus is on integers or rational numbers and an explicit subscript helps to avoid confusion. □

Definition 2.15 (Absolute value, positive and negative part)**.** For a real number x we define its

$$\textbf{absolute value:} \quad |x| \;=\; \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

$$\textbf{positive part:} \quad x^+ \;=\; \max(x,0) \;=\; \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

$$\textbf{negative part:} \quad x^- \;=\; \max(-x,0) \;=\; \begin{cases} -x & \text{if } x \leq 0, \\ 0 & \text{if } x > 0. \end{cases}$$

If f is a real–valued function then we define the functions $|f|, f^+, f^-$ argument by argument:

$$|f|(x) \;:=\; |f(x)|, \qquad f^+(x) \;:=\; \big(f(x)\big)^+, \qquad f^-(x) \;:=\; \big(f(x)\big)^-. \;\square$$

For completeness we also give the definitions of min and max.

Definition 2.16 (Minimum and maximum)**.** For two real number x, y we define

$$\textbf{maximum:} \quad x \vee y \;=\; \max(x,y) \;=\; \begin{cases} x & \text{if } x \geq y, \\ y & \text{if } x \leq y. \end{cases}$$

$$\textbf{minimum:} \quad x \wedge y \;=\; \min(x,y) \;=\; \begin{cases} y & \text{if } x \geq y, \\ x & \text{if } x \leq y. \end{cases}$$

If f and g is are real–valued function then we define the functions $f \vee g = \max(f,g)$ and $f \wedge g = \min(f,g)$ argument by argument:

$$f \vee g(x) \;:=\; f(x) \vee g(x) \;=\; \max\big(f(x), g(x)\big), \quad f \wedge g(x) \;:=\; f(x) \wedge g(x) \;=\; \min\big(f(x), g(x)\big). \;\square$$

Remark 2.9. You are advised to compute $|x|, x^+, x^-$ for $x = -5, x = 5, x = 0$ and convince yourself that the following is true:

$$\begin{aligned} x \;&= x^+ \,-\, x^-, \\ |x| \;&= x^+ \,+\, x^-, \end{aligned}$$

Thus any real–valued function f satisfies

$$\begin{aligned} f \;&= f^+ \,-\, f^-, \\ |f| \;&= f^+ \,+\, f^-, \end{aligned}$$

Get a feeling for the above by drawing the graphs of $|f|, f^+, f^-$ for the functon $f(x) = 2x$. $\square$

Assumption 2.1 (Square roots are always assumed nonnegative)**.** Remember that for any number a it is true that

$$a \cdot a \;=\; (-a)(-a) = a^2, \qquad \text{e.g.,} \quad 2^2 = (-2)^2 = 4,$$

or that, expressed in form of square roots, for any number $b \geq 0$

$$(+\sqrt{b})(+\sqrt{b}) \;=\; (-\sqrt{b})(-\sqrt{b}) = b.$$

We will always assume that "$\sqrt{b}$" is the **positive** value unless the opposite is explicitly stated.

Example: $\sqrt{9} = +3$, not -3. □

Remark 2.10. For any real number x we have

$$\sqrt{x^2} \;=\; |x|. \;\square \tag{2.29}$$

Proposition 2.4 (The Triangle Inequality for real numbers)**.** *The following inequality is used all the time in mathematical analysis to show that the size of a certain expression is limited from above:*

$$\textit{Triangle Inequality}: \qquad |\boldsymbol{a_1} + \boldsymbol{a_2} + \cdots + \boldsymbol{a_n}| \;\leq\; |\boldsymbol{a_1}| + |\boldsymbol{a_2}| + \cdots + |\boldsymbol{a_n}| \tag{2.30}$$

This inequality is true for any list of real numbers $a_1, a_2, \ldots, a_n$.

PROOF:

It is easy to prove this for $n = 2$: Just look separately at the three cases where both numbers are nonnegative, both are negative, or one of each is positive and negative. ■

2.4 Functions and Sequences

Introduction 2.2. You are familiar with functions from calculus. Examples are $f_1(x) = \sqrt{x}$ and $f_2(x, y) = \ln(x - y)$. Sometimes $f_1(x)$ means the entire graph, i.e., the entire collection of points $\big(x, \sqrt{x}\big)$ in the plane and sometimes it just refers to the function value $\sqrt{x}$ for a "fixed but arbitrary" number x. In case of the function $f_2(x)$: Sometimes $f_2(x, y)$ means the entire graph, i.e., the entire collection of points $\big((x, y), \ln(x - y)\big)$ in threedimensional space. At other times this expression just refers to the function value $\ln(x - y)$ for a pair of "fixed but arbitrary" numbers (x, y).

To obtain a usable definition of a function there are several things to consider. In the following $f_1(x)$ and $f_2(x, y)$ again denote the functions $f_1(x) = \sqrt{x}$ and $f_2(x, y) = \ln(x - y)$.

a. The source of all allowable arguments (x–values in case of $f_1(x)$ and (x, y)–values in case of $f_2(x, y)$) will be called the **domain** of the function. The domain is explicitly specified as part of a function definition and it may be chosen for whatever reason to be only a subset of all arguments for which the function value is a valid expression. In case of the function $f_1(x)$ this means that the domain must be a subset of the interval $[0, \infty[$ because the square root of a negative number cannot be taken. In case of the function $f_2(x, y)$ this means that the domain must be a subset of

$$\{\, (x, y) : \; x, y \in \mathbb{R} \text{ and } x - y > 0 \,\},$$

because logarithms are only defined for strictly positive numbers.

b. The set to which all possible function values belong will be called the **codomain** of the function. As is the case for the domain, the codomain also is explicitly specified as part of a function definition. It may be chosen as any superset of the set of all function values for which the argument belongs to the domain of the function.
For the function $f_1(x)$ this means that we are OK if the codomain is a superset of the interval $[0, \infty[$. Such a set is big enough because square roots are never negative. It is OK to specify the interval $]-3.5, \infty[$ or even the set $\mathbb{R}$ of all real numbers as the codomain. In case of the function $f_2(x, y)$ this means that we are OK if the codomain contains $\mathbb{R}$. Not that it would make a lot of sense, but the set $\mathbb{R} \cup \{$ all inhabitants of Chicago $\}$ also is an acceptable choice for the codomain.

c. A function $y = f(x)$ is not necessarily something that maps (assigns) numbers or pairs of numbers to numbers. Rather domain and codomain can be a very different kind of animal. The following example will be very relevant for the remainder of the course:

> At the end of Section 1.2 (A First Look at Probability) We informally defined the probability associated with rolling a die as a function $A \mapsto P(A)$ which maps subsets A of $\Omega = \{1, 2, \ldots, 6\}$ to a real number $0 \leq P(A) \leq 1$. Thus, the domain here is 2^Ω, the power set of Ω; the codomain is $[0, 1]$ (or any superset of $[0, 1]$).

d. Considering all that was said so far one can think of the graph of a function $f(x)$ with domain D and codomain C (see earlier in this note) as the set

$$\Gamma_f := \{\big(x, f(x)\big) : x \in D\}.$$

Alternatively one can characterize this function by the assignment rule which specifies how $f(x)$ depends on any given argument $x \in D$. We write "$x \mapsto f(x)$" to indicate this. You can also write instead $f(x) =$ whatever the actual function value will be.
This is possible if one does not write about functions in general but about specific functions such as $f_1(x) = \sqrt{x}$ and $f_2(x, y) = \ln(x - y)$. We further write

$$f : D \longrightarrow C$$

as a short way of saying that the function $f(x)$ has domain D and codomain C.
In case of the function $f_1(x) = \sqrt{x}$ for which we might choose the interval $X := [\, 2.5, 7\,]$ as the domain (small enough because $X \subseteq [0, \infty[$) and $Y :=]1, 3[$ as the codomain (big enough because $1 < \sqrt{x} < 3$ for any $x \in X$) we specify this function as

$$\text{either } \; f_1 : [\, 2.5, 7\,] \to]1, 3[; \quad x \mapsto \sqrt{x} \qquad \text{or } \; f_1 : [\, 2.5, 7\,] \to]1, 3[; \quad f(x) = \sqrt{x}.$$

Let us choose $U := \{(x,y) : x,y \in \mathbb{R} \text{ and } 1 \leq x \leq 10 \text{ and } y < -2\}$ as the domain and $V := [0,\infty[$ as the codomain for $f_2(x,y) = \ln(x-y)$. These choices are OK because $x - y \geq 1$ for any $(x,y) \in U$ and hence $ln(x-y) \geq 0$, i.e., $f_2(x,y) \in V$ for all $(x,y \in U$. We specify this function as

$$\text{either } \; f_2 : U \to V, \quad (x,y) \mapsto \ln(x-y) \qquad \text{or } \; f_2 : U \to V, \quad f(x,y) = \ln(x-y). \; \square$$

We incorporate what we noted above into this definition of a function.

Definition 2.17 (Function)**.**

A **function** f consists of two nonempty sets X and Y and an assignment rule $x \mapsto f(x)$ which assigns any $x \in X$ uniquely to some $y \in Y$. We write $f(x)$ for this assigned value and call it the **function value** of the **argument** x. X is called the **domain** and Y is called the **codomain** of f. We write

$$f : X \to Y, \qquad x \mapsto f(x). \tag{2.31}$$

We read "$a \mapsto b$" as "a is assigned to b" or "a maps to b" and refer to $\mapsto$ as the **maps to operator** or **assignment operator**. The **graph** of such a function is the collection of pairs

$$\Gamma_f := \{\big(x, f(x)\big) : x \in X\}, \tag{2.32}$$

and the subset $f(X) := \{f(x) : x \in X\}$ of Y is called the **range** of the function f. $\square$

Note that the codomain) Y of f and its range $f(X)$ can be vastly different. For example, if $f : \mathbb{R} \to \mathbb{R}$ is given by the assignment $f(x) = \sin(x)$ then $f(\mathbb{R}) = [-1,1]$ is a very small part of the codomain!

Remark 2.11. The name given to the argument variable is irrelevant. Let f_1, f_2, X, Y, U, V be as defined in **d** of the introduction to ch.2.4 (A First Look at Functions and Sequences). The function

$$g_1 : X \to Y, \quad p \mapsto \sqrt{p}$$

is identical to the function f_1. The function

$$g_2 : U \to V, \quad (t,s) \mapsto \ln(t-s)$$

is identical to the function f_2 and so is the function

$$g_3 : U \to V, \quad (s,t) \mapsto \ln(s-t).$$

The last example illustrates the fact that you can swap function names as long as you do it consistently in all places. $\square$

We all know what it means that $f : \mathbb{R} \to]0,\infty]$; $x \mapsto e^x$ has $f^{-1}(x) = \ln(x)$ as its inverse function:

- The arguments of f^{-1} will be the function values of f and the function values of f^{-1} will be the arguments of f: $f(x) = e^x = y \;\Leftrightarrow\; g(y) = \ln(y) = x$.
- f and f^{-1} cancel each other, i.e.,

$$f^{-1}\big(f(y)\big) \;=\; y \qquad \text{and} \qquad f\big(f^{-1}(x)\big) \;=\; x\,.$$

- Not so obvious but very useful: We want both codomains to be so small that $f^{-1}\big(f(y)\big) = y$ is true for all y in the codomain of f and $f\big(f^{-1}(x)\big) = x$ is true for all x in the codomain of f^{-1}. One can show that this requires

$$\text{domain of } f \;=\; \text{codomain of } f^{-1} \qquad \text{and} \qquad \text{domain of } f^{-1} \;=\; \text{codomain of } f.$$

This leads to the following definition for the inverse of a function.

Definition 2.18 (Inverse function)**.**

Given are two nonempty sets X and Y and a function $f : X \to Y$ with domain X and codomain Y. We say that f has an **inverse function** if it satisfies all of the following conditions which uniquely determine this inverse function, so that we are justified to give it the symbol f^{-1}:

(a) $f^{-1} : Y \to X$, i.e., f^{-1} has domain Y and codomain X.

(b) $f^{-1}\big(f(x)\big) = x$ for all $x \in X$, and $f\big(f^{-1}(y)\big) = y$ for all $y \in Y$. □

Definition 2.19 (Surjective, injective and bijective functions)**.**

Given are two nonempty sets X and Y and a function $f : X \to Y$ with domain X and codomain Y. We say that

(a) f is "one–one" or **injective**, if for each $y \in Y$ there is at most one $x \in X$ such that $f(x) = y$.

(b) f is "onto" or **surjective**, if for each $y \in Y$ there is at least one $x \in X$ such that $f(x) = y$.

(c) f is **bijective**, f is both injective and surjective. □

Remark 2.12. that One can show that a function f has an inverse f^{-1} if and only if f is bijective. □

Remark 2.13. that If the inverse function f^{-1} exists and if $x \in X$ and $y \in Y$, then we have the relation

$$y \;=\; f(x) \quad \Leftrightarrow \quad x \;=\; f^{-1}(y)\,.$$

Example 2.7. If h is a function, we write Dom_h and Cod_h for its domain and codomain. Be sure you understand the following:

(a) $f : \mathbb{R} \to \mathbb{R};\; x \to e^x$ does not have an inverse $f^{-1}(y) = \ln(y)$ since its domain $Dom_{f^{-1}}$ would have to be the codomain $\mathbb{R}$ of f and $\ln(y)$ is not defined for $y \leq 0$.

(b) $g : \mathbb{R} \to]0,\infty[;\ x \to e^x$ has the inverse $g^{-1} :]0,\infty[\to \mathbb{R};\ g^{-1}(y) = \ln(y)$ since

$$Dom_{g^{-1}} = Cod_g =]0,\infty[, \qquad Cod_{g^{-1}} = Dom_g = \mathbb{R},$$
$$e^{\ln(y)} = y \text{ for } 0 < y < \infty, \qquad \ln(e^x) = x \text{ for all } x \in \mathbb{R}. \ \square$$

Definition 2.20 (Restriction/Extension of a function). ⋆ Given are three nonempty sets A, X and Y such that $A \subseteq X$, and a function $f : X \to Y$ with domain X. We define the **restriction of** f **to** A as the function

$$f\big|_A : A \to Y \quad \text{defined as} \quad f\big|_A(x) := f(x) \text{ for all } x \in A. \tag{2.33}$$

Conversely let $f : A \to Y$ and $\varphi : X \to Y$ be functions such that $f = \varphi\ |_A$. We then call φ an **extension** of f to X. $\square$

We now briefly address sequences and subsequences.

Definition 2.21. Let $n_\star$ be an integer and assume that an item x_j associated

- **either** with each integer $j \geq n_\star$, In other words, we have an item x_j assigned to each $j = n_\star, n_\star + 1, n_\star + 2, \ldots$.
- **or** with each integer j such that $n_\star \leq j \leq n^\star$. In this case an item x_j is assigned to each $j = n_\star, n_\star + 1, \ldots, n^\star$.

Such items can be anything, but we usually deal with numbers or outcomes or sets of outcomes of an experiment.

- In the first case we usually write $x_{n_\star}, x_{n_\star+1}, x_{n_\star+2}, \ldots$ or $(x_n)_{n \geq n_\star}$ for such a collection of items and we call it a **sequence** with **start index** $n_\star$.
- In the second case we speak of a **finite sequence**, which starts at $n_\star$ and ends at $n^\star$. We write $(x_n)_{n_\star \leq n \leq n^\star}$ or $x_{n_\star}, x_{n_\star+1}, \ldots, x_{n^\star}$ for such a finite collection of items.
- If we refer to a sequence $(x_n)_n$ without qualifying it as finite then we imply that we deal with an **infinite sequence**, $x_{n_\star}, x_{n_\star+1}, x_{n_\star+2}, \ldots$. $\square$

Example 2.8.

(1) If $u_k = k^2$ for $k \in \mathbb{Z}$, then $(u_k)_{k \geq -2}$ is the sequence of integers $4, 1, 0, 1, 4, 9, 16, \ldots$.

(2) If $A_j = [-1 - \frac{1}{j}, 1 + \frac{1}{j}] = \{x \in \mathbb{R} : -1 - \frac{1}{j} \leq x \leq 1 + \frac{1}{j}\}$, then $(A_j)_{j \geq 3}$ is the sequence of intervals of real numbers $[-\frac{4}{3}, \frac{4}{3}]$, $[-\frac{5}{4}, \frac{5}{4}]$, $[-\frac{6}{5}, \frac{6}{5}], \ldots$. This is a sequence of sets! $\square$

Remark 2.14 (Sequences are functions)**.** that

- One can think of a sequence $(x_i)_{i \geq n_\star}$ in terms of the assignment $i \mapsto x_i$. This sequence can then be interpreted as the function

$$x(\cdot) : [n_\star, \infty[_{\mathbb{Z}} \longrightarrow \text{ suitable codomain}; \quad i \mapsto x(i) := x_i\,,$$

  where that "suitable codomain" depends on the nature of the items x_i.
- In Example 2.8**(1)**, we could chose $\mathbb{Z}$ as that codomain. In Example 2.8**(2)** $2^{\mathbb{R}}$, the power set of $\mathbb{R}$ would be an appropriate choice. $\square$

Definition 2.22.

- If $(x_n)_n$ is a finite or infinite sequence and one pares down the full set of indices to a subset $\{n_1,\ n_2,\ n_3, \dots\}$ such that $n_1 < n_2 < n_3 < \dots$, then we call the corresponding thinned out sequence $(x_{n_j})_{j\in\mathbb{N}}$ a **subsequence** of that sequence.
- If this subset of indices is finite, i.e., we have $n_1 < n_2 < \cdots < n_K$ for some suitable $K \in \mathbb{N}$, then we call $(x_{n_j})_{j\le K}$ a **finite subsequence** of the original sequence. □

Note that subsequences of finite sequences are necessarily finite whereas subsequences of infinite sequences can be finite or infinite.

Remark 2.15. Does it matter whether we look at a sequence $\big(x_j\big)_{j\in J}$ or at the corresponding set $\{x_j : j \in J\}$? The answer: **THIS CAN MATTER GREATLY!** Consider the sequence

$$x_1 = -1,\ x_2 = 1,\ x_3 = -1,\ x_1 = -1,\ \dots\ ; \qquad \text{i.e.,}\quad x_n = (-1)^n \text{ for } n \in \mathbb{N}$$

- The sequence is infinite, since the index set $\mathbb{N}$ is infinite
- Let $A := \{x_j : j \in \mathbb{N}\}$. Since **sets have no duplicates**, $A = \{-1, 1\}$ has only two elements.
- The ordering of the indices j is lost when considering the set: There is no difference between $\{-1, 1\}$ and $\{1, -1\}$!

Considering the last point, do not confuse the ordering of the indices j with a possible ordering of the x_j! The order may be reversed (e.g., $x_j = 5 - j$), neither increasing nor decreasing ($x_j = \sin(j)$), or there is no ordering (x_j = eye color of person j). □

Definition 2.23. We give some convenient definitions and notations for monotone sequences of numbers, functions and sets.

(a) Let x_n be a sequence of extended real–valued numbers.
- We call x_n a **nondecreasing** or **increasing** sequence, if $\quad j < n \ \Rightarrow\ x_j \le x_n$.
- We call x_n a **strictly increasing** sequence, if $\quad j < n \ \Rightarrow\ x_j < x_n$.
- We call x_n a **nonincreasing** or **decreasing** sequence, if $\quad j < n \ \Rightarrow\ x_j \ge x_n$.
- We call x_n a **strictly decreasing** sequence, if $\quad j < n \ \Rightarrow\ x_j > x_n$.
- We write $x_n \uparrow$ for nondecreasing x_n, and $x_n \uparrow x$ to indicate that $\lim\limits_{n\to\infty} x_n = x$,
- We write $x_n \downarrow$ for nonincreasing x_n, $x_n \downarrow x$ to indicate that $\lim\limits_{n\to\infty} x_n = x$. □

(b) Let A_n be a sequence of sets.
- We call A_n a **nondecreasing** or **increasing** sequence, if $\quad j < n \ \Rightarrow\ A_j \subseteq A_n$.
- We call A_n a **strictly increasing** sequence, if $\quad j < n \ \Rightarrow\ A_j \subsetneq A_n$.
- We call A_n a **nonincreasing** or **decreasing** sequence, if $\quad j < n \ \Rightarrow\ A_j \supseteq A_n$.
- We call A_n a **strictly decreasing** sequence, if $\quad j < n \ \Rightarrow\ A_j \subsetneq A_n$.
- We write $A_n \uparrow$ for nondecreasing A_n, and $A_n \uparrow A$ to indicate that $\bigcup_n A_n = A$,
- We write $A_n \downarrow$ for nonincreasing A_n, $A_n \downarrow A$ to indicate that $\bigcap_n A_n = A$. □

Example 2.9.

(a) The sequence $x_n = -\frac{1}{n}$ is strictly increasing.
(b) The sequence $y_n = \frac{1}{n}$ is strictly decreasing.
(c) The sequence $a_1 = 1$, $a_{n+1} = a_n$ for even n and $a_{n+1} = -\frac{1}{n}$ for odd n, is nonincreasing.
(c) The sequence $b_1 = 1$, $b_{n+1} = b_n$ for even n and $b_{n+1} = \frac{1}{n}$ for odd n, is nondecreasing. □

There are different degrees of infinity for the size of a set. Finite sets and many inifinite sets are "small enough" to list all their elements in a finite or infinite sequence. Other infinite sets are too big for that.

Definition 2.24 (Countable and uncountable sets)**.** Let X be a set.

(a) We call X **countable** if its elements can be written as a finite sequence (those are the finite sets) $X = \{x_1, x_2, \dots, x_n\}$ or as an infinite sequences. $X = \{x_1, x_2, \dots\}$.
(b) We call X **countably infinite** X is both countable and infinite, i.e., there is an infinite sequence. $X = \{x_1, x_2, \dots\}$. of distinct items x_j.
(c) We call a nonempty set **uncountable** if it is not countable, i.e., its elements cannot be sequenced.
(d) By convention the empty set, $\emptyset$, is countable. □

Fact 2.1. *One can prove the following important facts:*

(a) *The integers are countable. (Easy: $\mathbb{Z} = \{0, -1, 1, -2, 2, -3, 3, \dots\}$) lists all elements of $\mathbb{Z}$ in a sequence.*
(b) *Subsets of countable sets are countable. (Easy: If $X = \{x_1, x_2, \dots\}$ and $A \subseteq X$, then remove all x_j that are not in A. That subsequence lists the elements of A.*
(c) *Countable unions of countable sets are countable: If $A_1, A_2, \dots$ is a finite or infinite sequence of sets, then $A_1 \cup A_2 \cup \cdots$ is countable.*
(d) *The rational numbers $\mathbb{Q}$ are countable. A proof is given below.*
(e) *The real numbers $\mathbb{R}$ are uncountable!* □

⋆ Here is a proof that $\mathbb{Q}$ is countable. For fixed $d \in \mathbb{N}$, let $A_d := \{n/d : n \in \mathbb{Z}\}$ ("d" for denominator). Then is countable since it can be sequenced as follows.

$$A_d = \{0, -\frac{1}{d}, \frac{1}{d}, -\frac{2}{d}, \frac{2}{d}, \dots\}$$

The assertion follows from fact **(c)** and $\mathbb{Q} = \bigcup_{d=1}^{\infty} A_d$ (WHY?)

Example 2.10. ⋆ For $a, b, r \in \mathbb{R}$, let $A_{(a,b,r)} := \{(x, y) \in \mathbb{R}^2\}$ such that $(x-a)^2 + (y-b)^2 = r^2$, i.e., $A_{(a,b,r)}$ is the circle with radius $|r|$ around the point (a, b) in the plane. It is not possible to write the indexed collection

$$\big(A_{(a,b,r)}\big)_{(a,b,r)\in\mathbb{R}^3}$$

as a sequence, since $\mathbb{R}^3$ is bigger than the uncountable set $\mathbb{R}$, hence cannot be sequenced. □

There is a name for those "generalized sequences" $\big(x_i\big)_{i\in I}$ which have an index set that not necessarily consists of integers $n_\star, n_\star + 1, \dots, n^\star$ or $n_\star, n_\star + 1, \dots$ or of a subset of such a set. The next definition is marked as optional and you not need remember it for quizzes or exams. But you must remember it well enough to understand problems and propositions which refer to families.

Definition 2.25 (Families). ⋆

Let I and X be nonempty sets such that each $i \in I$ is associated with some $x_i \in X$. Then
a. $(x_i)_{i\in I}$ is called an **indexed family** or simply a **family** in X.
b. I is called the **index set** of the family.
c. For each $i \in IJ$, x_i is called a **member of the family** $(x_i)_{i\in I}$. □

Remark 2.16 (Families are functions). that

We saw in example 2.14 on p.35 that sequences $(x_n)_n$ can be interpreted as functions with domain = index set and codomain = a set that contains all members x_n. This also holds true for families and is particularly easily understood if the family $(x_i)_{i\in I}$ in X is written in a way that each member explicitly tracks the index that it is associated with, i.e., we write $(i, x_i)_{i\in I}$. The set

$$\Gamma_f := \{ (i, x_i) : i \in I \}$$

is the graph Γ_f of the function

$$f : I \longrightarrow X ; \quad i \mapsto f(i) := x_i .$$

At the end of Definition 2.4 on p.20 we defined unions and intersections of any collection of sets $(A_i)_{i\in J}$ which is indexed by integers, i.e., $J \subseteq \mathbb{Z}$. We did so by saying that [17]

$$\bigcup_{i\in J} A_i = \{x : \exists\, i_0 \in J \text{ s.t. } x \in A_{i_0}\} \qquad \text{and} \qquad \bigcap_{i\in J} A_i = \{x : \forall\, i \in J : x \in A_i\} .$$

This allows us to generalize unions and intersections of finite and infinite sequences of sets to collections of sets with an arbitrary index set. Note the following:

- The next definition is NOT marked as OPTIONAL
- It contains Definition 2.4 as a special case!

Definition 2.26 (Arbitrary unions and intersections of families of sets). Let J be an arbitrary, nonempty set and $(A_j)_{j\in J}$ a family of sets with index set J. We define

- The **union** $\bigcup_{j\in J} A_j := \{x : \exists\, i_0 \in J \text{ s.t. } x \in A_{i_0}\}$.
- The **intersection** $\bigcap_{j\in J} A_j = \{x : \forall\, i \in J : x \in A_i\}$.
- If the sets A_i are disjoint, we often write $\biguplus_{j\in J} A_j$ rather than $\bigcup_{j\in J} A_j$.
- Let $(B_j)_{j\in J}$ be a family of subsets of a set X. We call this family a **partition** or a **partitioning** of X if the corresponding set of sets $\{B_i : i \in J\}$ is a partition of X:
  (a) $i \neq j \Rightarrow B_i \cap B_j = \emptyset$ **(b)** $X = \biguplus_{j\in J} B_j$. See Definition 2.10 on p.25. □

[17] See paragraph 2.2.0.4 (Some Convenient Shorthand Notation) on p.26 about ∀ and ∃.

Notation 2.2. Empty unions and intersections: If $J = \emptyset$, it seems reasonable to define $\bigcup_{j\in\emptyset} A_j :=$ $\emptyset$, since there is no x for which one can find $i_0 \in \emptyset$ such that $x \in A_{i_0}$. Also, since there are no indices $i \in \emptyset$, any x, no matter what it might be, satisfies $x \in A_i$ for all $i \in \emptyset$. So should one define the intersection of an empty family as $\bigcap_{j\in\emptyset} A_j :=$ { everything }? It turns out that the use of an "everything" set leads to contradictions. However, if there is a universal set Ω which one can interpret as "everything under consideration", then $\bigcap_{j\in\emptyset} A_j := \Omega$ looks reasonable. Thus, we define

(2.34) $$\bigcup_{i\in\emptyset} A_i := \emptyset, \text{ always}; \qquad \bigcap_{i\in\emptyset} A_i := \Omega, \text{ if there is a universal set, } \Omega.$$

For nonempty index sets I and J, unions and intersections are monotone:

(2.35) $$I \subseteq J \;\Rightarrow\; \Big[\bigcup_{i\in I} A_i \subseteq \bigcup_{j\in J} A_j, \quad \bigcap_{i\in I} A_i \supseteq \bigcup_{j\in J} A_j \Big].$$

Note that (2.34) respects monotoneness, since (2.35) holds for $I = \emptyset$. □

Remark 2.17. ⋆ For typographical reasons we sometimes use the following notation.

$$\bigcup \big[A_i; i \in I\big] := \bigcup_{i\in I} A_i.$$

Analogous notation exists for $\bigcap$, $\biguplus$ and even summation. For example, assume that $g : \mathbb{R} \to \mathbb{R}$ is some rel–valued function of real numbers, and that the indices of interest are

$$I := \{x \in \mathbb{R} : x > 5 \text{ and } 0 \le g(x) < 5\}.$$

Then $\bigcap_{x\in I} B_x$ can also be expressed as follows:

$$\bigcap_{x\in I} B_x = \bigcap \big[B_x : x > 5 \text{ and } 0 \le g(x) < 5\big] = \bigcap_{x>5 \text{ and } 0\le g(x)<5} B_x. = \bigcap_{\substack{x>5\\ 0\le g(x)<5}} B_x. \;\square$$

Be sure that you understand how to solve the following problem. (Draw a picture!)

Problem 2.1. ⋆ For $a, b \in \mathbb{R}$, let $Q_{(a,b)} := \{(x, y) \in \mathbb{R}^2 : |x - a| \le 3/2, \ |y - b| \le 3/2\}$. Thus, $Q_{(a,b)}$ is the square in the plane with center (a, b) and side length 3. Compute $\bigcap_{(a,b)\in K} Q_{(a,b)}$ and $\bigcup_{(a,b)\in K} Q_{(a,b)}$.

For $K = \{(a, b) \in \mathbb{R}^2 : -1 \le a, b \le 1\}$, compute $\bigcap_{(a,b)\in K} Q_{(a,b)}$ and $\bigcup_{(a,b)\in K} Q_{(a,b)}$.

Solution:

Let $U := \bigcap_{(a,b)\in K} Q_{(a,b)}$ and $V := \bigcup_{(a,b)\in K} Q_{(a,b)}$.

Fix $b_0 \in [-1,1]$ and consider the squares $Q_{(a,b_0)}$ moving from the left ($a = -1$) all the way to the right ($a = +1$). Even $Q_{(-1,b_0)}$ as the leftmost square has x values as big as $1/2$, and $Q_{(1,b_0)}$ as the rightmost square has x values as small as $-(1/2)$, Thus,

$$(x,y) \in \bigcap_{-1\le a\le 1} Q_{(a,b_0)} \quad \Leftrightarrow \quad \left[-\frac{1}{2} \le x \le \frac{1}{2} \quad \text{and} \quad b_0 - \frac{3}{2} \le y \le b_0 + \frac{3}{2}\right].$$

Likewise, if we now also move the squares vertically from $b = -1$ to $b = 1$, then the y values of points in the intersection are exactly those that satisfy $-(1/2) \le y \le 1/2$. Thus,

$$U = \{(x,y) : |x| \le 1/2 \text{ and } |y| \le 1/2\}.$$

One sees in likewise faxhion that the points in the union V are exactly those with x values and y values between $-1-(3/2) = -5/2$ and $1+(3/2) = 5/2$. Thus,

$$V = \{(x,y) : ||x| \le 5/2 \text{ and } |y| \le 5/2\}. \ \square$$

We finish this section with two very useful propositions. The first one (De Morgan) you already have encountered for two sets (see Proposition 2.3 on p.2.3).

Recall for the next theorem that we have defined unions and intersections for arbitrary collections, $(A_j)_{j\in J}$, of sets. See Definition 2.5 on p.20.

Theorem 2.1 (De Morgan's Law)**.** *Let J be an arbitrary, nonempty set. Let $(A_j)_{j\in J}$ be a collection of subsets of a set Ω. Then the complement of the union is the intersection of the complements, and the complement of the intersection is the union of the complements:*

$$(2.36) \qquad \textbf{(a)} \quad \left(\bigcup_{j\in J} A_j\right)^{\complement} = \bigcap_{j\in J} A_j^{\complement}; \qquad \textbf{(b)} \quad \left(\bigcap_{j\in J} A_j\right)^{\complement} = \bigcup_{k} A_k^{\complement};$$

PROOF of De Morgan's law, formula **(a)**: ⋆

1) First we prove that $(\bigcup_\alpha A_\alpha)^{\complement} \subseteq \bigcap_\alpha A_\alpha^{\complement}$:

Assume that $x \in (\bigcup_\alpha A_\alpha)^{\complement}$. Then $x \notin \bigcup_\alpha A_\alpha$ which is the same as saying that x does not belong to any of the A_α. That means that x belongs to each $A_\alpha^{\complement}$ and hence also to the intersection $\bigcap_\alpha A_\alpha^{\complement}$.

2) Now we prove that $(\bigcup_\alpha A_\alpha)^{\complement} \supseteq \bigcap_\alpha A_\alpha^{\complement}$:

Let $x \in \bigcap A_\alpha^{\complement}$. Then x belongs to each of the $A_\alpha^{\complement}$ and hence to none of the A_α. Then it also does not belong to the union of all the A_α and must therefore belong to the complement $(\bigcup_\alpha A_\alpha)^{\complement}$. This completes the proof of formula **(a)**. The proof of **(b)** is similar. ■

Remark 2.18. Note that (2.36) holds true for any index set J. In particular, for finite and infinite sequences of sets. □

Proposition 2.5 (Distributivity of unions and intersections)**.** *Let $(A_n)_n$ be a finite or infinite sequence of sets and let B be a set. Then*

$$\bigcup_j (B \cap A_j) = B \cap \bigcup_j A_j, \tag{2.37}$$

$$\bigcap_{j\in I} (B \cup A_j) = B \cup \bigcap_j A_j. \tag{2.38}$$

PROOF: ■

The next proposition shows how to rewrite any countable union (finite or infinite) as a DISJOINT union.

Proposition 2.6 (Rewrite unions as disjoint unions)**.** *Let $(A_j)_{j\in\mathbb{N}}$ be a sequence of sets which all are contained within the universal set Ω. Let*

$$B_n := \bigcup_{j=1}^{n} A_j = A_1 \cup A_2 \cup \cdots \cup A_n \ (n \in \mathbb{N}),$$
$$C_1 := A_1 = B_1, \quad C_{n+1} := A_{n+1} \setminus B_n \ (n \in \mathbb{N}).$$

Then

(a) *The sequence $(B_j)_j$ is increasing: $m < n \Rightarrow B_m \subseteq B_n$.*

(b) *For each $n \in \mathbb{N}$, $\bigcup_{j=1}^{n} A_j = \bigcup_{j=1}^{n} B_j$. Further, $\bigcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} B_j$.*

(c) *The sets C_j are mutually disjoint, $\bigcup_{j=1}^{n} A_j = \biguplus_{j=1}^{n} C_j$ for all n, and $\bigcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} C_j$.*

(d) *The sets C_j $(j \in \mathbb{N})$ form a partition of the set $\bigcup_{j=1}^{\infty} A_j$.*

PROOF: ⋆ **(a)** and **(b)** are trivial. For the proof of **(c)** and **(d)**, convince yourself that

$$C_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1}).$$

Thus, C_n precisely contains those elements of A_n that have not previously been encountered! ■

2.5 Preimages

Introduction 2.3. The major part of this course will be about functions

$$X : (\Omega, P) \longrightarrow \Omega'; \qquad \omega \mapsto X(\omega)$$

which assign the outcomes (= elements) ω of a probability space (Ω, P) to items $X(\omega) \in \Omega'$. In the context of probability theory, such functions will be called **random elements**. [18] Usually, those function values are numbers or vectors of numbers. In other words, the codomain often is (a subset of) $\mathbb{R}$ or $\mathbb{R}^n$. It is customary to call a real–valued random element

$$Y : (\Omega, P) \longrightarrow B \quad (B \subseteq \mathbb{R}); \qquad \omega \mapsto Y(\omega)$$

a **random variable**. [19]

Let us take another look at Examples 1.3 (Two rolls of a die), 1.4 (Sum of two die rolls), 1.6, and Remark 1.7. This material begins on p.10 of Section 1.2 (A First Look at Probability). There,

- $\Omega = [1,6]^2_{\mathbb{Z}}, \quad \Omega' = [2,12]_{\mathbb{Z}}, \qquad P$ was determined by $P\{\omega\} = \dfrac{1}{36}$ $(\omega \in \Omega)$,
- $Y : \Omega \to \Omega'; \qquad \omega = (\omega_1, \omega_2) \mapsto Y(\omega_1, \omega_2) := \omega_1 + \omega_2$.[20]

The probability space (Ω, P) represent the outcomes of two rolls of a fair die:

- Interpret $\omega = (\omega_1, \omega_2)$ as follows: Die$_1$ yields ω_1, die$_2$ yields ω_2.
  ⊡ Thus, $\omega = (5, 2)$ represents the outcome of die$_1$ giving a 5 and die$_2$ giving a 2.

The function Y was used to "transport" the probability measure P, defined on the powerset of the domain, Ω, to a probability measure P_Y, defined on the powerset of the Codomain, Ω., by means of the formula (1.25) (see Remark 1.7, p.17). We repeat it here:

(2.39) $\qquad P_Y(B) := P\{Y \in B\}$, i.e., $P_Y(B) = P\{\omega \in \Omega : Y(\omega) \in B\}$, for $B \subseteq \Omega'$.

This formula makes those sets so important that they warrant their own definition. □

Since the following definition is of interest not only for probabilistic topics, we now switch from the probabilistic function notation $Y : \Omega \to \Omega'$, to the more familiar $f : X \to Y$.

Definition 2.27.

Let X, Y be two nonempty sets. Let $f : X \to Y$ and $B \subseteq Y$. Then

(2.40) $$f^{-1}(B) := \{x \in X : f(x) \in B\}$$

is a subset of X which we call the **preimage** of B under f. □

Remark 2.19.

(a) ⋆ If we vary $B \subseteq Y$, i.e., $B \in 2^Y$, we can think of the preimage as a function $2^Y \to 2^X$ (since $f^{-1}(B) \in 2^X$).

(b) The symbol f^{-1} is the same for the preimage function $f^{-1} : 2^Y \to 2^X$ and for the ordinary inverse function $f^{-1} : Y \to X$, **if this inverse function exists!** DO NOT CONFUSE THOSE TWO CONCEPTS:
⊡ Arguments and function values of the inverse function are elements of X and Y,
⊡ Arguments and function values of the preimage function are subsets of X and Y.

(c) The preimage $f^{-1}(B)$ exists for any choice of $X, Y, f : X \to Y$, and $B \subseteq Y$, even if the inverse function does not exist! □

[18] See Definition 5.15 (Random element) on p.125

[19] See Definition 5.14 (Random Variables and Random Vectors) on p.124. We are trying to adhere to the probability theory conventions of using capital letters U, V, W, X, Y, Z rather than f, g, h for random elements and in particular the letter Y for random variables.

[20] We often prefer to write ω rather than $\vec{\omega}$ if the the symbol Ω is involved, even if it represents a vector.

□

Example 2.11. This example illustrates the point made in Remark 2.19**(c)**. Let

$$f : \mathbb{R} \to [-1, \infty[\,; \qquad f(x) = x^2 \,.$$

If there was an inverse function, f^{-1}, then its domain must be the codomain of f, and its codomain must be the domain of f. In other words,

$$f^{-1} : [-1, \infty[\to \mathbb{R}\,; \qquad f^{-1}(y) = \sqrt{y}\,.$$

it would have to assign to EACH $y \in [-1, \infty[$ a UNIQUE $x \in \mathbb{R}$ (that x would be $f^{-1}(y)$) such that $f(x) = y$. But such is not the case:

- If $y = -0.5$, then there is no $x \in \mathbb{R}$ such that $x^2 = y$
- If $y = 10$, then there are too many $x \in \mathbb{R}$ such that $x^2 = y$:
  Both $x = \sqrt{10}$ and $x = -\sqrt{10}$ satisfy $x^2 = 10$.
- Note that, for the preimages, we obtain $f^{-1}\big(\{-0.5\}\big) = \emptyset$
  and $f^{-1}\big(\{10\}\big) = \{-\sqrt{10}, \sqrt{10}\}$. Coincidence? □

Example 2.12. For a more extreme example, consider

$$g : [0, \infty[\to \mathbb{R}\,; \qquad g(x) = \sin(x)\,.$$

If $B_1 = [5, 10]$, $B_2 = \{0\}$, what are $g^{-1}(B_1)$ and $g^{-1}(B_2)$? So, does each $y \in \mathbb{R}$ have a unique $x \in [0, \infty[$ such that $g(x) = y$? □

Example 2.13. For an even more extreme example, consider the constant function

$$h : \mathbb{R} \to \mathbb{R}\,; \qquad h(x) = 2\pi\,.$$

If $B_1 = [5, 10]$, $B_2 = \{2\pi\}$, $B_3 = [-500, 5]$, what are $h^{-1}(B_j)(j = 1, 2, 3)$? Again, does each $y \in \mathbb{R}$ have a unique $x \in [0, \infty[$ such that $h(x) = y$? □

Example 2.14. Let

$$h :]0, 3[\to]0, 9[\,; \qquad h(x) = x^2\,.$$

Does h have an inverse? The answer is Yes. The inverse function of h is

$$h^{-1} :]0, 9[\to]0, 3[\,; \qquad h^{-1}(y) = \sqrt{y}\,,$$

since for each $0 < y < 9$, $x = \sqrt{y}$ is the unique solution of the equation $h(x) = y$.
Note the following:

- $h^{-1}(4) = 2$, but $h^{-1}\{4\} = \{2\}$ and NOT 2.
- $h^{-1}(-4)$ does not exist, but $h^{-1}\{-4\} = \emptyset$! □

Notation 2.3 (Notational conveniences for preimages - I)**.**

If we have a set that is written as $\{\dots\}$ then we may write $f^{-1}\{\dots\}$ instead of $f^{-1}(\{\dots\})$.
Specifically for singletons $\{y\}$ such that $y \in Y$, it is OK to write $f^{-1}\{y\}$.
You also are allowed to write $f^{-1}(y)$ instead of $f^{-1}\{y\}$, even though this author thinks that it is not a good idea to confound elements y and subsets $\{y\}$ of Y.

VERY IMPORTANT: Work the following examples closed book and then check that your solutions are correct!

Example 2.15 (Preimages)**.** Let $f : \mathbb{R} \to \mathbb{R}$; $\quad f(x) = x^2$. Determine

a. $f^{-1}(]-4,-2[)$, **b.** $f^{-1}([1,2])$, **c.** $f^{-1}([5,6])$, **d.** $\{-4 < f < -2$ **or** $1 \le f \le 2$ **or** $5 \le f < 6]\}$.

Solution:

a. $f^{-1}(]-4,-2[) = \{\, x \in \mathbb{R} : x^2 \in]-4,-2[\,\} = \{\, -4 < f < -2 \,\} = \emptyset$.

b. $f^{-1}([1,2]) = \{\, x \in \mathbb{R} : x^2 \in [1,2] \,\} = \{\, 1 \le f \le 2 \,\} = [-\sqrt{2},-1] \cup [1,\sqrt{2}]$.

c. $f^{-1}([5,6]) = \{\, x \in \mathbb{R} : x^2 \in [5,6] \,\} = \{\, 5 \le f \le 6 \,\} = [-\sqrt{6},-\sqrt{5}] \cup [\sqrt{5},\sqrt{6}]$.

d. $\{-4 < f < -2$ **or** $1 \le f \le 2$ **or** $5 \le f < 6]\} = f^{-1}(]-4,-2[\cup [1,2] \cup [5,6])$
$= \{\, x \in \mathbb{R} : x^2 \in]-4,-2[$ or $x^2 \in [1,2]$ **or** $x^2 \in [5,6] \,\}$
$= [-\sqrt{2},-1] \cup [1,\sqrt{2}] \cup [-\sqrt{6},-\sqrt{5}] \cup [\sqrt{5},\sqrt{6}]$. □

Example 2.16 (Preimages)**.** Let $f : \mathbb{R} \to \mathbb{R}$; $\quad f(x) = x^2$. Determine

a. $f^{-1}(]-4,2[)$, **b.** $f^{-1}([1,3])$, **c.** $\{-4 < f < 2$ **and** $1 \le f \le 3\}$.

Solution:

a. $f^{-1}(]-4,2[) = \{\, x \in \mathbb{R} : x^2 \in]-4,2[\,\} = \{\, x \in \mathbb{R} : -4 < x^2 < 2 \,\} =]-\sqrt{2},\sqrt{2}[$.

b. $f^{-1}([1,3]) = \{\, x \in \mathbb{R} : x^2 \in [1,3] \,\} = \{\, x \in \mathbb{R} : 1 \le x^2 \le 3 \,\} = [-\sqrt{3},-1] \cup [1,\sqrt{3}]$.

c. $\{-4 < f < 2$ **and** $1 \le f \le 3\} = f^{-1}(]-4,2[\cap [1,3])$
$= \{\, x \in \mathbb{R} : x^2 \in]-4,2[$ **and** $x^2 \in [1,3] \,\}$
$= \{\, x \in \mathbb{R} : 1 \le x^2 < 2 \,\} =]-\sqrt{2},-1] \cup [1,\sqrt{2}[$. □

Proposition 2.7. *Some simple properties:*

(2.41) $\quad f^{-1}(\emptyset) = \emptyset$

(2.42) $\quad B_1 \subseteq B_2 \subseteq Y \Rightarrow f^{-1}(B_1) \subseteq f^{-1}(B_2)$ $\quad$ (***monotonicity of*** $f^{-1}\{\dots\}$)

(2.43) $\quad f^{-1}(Y) = X$ $\quad$ *always!*

PROOF of 2.42:
We show that $x \in f^{-1}(B_1) \Rightarrow f^{-1}(B_1)$ as follows.

$$x \in f^{-1}(B_1) \overset{(a)}{\Rightarrow} f(x) \in B_1 \overset{(b)}{\Rightarrow} f(x) \in B_2 \overset{(c)}{\Rightarrow} x \in f^{-1}(B_2)$$

In the above, (a) and (c) state the definition of a preimage and (b) follows from $B_1 \subseteq B_2$
The proof of of 2.41 and 2.42 is left as an exercise. ■

Notation 2.4 (Notational conveniences for preimages - II).

If we have a set that is written as $\{\dots\}$ then we may write $f^{-1}\{\dots\}$ instead of $f^{-1}(\{\dots\})$.
Specifically for singletons $\{y\}$ such that $y \in Y$, it is OK to write $f^{-1}\{y\}$.
You also are allowed to write $f^{-1}(y)$ instead of $f^{-1}\{y\}$, even though this author thinks that it is not a good idea to confound elements y and subsets $\{y\}$ of Y.

Example 2.17. Consider the random variable $Y : (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2$ of the introduction to this section.

- $P_Y(\{10\}) = P\big(\{(\omega_1,\omega_2) \in \Omega : Y(\omega_1,\omega_2) = 10\}\big)$
  can be written $P_Y(\{10\}) = P\big(Y^{-1}\{10\}\big) = P\{Y = 10\}$.
- $P_Y(\{(\omega')\}) = P\big(\{(\omega_1,\omega_2) \in \Omega : Y(\omega_1,\omega_2) = \omega'\}\big)$.
  can be written $P_Y(\{\omega'\}) = P\big(Y^{-1}\{\omega'\}\big) = P\{Y = \omega'\}$.
- $P_Y(B) = P\big(\{\omega \in \Omega : Y(\omega) \in B\}\big)$
  can be written $P_Y(B) = P\big(Y^{-1}(B)\big) = P\{Y \in B\}$.

It is very important that you **remember the first four** of the six formulas of the Theorem 2.2 below. In the proof of Theorem 5.11 on p.123 they show the following: For a function $X : (\Omega, P) \to \Omega'$, the assignment

$$A \mapsto P_X(A) := P\left(f^{-1}(A)\right)$$

defines a probability measure [21] on Ω'.

Theorem 2.2 (f^{-1} is compatible with all basic set ops). *Assume that X, Y be nonempty, $f : X \to Y$, J is an arbitrary index set.* [22] *Further assume that $B \subseteq Y$ and that $B_j \subseteq Y$ for all j. Then*

$$f^{-1}\Big(\bigcap_{j\in J} B_j\Big) = \bigcap_{j\in J} f^{-1}(B_j) \tag{2.44}$$

$$f^{-1}\Big(\bigcup_{j\in J} B_j\Big) = \bigcup_{j\in J} f^{-1}(B_j) \tag{2.45}$$

$$f^{-1}(B^{\complement}) = \big(f^{-1}(B)\big)^{\complement} \tag{2.46}$$

$$B_1 \cap B_2 = \emptyset \;\Rightarrow\; f^{-1}(B_1) \cap f^{-1}(B_2) = \emptyset. \tag{2.47}$$

$$f^{-1}(B_1 \setminus B_2) = f^{-1}(B_1) \setminus f^{-1}(B_2) \tag{2.48}$$

$$f^{-1}(B_1 \Delta B_2) = f^{-1}(B_1) \Delta f^{-1}(B_2) \tag{2.49}$$

Note that (2.47) implies that the preimages of a disjoint family form a disjoint family.

PROOF: ⋆ MF330 notes, ch.8 ■

Proposition 2.8 (Preimages of function composition). *Let X, Y, Z be arbitrary, nonempty sets.*

[21] the so–called distribution of X with respect to P. See Definition 5.13 (Probability Distribution) on p.124.

[22] If you have problems with the concept of a family, think of J as a set of integers which are bounded below, i.e., that J is the index set of a finite or infinite sequence or subsequence of sets

Let $f : X \to Y$ and $g : Y \to Z$ and $h : X \to Z$ the composition

$$h(x) \;=\; g \circ f(x) \;=\; g\big(f(x)\big).$$

Let $U \subseteq X$ and $W \subseteq Z$. Then

(2.50) $\quad (g \circ f)^{-1} \;=\; f^{-1} \circ g^{-1}$, *i.e.,* $\quad (g \circ f)^{-1}(W) \;=\; f^{-1}\big(g^{-1}(W)\big)$ *for all* $W \subseteq Z$.

PROOF: ⋆ MF330 notes, ch.8 ■

Try to understand the above with a simple example, such as $X = Y = R$, $f(x) = 3x - 1, g(y) = y^2$, and $W = [0, 1]$, $W = \{-10\}$ $W = \{10\}$ (three different choices for W).

Given a function $f : X \to Y$, the preimage acts as a set function which assigns subsets B of the codomain to the subsets $f^{-1}(B)$ of the domain. There is a "dual" definition which goes the other way: It assigns set in the domain to sets in the codomain.

Definition 2.28 (Direct image). ⋆

Let X, Y be two nonempty sets and $f : X \to Y$. Let $A \subseteq X$. We call the set

(2.51) $$f(A) \;:=\; \{f(a) : a \in A\}\,.$$

which consists of all function values of arguments in A, the **direct image** of A under f. □

Note that the range $f(X)$ of f (see Definition 2.17 (Function) on p.33) is a special case of a direct image.

Notation 2.5 (Notational conveniences for direct images).

As we do for preimages, if we deal with a set that is written as $\{\dots\}$, then we may write $f\{\dots\}$ instead of $f(\{\dots\})$. In particular, we can write $f\{x\}$ for singletons $\{x\} \subseteq X$. □

The same symbol f is used for the original function $f : X \to Y$ and the direct image which we can think of as a function

$$2^X \;\to\; 2^Y\,; \qquad A \;\mapsto\; f(A) \;=\; \{f(a) : a \in A\}\,, \quad (A \subseteq X)\,.$$

Be careful not to let this confuse you! □

Example 2.18 (Direct images). ⋆ Let $f : \mathbb{R} \to \mathbb{R}$; $\quad f(x) = x^2$.

(a) $f(]-4,-2[) \;=\; \{\, x^2 : x \in\,]-4,-2[\,\} \;=\; \{\, x^2 : -4 < x < -2\,\} \;=\;]4, 16[$.
(b) $f([1,2]) \;=\; \{\, x^2 : x \in [1,2]\,\} \;=\; \{\, x^2 : 1 \le x \le 2\,\} \;=\; [1,4]$.
(c) $f([5,6]) \;=\; \{\, x^2 : x \in [5,6]\,\} \;=\; \{\, x^2 : 5 \le x \le 6\,\} \;=\; [25,36]$.
(d) $f(]-4,-2[\cup [1,2] \cup [5,6]) \;=\; \{\, x^2 : x \in\,]-4,-2[$ or $x \in [1,2]$ or $x \in [5,6]\,\}$
$=]4, 16[\cup [1,4] \cup [25,36] \;=\; [1, 16[\cup [25,36]$. □

Example 2.19 (Direct images). ⋆ Let $f : \mathbb{R} \to \mathbb{R}; \quad f(x) = x^2$.

(a) $f(]-4,2[) = \{ x^2 : x \in]-4,2[\} = \{ x^2 : -4 < x < 2 \} =]4,16[$.
(b) $f([1,3]) = \{ x^2 : x \in [1,3] \} = \{ x^2 : 1 \le x \le 3 \} = [1,9]$.
(c) $f(]-4,2[\cap [1,3]) = \{ x^2 : x \in]-4,2[\text{ and } x \in [1,3] \} = \{ x^2 : 1 \le x < 2 \} = [1,4[$. □

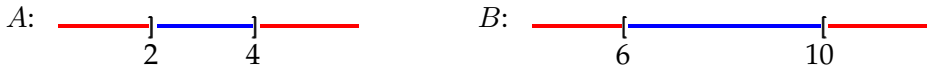
2.6 Infimum and Supremum: Generalized Minimum and Maximum

Introduction 2.4. Let $A :=]2,4]$ and $B := [6,10[$. Then A possesses 4 as its maximum, and the minimum of B is 6.

It is just as obvious that 2 plays a role for A very similar to the one that $\min(B) = 6$ plays for B, and that 10 plays a role for B very similar to the one that $\max(A) = 4$ plays for A.

But is $\min(A) = 2$ and $\max(B) = 10$? The answer is NO: The minimum and the maximum of a set must belong to that set, and neither is $2 \in A$, nor is $10 \in B$.

Let us find some appropriate names for those two numbers. In the pictures below the sets A and B are colored blue and their upper and lower bounds are colored red.



Upper bounds are those numbers so far "up" to the right that they dominate each item in the set. For example, 20 is an upper bound of both A and B, since $20 \ge a$ for each $a \in A$ and $20 \ge b$ for each $b \in B$. Clearly any $x > 20$ also is an upper bound for both A and B.

What about $x = 2\pi \approx 6.28$? That one is smaller than 20 but still an upper bound of A, since $2\pi \ge a$ for each $a \in A$. However, it is not an upper bound of B since, e.g., $7.5 \in B$ and $6 > 7.5$ is false.

Lower bounds are the opposite of upper bounds. They are so far "down" to the left that they are dominated by each item in the set. For example, $-\sqrt{2}$ is a lower bound of both A and B, since $-\sqrt{2} \le a$ for each $a \in A$ and $-\sqrt{2} \le b$ for each $b \in B$. Clearly any $x < -\sqrt{2}$ also is a lower bound for both A and B. Matter of fact, any negative number is a lower bound of both A and B.

What about $x = \pi$? That one is larger than $-\sqrt{2}$ but still a lower bound of B, since $\pi \le b$ for each $b \in B$. However, π is too large for a lower bound of A. For example, $3 \in A$ and $\pi < 3$ is false.

Our goal was to find appropriate names for 2 in relation to A and for 10 in relation to B.

2 is similar to $6 = \min(B)$ in the following sense:

- 2 is a lower bound of A, just as $\min(B)$ is a lower bound of B
- Not only that, but 2 is the GREATEST lower bound of A, just as $\min(B)$ is the greatest lower bound of B

Similarly, 10 is similar to $4 = \max(A)$ in the following sense:

- 10 is an upper bound of B, just as $\max(A)$ is an upper bound of A.
- Not only that, but 10 is the LEAST (smallest) upper bound of A, just as $\max(A)$ is the least upper bound of A.

In summary, greatest lower bound and least upper bound or something equivalent seems to be bood names. □

We give mathematical precision to our findings in the next definition. You will not be asked to

recite it from memory, but you are expected to determine the min/max/inf/sup of a given set of real numbers.

Definition 2.29 (Minimum, maximum, infimum, supremum). ⋆ Let $A \subseteq \mathbb{R}$, $A \neq \emptyset$, and let l and u be real numbers.

(a) We call l a **lower bound** of A if $l \leq a$ for all $a \in A$.
(b) We call u an **upper bound** of A if $u \geq a$ for all $a \in A$.
(c) We call A **bounded above** if this set has an upper bound.
(d) We call A **bounded below** if A has a lower bound.
(e) We call A **bounded** if A is both bounded above and bounded below.

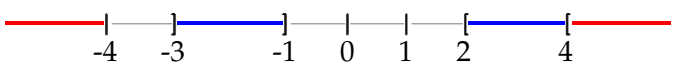
(f) The **minimum** of A, if it exists, is the unique lower bound l of A such that $l \in A$.
(g) A **maximum** of A, if it exists, is the unique upper bound u of A such that $u \in A$.

Since they are uniquely determined by A, we may write $\min(A)$ for the minimum of A and $\max(A)$ for the maximum of A.

(h) If A is bounded below (i.e., A has lower bounds), we call the maximum of those bounds the **infimum** of A. Thus, it is the **greatest lower bound** of A. We write $\inf(A)$ or g.l.b.(A). Otherwise (A is not bounded below), we define $\inf(A) := -\infty$.
(i) If A is bounded above (i.e., A has upper bounds), we call the minimum of those bounds the **supremum** of A. Thus, it is the **least upper bound** of A. We write $\sup(A)$ or l.u.b.(A). Otherwise (A is not bounded above), we define $\sup(A) := \infty$. □

Problem 2.2. Let $A =]-3,-1] \cup [2,4[\cup \{-4,0,1\}$. Determine $\min(A)$, $\max(A)$, $\inf(A)$, $\sup(A)$.

Solution: In the picture below the segments belonging to A are colored blue, upper and lower bounds are colored red,



lower bounds of $A =]-\infty,-4]$
upper bounds of $A = [4,\infty[$

- $\inf(A) = \min(A) = -4 =$ greatest lower bound $= \max\{$ lower bounds $\}$
- $\sup(A) = 4 =$ least upper bound $= \min(\{$ upper bounds $\}$; $\max(A) =$ DNE, since $4 \notin A$

Remark 2.20. Here is the cookbook approach to infima and suprema. (NOT OPTIONAL!)

- Infima are generalized minima and suprema are generalized maxima.
- Think of $\inf(A)$ as a minimum that does not need to belong to A.
- Traverse the lower bounds of A from the left (from $-\infty$) to the rigt until you "hit" A. That's the greatest lower bound. That's $\inf(A)$.
- Think of $\sup(A)$ as a maximum that does not need to belong to A.
- Traverse the upper bounds of A from the right ($+\infty$) to the left until you "hit" A. That's the least (smallest) upper bound. That's $\sup(A)$. □

The min/max/inf/sup of a function or family or sequence which takes values in $\mathbb{R}$, is the min/max/inf/sup of the set of all values that this entity can have. We give explicit definitions of the notation of those items only for infimum and supremum. It is obvious how to define their maximum and minimum. (But remember: max and min are not guaranteed to exist!)

Definition 2.30. ⋆ This is marked optional, but be sure you can work with the most common notation introduced here, including the counterparts for min and max!

Let X be an arbitrary set (need not be numbers or elements of $\mathbb{R}^d$!) and $A \subseteq X$.
Let $f : X \to \mathbb{R}$ be real–valued. The **supremum** and **infimum** of f on A are defined as

$$\sup_A f := \sup_{x\in A} f(x) := \sup\{f(x) : x \in A\} \tag{2.52}$$

$$\inf_A f := \inf_{x\in A} f(x) := \inf\{f(x) : x \in A\}. \tag{2.53}$$

The **supremum** and **infimum** of a family of real numbers $(x_i)_{i\in I}$ $(x_i)_{i\in I}$ are defined as

$$\sup(x_i) := \sup_i (x_i) := \sup(x_i)_i := \sup(x_i)_{i\in I} := \sup_{i\in I} x_i := \sup\{x_i : i \in I\}. \tag{2.54}$$

$$\inf(x_i) := \inf_i (x_i) := \inf(x_i)_i := \inf(x_i)_{i\in I} := \inf_{i\in I} x_i := \inf\{x_i : i \in I\}. \quad \square \tag{2.55}$$

The definition above for families extends to sequences x_n, defined for $n = n_*, n_* + 1, n_* + 2, \ldots$.

The **supremum** and **infimum** of a sequence of real numbers $(x_n)_{n\geq n_*}$ are defined as

$$\sup(x_n) := \sup(x_n)_{n\geq n_*} := \sup_{n\geq n_*} x_n = \sup\{x_n : n = n_*, n_* + 1, n_* + 2, \ldots\} \tag{2.56}$$

$$\inf(x_n) := \inf(x_n)_{n\geq n_*} := \inf_{n\geq n_*} x_n = \inf\{x_n : n = n_*, n_* + 1, n_* + 2, \ldots\} \quad \square \tag{2.57}$$

Problem 2.3.

(a) Let $f(x) := |\sin x|$; Determine min, max, inf and sup of f on $\mathbb{R} \setminus \{k\pi : k \in \mathbb{Z}\}$.

(b) Let $(x_\alpha)_{\alpha\in J}$ be the family defined by $x_\alpha := \cos\alpha$; $J := \mathbb{R} \setminus \{k\pi : k \in \mathbb{Z}\}$.
Determine min, max, inf and sup of $(x_\alpha)_{\alpha\in J}$.

(c) Let $(a_n := \dfrac{n}{n+1}; n = 0, 1, 2, \ldots$. Determine min, max, inf and sup of the sequence $(a_n)_{n=0}^{\infty}$.

Solution:

(a) Let $A := \mathbb{R} \setminus \{k\pi : k \in \mathbb{Z}\}$. Then $\max_A f = \sup_{x\in A} f(x) = 1$,
$\min\{f(x) : x \in A\}$ DNE, $\inf\{f(x) : x \in \mathbb{R}$ and $x \neq k\pi$ for $k \in \mathbb{Z}\} = 0$.

(b) $\min(x_\alpha)_{\alpha\in J}$ DNE, $\max_{\alpha\in J} x_\alpha$ DNE, $\inf_\alpha(x_\alpha) = -1$, $\sup\{x_\alpha : \alpha \in J\} = 1$.

(c) $\min(a_n) = \inf_{n\geq 0} a_n = 0$, $\max_{n=0}^{\infty} a_n$ DNE, $\sup\{x_n : n = 0, 1, 2, \ldots\} = 1$. $\square$

Theorem 2.3. ⋆

Let $\alpha_1 \geq \alpha_2 \geq \cdots$ be a nonincreasing sequence and $\beta_1 \leq \beta_2 \leq \cdots$ a nondecreasing sequence of real numbers. Then

(a) $\lim_{n\to\infty} \alpha_n$ *exists (might be $-\infty$) and equals* $\inf_{n\in\mathbb{N}} \alpha_n$.

(b) $\lim_{n\to\infty} \beta_n$ *exists (might be ∞) and equals* $\sup_{n\in\mathbb{N}} \beta_n$.

Let $\emptyset \neq A \subseteq \mathbb{R}$ and $f_n, g_n : A \to \mathbb{R}$ two sequences of real–valued functions on A, such that

$(f_n)_n$ *is nonincreasing, i.e.,* $f_1 \geq f_2 \geq \cdots$, *i.e.,* $f_1(x) \geq f_2(x) \geq \cdots$, *for all* $x \in A$,

$(g_n)_n$ *is nonincreasing, i.e.,* $g_1 \leq g_2 \leq \cdots$, *i.e.,* $g_1(x) \leq g_2(x) \leq \cdots$, *for all* $x \in A$,

Then

(c) $x \to \lim_{n\to\infty} f_n(x)$ *exists (might be $-\infty$ for some or all $x \in A$) and equals* $x \to \inf_{n\in\mathbb{N}} f_n(x)$.

(d) $x \to \lim_{n\to\infty} g_n(x)$ *exists (might be ∞ for some or all $x \in A$) and equals* $x \to \sup_{n\in\mathbb{N}} g_n(x)$.

PROOF: Will not be given here. Note though, that **(c)** follows from **(a)** and **(d)** follows from **(b)**, simply by freezing x and examining the sequences of numbers, $\alpha_n := f_n(x)$ and $\beta_n := g_n(x)$. ■

Example 2.20. Let $f_n :]-1, 0] \to \mathbb{R};\ f_n(x) := \sum_{j=0}^{n} x^n$, and $g_n : [0, \infty[\to \mathbb{R};\ g_n(x) := \sum_{j=0}^{n} x^n$. (Same function (geometric series with quotient x), but different domains!) Then

$$\begin{aligned}
\lim_{n\to\infty} f_n(x) &= \lim_{n\to\infty} \frac{1-x^{n+1}}{1-x} \downarrow \frac{1}{1-x} = \inf_{n\geq 0} f_n(x) \quad (\downarrow, \text{since } -x^{n+1} \geq 0)\,, \\
\lim_{n\to\infty} g_n(x) &= \lim_{n\to\infty} \frac{1-x^{n+1}}{1-x} \uparrow \frac{1}{1-x} = \sup_{n\geq 0} g_n(x) \text{ for } 0 \leq x < 1, \\
&\uparrow \infty = \sup_{n\geq 0} g_n(x) \text{ for } x \geq 1 \text{ (since } x^n \geq 1). \ \square
\end{aligned}$$

2.7 Cartesian Products

We next define cartesian products of sets. Those mathematical objects generalize rectangles

$$[a_1, b_1] \times [a_2, b_2] = \{(x, y) : x, y \in \mathbb{R}, a_1 \leq x \leq b_1 \text{ and } a_2 \leq y \leq b_2\}$$

and quads

$$[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] = \{(x, y, z) : x, y, z \in \mathbb{R}, a_1 \leq x \leq b_1, a_2 \leq y \leq b_2 \text{ and } a_3 \leq z \leq b_3\}.$$

which you certainly have encountered in multivariable calculus.

Definition 2.31 (Cartesian Product)**.** Let X and Y be two sets The set

$$X \times Y := \{(x, y) : x \in X, y \in Y\} \tag{2.58}$$

is called the **cartesian product** of X and Y. We write X^2 as an abbreviation for$X \times X$.

Note that the order is important: (x, y) and (y, x) are different unless $x = y$.

This definition generalizes to more than two sets as follows:

Let $X_1, X_2, \ldots, X_n$ be sets. The set

$$X_1 \times X_2 \cdots \times X_n := \{(x_1, x_2, \ldots, x_n) : x_j \in X_j \text{ for each } j = 1, 2, \ldots n\} \tag{2.59}$$

is called the cartesian product of $X_1, X_2, \ldots, X_n$.
We write X^n as an abbreviation for$X \times X \times \cdots \times X$.

Example 2.21. In your multivariable calculus course you have learned about twodimensional vectors and threedimensional vectors. Convenient notations would often be

$$(x, y) \in \mathbb{R}^2\,, \quad (a, b) \in \mathbb{R}^2\,, \quad (x, y, z) \in \mathbb{R}^3\,, \quad (a, b, c) \in \mathbb{R}^3\,. \tag{2.60}$$

Note that those vectors are elements of the cartesian products $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

In general, any finite list of real numbers $(\beta_1, \beta_2, \ldots, \beta_d)$ is an element of $\mathbb{R}^d$ which we call a d–dimensional **vector** of real numbers. You probably are used to write $\mathbb{R}^n$ rather than $\mathbb{R}^d$. We choose the letter d (first letter of "dimension"), to keep the symbol n free for other purposes, such as denoting the size of a sample.

Here is an example.

$$(8, -3, 0, 4, -7)$$

is a 5–dimensional vector of Integers. Since integers are special cases of rational numbers which themselves are also real numbers, this vector is an element of each one of $\mathbb{Z}^5, \mathbb{Q}^5, \mathbb{R}^5$.

The notation used in (2.60) does not scale for higher dimensional vectors. On the other hand, the expression $(\beta_1, \beta_2, \ldots, \beta_d)$ is very suitable. However, this is very lengthy notation, so we use the symbol for the subscripted components (that's β) and write an arrow on top of that symbol to indicate that we are dealing with a vector. [23]

We will use this arrow notation for vectors very frequently. Here are some examples.

$$\vec{x} = (x_1, x_2, \ldots, x_n)\,, \quad \vec{b} = (b_1, b_2, b_3, b_4)\,, \quad \vec{Z} = (Z_1, Z_2, \ldots, Z_d)\,.$$

Assuming that each subscripted item belongs to $\mathbb{R}$ we have $\vec{x} \in \mathbb{R}^n, \vec{b} \in \mathbb{R}^4, \vec{Z} \in \mathbb{R}^d$. □

Notation 2.6. Notational conveniences for vectors: Unless something else is stated, we will always assume the following. If X is a nonempty set (usually, X is a set of numbers),

$\vec{x} \in X^d$ is shorthand for $\vec{x} = (x_1, x_2, \ldots, x_d) \in X^d$ (i.e., $x_j \in X$ for $j = 1, 2, \ldots, d$.)

We extend this convention to the case $X_1 \times \cdots \times X_d$ with potentially different sets X_j. □

[23] We borrow that notation from physics.

This is best explained by example.

Example 2.22. Let $a_1 < b_1$, $a_2 < b_2$, $\ldots$, $a_d < b_d$, be d pairs of numbers ($d \in \mathbb{N}$). We apply the notation established above to $X_j :=]a_j, b_j]$ and see that

$$\begin{aligned} \vec{y} \in]a_1, b_1\,] \times \cdots \times]a_d, b_d\,] \qquad & \text{is shorthand for} \\ \vec{y} \;=\; \big(y_1, y_2, \ldots, y_d\big), \;\; & \text{where } a_i < y_i \leq b_i, \text{ for } i = 1, \ldots, d. \end{aligned}$$

It is customary to call sets of the form

- ⊡ $]a_1, b_1[\times \cdots \times]a_d, b_d[$, ⊡ $]a_1, b_1\,] \times \cdots \times]a_d, b_d\,]$,
- ⊡ $[\,a_1, b_1[\times \cdots \times \,[a_d, b_d[$, ⊡ $[\,a_1, b_1\,] \times \cdots \times \,[a_d, b_d\,]$,

d**–dimensional rectangles**. □

Example 2.23. Cartesian products occur in a natural manner in probability theory when one models the outcomes of repeated experiments.

(a) If the experiment is three rolls of a die, then the set

$$\Omega \;=\; \big([1,6]_{\mathbb{Z}}\big)^3 \;=\; \{1,2,3,4,5,6\}^3$$

is a natural container for the outcomes of this experiment. For example, $(4,2,6) \in \Omega$ is the outcome of having rolled a 4 followed by a 2 followed by a 6.

(b) n tosses of a coin ($n \in \mathbb{N}$) are modeled as follows. Let H stand for Heads and T for Tails. Then let

$$\Omega \;=\; \{H, T\}^n$$

For example, if $n = 5$, then $(H, H, T, H, T) \in \Omega$ models the outcome of having tossed Heads followed by Heads followed by Tails followed by Heads followed by Tails. This example demonstrates that cartesian products are also defined for sets that do not necessarily consist of numbers □

Here is an abstract example.

Example 2.24. The graph Γ_f of a function with domain X and codomain Y (see def.2.32) is a subset of the cartesian product $X \times Y$. □

Proposition 2.9. *Let X_1, X_2, X_n be finite, nonempty sets. Then,*

The size of the cartesian product is the product of the sizes of its factors, i.e.,

$$\big|\, X_1 \times X_2 \times \cdots \times X_n \,\big| \;=\; \big|X_1\big| \cdot \big|X_2\big| \cdot \big|X_3\big| \cdots \big|X_n\big|\,. \tag{2.61}$$

PROOF:

Case $n = 2$: This trivial for two sets, since the proposition simply states that a matrix (a rectangular grid) of m rows and n columns possesses mn entries.

Case $n = 3$: For three sets X_1, X_2, X_3, we arrange the $|X_1|\cdot|X_2|$ entries of $X_1 \times X_2$ into a single row. In other words, we consider the members $(x_i^{(1)}, x_j^{(2)}, x_k^{(3)})$ of $X_1 \times X_2 \times X_3$ as members $\big((x_i^{(1)}, x_j^{(2)}), x_k^{(3)}\big)$

of $\big(X_1 \times X_2\big) \times X_3$. We apply the result for two sets to the cartesian product of $X_1 \times X_2$ and X_3 and obtain

$$\big|X_1 \times X_2 \times X_3\big| \;=\; \big|(X_1 \times X_2) \times X_3\big| \;=\; \big|X_1 \times X_2\big| \cdot \big|X_3\big| \,. \;=\; \big|X_1\big| \cdot \big|X_2\big| \cdot \big|X_3\big| \,.$$

We repeat this procedure for $n = 3, 4, 5, \ldots$ sets.

Case n: We arrange the elements of $X_1 \times X_2 \times \times X_{n-1}$ into a single row and

$$\text{interpret each} \quad (x_1, \ldots, x_n) \in X_1 \times X_n \quad \text{as} \quad \big((x_1, \ldots, x_{n-1}), x_n\big) \in (X_1 \times X_{n-1}) \times X_n.$$

Thus, the sets $X_1 \times X_n$ and $(X_1 \times X_{n-1}) \times X_n$ have the same size. We know from the prior step, case $n-1$, that $\big|X_1 \times \cdots \times X_{n-1}\big| \;=\; \big|X_1\big| \cdots \big|X_{n-1}\big|$. Hence,

$$\begin{aligned} \big|\, X_1 \times \cdots \times X_n \,\big| \;&= \big|\, (X_1 \times \cdots X_{n-1}) \times X_n \,\big| \;=\; \big(\big|\, X_1 \times \cdots X_{n-1} \,\big|\big) \cdot \big|X_n\big| \\ &= \big(\big|X_1\big| \cdots \big|X_{n-1}\big|\big)\big|X_n\big| \;=\; \big|X_1\big| \cdot \big|X_2\big| \cdot \big|X_3\big| \cdots \big|X_n\big| \,. \;\blacksquare \end{aligned}$$

2.8 Indicator Functions

Indicator functions often are a great notational convenience, for example, when dealing with functions that are defined differently in two or more parts of the domain.

Definition 2.32 (indicator function for a set)**.**

Let Ω be a nonempty set and $A \subseteq \Omega$. Let $\mathbf{1}_A : \Omega \to \{0, 1\}$ be the function defined as

$$\mathbf{1}_A(\omega) \;:=\; \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{2.62}$$

$\mathbf{1}_A$ is called the **indicator function** of the set A. [24] $\square$

Example 2.25. The following examples demonstrate the usefulness of indicator functions.

(a) Let $f : \mathbb{R} \to \mathbb{R}$ be the function

$$f(x) \;:=\; \begin{cases} 3x & \text{if } -10 < x \le 0, \\ \sin(7x) & \text{if } 2 \le x \le 4, \\ 4x^3 + 6, & \text{if } x > 10, \\ 0, & \text{else.} \end{cases}$$

More compactly, $f(x) \;=\; 3x \cdot \mathbf{1}_{]-10,0]} \;+\; \sin(7x) \cdot \mathbf{1}_{[2,4]} \;+\; (4x^3 + 6) \cdot \mathbf{1}_{]10,\infty[}$.

(b) The so-called density function of the exponential distribution with parameter $\beta > 0$ is [25]

$$f(y) \;=\; \begin{cases} \frac{1}{\beta}\, e^{-y/\beta}\,, & 0 \le y < \infty\,, \\ 0\,, & \text{elsewhere}\,. \end{cases}$$

This can also be written as $f(y) \;=\; \frac{1}{\beta}\, e^{-y/\beta}\, \mathbf{1}_{[0,\infty[}(y)$.

[25] See definition 10.12 (Exponential distribution) on p.235.

(c) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function $f(x,y) := 2x^2 - xy$. Let $A := \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 9\}$, i.e., A is the circle centered at the origin with radius 3. Recall from (multivariable) calculus that the integral of f on the set A is defined (by means of Riemann sums) as follows: $\int_A f(x,y)d(x,y) = \int_R \mathbf{1}_A(x,y)f(x,y)d(x,y)$. Here, R is some rectangle $[a_1,b_1]\times[a_1,b_1]$ big enough to contain A. [26] For example, once could chose $R = [-3,3]\times[-4,8]$. □

Proposition 2.10. *Let $A_1, A_2, \ldots$ be subsets of Ω. Then*

$$A_1 \subseteq A_2 \;\Rightarrow\; \mathbf{1}_{A_1} \le \mathbf{1}_{A_2}, \tag{2.63}$$

$$\mathbf{1}_{A_1\cap A_2} = \min(\mathbf{1}_{A_1}, \mathbf{1}_{A_2}), \qquad \mathbf{1}_{\bigcap[A_n : n\in\mathbb{N}]} = \inf_{n\in\mathbb{N}} \mathbf{1}_{A_n}, \tag{2.64}$$

$$\mathbf{1}_{A_1\cup A_2} = \max(\mathbf{1}_{A_1}, \mathbf{1}_{A_2}), \qquad \mathbf{1}_{\bigcup[A_n : n\in\mathbb{N}]} = \sup_{n\in\mathbb{N}} \mathbf{1}_{A_n}, \tag{2.65}$$

$$\mathbf{1}_{A_1^{\complement}} = 1 - \mathbf{1}_{A_1}, \tag{2.66}$$

$$\mathbf{1}_{A_1\uplus A_2} = \mathbf{1}_{A_1} + \mathbf{1}_{A_2}\,, \qquad \mathbf{1}_{\biguplus[A_n : n\in\mathbb{N}]} = \sum_{n\in\mathbb{N}} \mathbf{1}_{A_n}\,, \quad (A_1, A_2, \ldots \textit{ disjoint}). \tag{2.67}$$

PROOF: The proof is an easy exercise.

2.9 Exercises for Ch.2

2.9.1 Exercises for Sets

Exercise 2.1. Prove (2.12) of prop.2.2 on p.23.

Exercise 2.2. Prove the set identities of prop.2.1.

Exercise 2.3. Prove that for any three sets A, B, C it is true that $(A \setminus B) \setminus C = A \setminus (B \cup C)$.
Hint: use De Morgan's formula (2.13.**a**). ■

Exercise 2.4. Let $X = \{x, y, \{x\}, \{x,y\}\}$. True or false?

a. $\{x\} \in X$ **c.** $\{\{x\}\} \in X$ **e.** $y \in X$ **g.** $\{y\} \in X$
b. $\{x\} \subseteq X$ **d.** $\{\{x\}\} \subseteq X$ **f.** $y \subseteq X$ **h.** $\{y\} \subseteq X$ □

For the subsequent exercises refer to Definition 2.11 on p.25 of the size $|A|$ of a set A and to Definition 2.31 on p.50 of Cartesian products.

Exercise 2.5. Find the size of each of the following sets:

a. $A = \{x, y, \{x\}, \{x,y\}\}$ **c.** $C = \{u, v, v, v, u\}$ **e.** $E = \{\sin(k\pi/2) : k \in \mathbb{Z}\}$
b. $B = \{1, \{0\}, \{1\}\}$ **d.** $D = \{3z - 10 : z \in \mathbb{Z}\}$ **f.** $F = \{\pi x : x \in \mathbb{R}\}$ □

Exercise 2.6. Let $X = \{x, y, \{x\}, \{x,y\}\}$ and $Y = \{x, \{y\}\}$. True or false?

a. $x \in X \cap Y$ **c.** $x \in X \cup Y$ **e.** $x \in X \setminus Y$ **g.** $x \in X \Delta Y$
b. $\{y\} \in X \cap Y$ **d.** $\{y\} \in X \cup Y$ **f.** $\{y\} \in X \setminus Y$ **h.** $\{y\} \in X \Delta Y$ □

[26] A review of some aspects of classical (Riemann) integrals will be given in Chapter 3 (Calculus Revisited).

Exercise 2.7. Let $X = \{1, 2, 3, 4\}$ and let $Y = \{x, y\}$.

a. What is $X \times Y$? **c.** What is $\big|X \times Y\big|$? **e.** Is $(x, 3) \in X \times Y$? **g.** Is $3 \cdot x \in X \times Y$?
b. What is $Y \times X$? **d.** What is $\big|X \times Y\big|$? **f.** Is $(x, 3) \in Y \times X$? **h.** Is $2 \cdot y \in Y \times X$? □

Exercise 2.8. Let $X = \{8\}$. What is $2^{(2^X)}$?

Exercise 2.9. Let $A = \{1, \{1, 2\}, 2, 3, 4\}$ and $B = \{\{2, 3\}, 3, \{4\}, 5\}$. Compute the following.

a. $A \cap B$ **b.** $A \cup B$ **c.** $A \setminus B$ **d.** $B \setminus A$ **e.** $A \triangle B$ □

Exercise 2.10. Let A, X be sets such that $A \subseteq X$ and let $x \in X$. Prove the following:

a. If $a \in A$ then $A = (A \setminus \{a\}) \uplus \{a\}$.
b. If $a \notin A$ then $A = (A \uplus \{a\}) \setminus \{a\}$.
□

2.9.2 Other Exercises

Exercise 2.11. Let $D := \{(y_1, y_2) \in \mathbb{R}^2 : z_1^2 + z_2^2 = 1\}$. Let $h := \mathbb{Z} \to D$ be defined by $k \mapsto h(k) := \big(\cos(\frac{k\pi}{2}), \sin(\frac{k\pi}{2})\big)$. Compute the preimage $h^{-1}\{(0, 1), (1, 0)\}$.
Hint: What is $h(k)$ for $k = -4, -3, \ldots, 3, 4$? Draw a picture!

Solution: A: First, we compute $h^{-1}\{(0, 1)\}$.

$$h^{-1}\{(0,1)\} \;=\; \{k \in \mathbb{Z} : h(k) = (0,1)\} \;=\; \left\{k \in \mathbb{Z} : \cos\left(\frac{k\pi}{2}\right) = 0, \text{ and } \sin\left(\frac{k\pi}{2}\right) = 1\right\}.$$

To find all $k \in \mathbb{Z}$ such that both $\cos\left(\frac{k\pi}{2}\right) = 0$ and $\sin\left(\frac{k\pi}{2}\right) = 1$, we abbreviate $\theta := \frac{k\pi}{2}$ and look for all angles θ such that both $\cos(\theta) = 0$, and $\sin(\theta) = 1$. The answer:

$$\theta = \frac{\pi}{2} + 2n\pi\,, \quad \text{for some integer } n.$$

Going back to the original formulation of the problem, we must find all integers k such that

$$\frac{k\pi}{2} \;=\; \frac{\pi}{2} + 2n\pi \;=\; \frac{1 + 4n\pi}{2}\,, \quad \text{for some } n \in \mathbb{Z}.$$

This is equivalent to $k = 1 + 4n$, for some $n \in \mathbb{Z}$. Thus,

$$h^{-1}\{(0,1)\} \;=\; \{4n + 1 : n \in \mathbb{Z}\}\,.$$

B: Next, we compute $h^{-1}\{(1, 0)\}$. We follow the same pattern.

$$h^{-1}\{(1,0)\} \;=\; \{k \in \mathbb{Z} : h(k) = (1,0)\} \;=\; \left\{k \in \mathbb{Z} : \cos\left(\frac{k\pi}{2}\right) = 1, \text{ and } \sin\left(\frac{k\pi}{2}\right) = 0\right\}.$$

To find all $k \in \mathbb{Z}$ such that both $\cos\left(\frac{k\pi}{2}\right) = 1$ and $\sin\left(\frac{k\pi}{2}\right) = 0$, we abbreviate $\theta := \frac{k\pi}{2}$ and look for all angles θ such that both $\cos(\theta) = 1$, and $\sin(\theta) = 0$. The answer:

$$\theta \;=\; 2n\pi\,, \quad \text{for some integer } n.$$

Going back to the original formulation of the problem, we must find all integers k such that

$$\frac{k\pi}{2} \;=\; 2n\pi \;=\; \frac{4n\pi}{2}\,, \quad \text{for some } n \in \mathbb{Z}.$$

This is equivalent to $k \;= 4n$, for some $n \in \mathbb{Z}$. Thus,

$$h^{-1}\{(1,0)\} \;=\; \{4n : n \in \mathbb{Z}\}\,.$$

C: Since the preimage of a union is the union of the preimages,

$$h^{-1}\{(0,1),(1,0)\} \;=\; \{4n, 4n+1 : n \in \mathbb{Z}\}\,. \;\square$$

2.10 Blank Page after Ch.2

This page is intentionally left blank!

3 Calculus Revisited

We list here some avanced calculus topics. With the exception of the formulation of some of this material in dimensions arbitrary dimensions and the conditional convergence of series, all of it can be found in [12] Stewart, J: Single Variable Calculus and [12] Stewart, J: Multivariable Calculus. However, the notation needs getting used to, and some of the material is explained from an unfamiliar point of view that is more suitable for its application to probability theory.

3.1 Absolute Convergence of Series

You should be familiar with the next definition from your calculus class. See [12] Stewart, J: Single Variable Calculus.

Definition 3.1 (Absolute Convergence)**.**

We say that an infinite series $\sum a_j (a_j \in \mathbb{R})$ is **absolutely convergent** and also, that it **converges absolutely**, if

$$\sum_{j=1}^{\infty} |a_j| \;=\; |a_1| + |a_2| + |a_3| + \cdots \;<\; \infty\,, \quad \square$$

Theorem 3.1.

If the series $\sum a_j (a_j \in \mathbb{R})$ is absolutely convergent, then the following holds true:

(a) *The series $\sum a_j$ itself converges, i.e., there is $-\infty < a < \infty$ such that $\sum_{j=1}^{\infty} a_j = a$,*

(b) *ANY rearrangement $\sum_{j=1}^{\infty} a_{n_j} = a_{n_1} + a_{n_2} + \cdots$ converges to the same limit as $\sum a_j$.*

*We speak of a **rearrangement** of a sequence $(a_j)_{n\in\mathbb{N}}$ (a series $\sum a_j$) if its members are reshuffled into a sequence $(b_j)_{n\in\mathbb{N}}$ (a series $\sum b_j$) as follows: There are indices $n_j \in \mathbb{N}$ such that*

$$b_1 = a_{n_1},\; b_2 = a_{n_2},\; b_3 = a_{n_3},\; \ldots,$$

and those indices satisfy the following:

(1) *They are distinct: $i \neq j \;\Rightarrow\; n_i \neq n_j$.*

(2) *They leave no gaps in the set $\mathbb{N}$ of all indices: For each $k \in \mathbb{N}$ there is $j \in \mathbb{N}$ such that $k = n_j$.* [27]

PROOF: See your calculus book. ■

[27] ⋆ We could have expressed **(1)** and **(2)** by stating that the assignment $j \mapsto n_j$ is a bijection $\mathbb{N} \to \mathbb{N}$. (See Definition 2.19 (Surjective, injective and bijective functions) on p.34.)

Theorem 3.2.

If the series $\sum a_j (a_j \in \mathbb{R})$ satisfies $a_j \geq 0$ for all j, then

- *ANY rearrangement $\sum_{j=1}^{\infty} a_{n_j}$ possesses the same limit, finite or infinite, as $\sum_{j=1}^{\infty} a_j$.*
- *In particular, if $\sum a_j$ is not convergent, then $\sum_{j=1}^{\infty} a_{n_j} = \infty$ for each rearrangement.*

PROOF: ⋆

Case 1: The series has a finite limit. Then it converges absolutely, and the assertion follows from Theorem 3.1.

Case 2: Otherwise, since $a_j \geq 0$ for all j, $k \mapsto \sum_{j=1}^{k} a_j$ and $k \mapsto \sum_{j=1}^{k} a_{n_j}$ both are nondecreasing and nonnegative. By Theorem 2.3**(a)**, both have a limit. Let $a := \sum a_j$, $b_j := a_{n_j}$, $b := \sum b_j$. Note that $a \geq 0$ and $b \geq 0$, because $a_j \geq 0$ and $b_j \geq 0$ for all j.

Assume to the contrary that $b \neq a$. We assumed that $\sum a_j$ is not convergent, i.e.,

(⋆) $$a = \sum a_j = \infty\,.$$

Since $b \neq a$, this means that $0 \leq b \neq \infty$. Thus, $b \in \mathbb{R}$. Thus, $\sum b_j$ is absolutely convergent. By Theorem 3.1, each rearrangement $\sum b_{m_i}$ of $\sum b_j$ has the same limit b. Since $\sum a_j$ is a rearrangement of $\sum b_j$, it has the same limit $b < \infty$. However, by **(⋆)**, this limit is ∞.

In summary, the assumption $a \neq b$ led us to a contradiction and we conclude that it is not true. Thus, $b = a = \infty$. In other words, $\sum a_{n_j} = \sum a_j = \infty$. ■

Remark 3.1. ⋆ This remark might seem very strange to you. First, a definition.

A series $\sum a_j$ is called **conditionally convergent**, if it is convergent but not absolutely convergent.

This can be formulated as follows: There is some $a \in \mathbb{R}$ (thus, $-\infty < a < \infty$) such that $\sum_{j=1}^{\infty} a_j = a$, but $\sum_{j=1}^{\infty} |a_j| = \infty$.

The following is known as Riemann's rearrangement theorem: [28] Assume that the series $\sum a_j$ is conditionally convergent, but not absolutely convergent: Pick any $-\infty \leq b \leq \infty$. The terms a_j can be rearranged in such a way that the rearranged sequence, call it $\sum_{j=1}^{\infty} a_{n_j}$, converges to b. In other words, you can jumble the terms such that the limit is π. Some other rearrangement yields 0 as the limit, for yet another, $\sum_{j=1}^{\infty} a_{n_j} = -\sqrt{e^{30}}$, ... □

[28] This was proved by the German mathematician Bernhard Riemann (1826-1866). The integral that is being taught in calculus, the Riemann integral, also is named after him.

Example 3.1 (Harmonic series)**.** It is known from calculus that

$$\sum_{j=1}^{\infty} \frac{1}{n} \;=\; \infty \text{ (harmonic series) and that } \sum_{j=1}^{\infty} \frac{(-1)^n}{n} \text{ has a real limit.}$$

Thus, the series $\sum \frac{(-1)^n}{n}$ converges conditionally. □

Proposition 3.1.

(1) *A series which only has finitely many nonzero terms converges absolutely.*
(2) *If $|a_n| \le |b_n|$ for all n and $\sum b_n$ converges absolutely, then $\sum a_n$ converges absolutely.*

PROOF: ⋆

PROOF of **(1)**: Let the nonzero terms be $a_{n_1}, a_{n_2}, \ldots, a_{n_k}$,. Then, $\sum_{n=1}^{\infty} |a_n| = \sum_{j=1}^{k} |a_{n_j}| < \infty$.

PROOF of **(2)**: $|a_n| \le |b_n|$ for all $n \;\Rightarrow\; \sum_{n=1}^{\infty} |a_n| \le \sum_{n=1}^{\infty} |b_n| < \infty$. ■

Theorem 3.3.

Let S be some (abstract) nonempty set and $f : S \to \mathbb{R}$ some real–valued function on S.
Assume that $S^* := \{x \in S : f(x) \neq 0\}$ is countable, i.e. $S^* = \{x_1, x_2, \cdots\}$ for some finite or infinite sequence $x_1, x_2, \cdots$ of elements of S and that at least one of the following two is true:

(a) *$f(x_j) \ge 0$, for all j,* ***(b)*** *the series $\sum f(x_j)$ is absolutely convergent.*

- *Then, ANY rearrangement $\sum_{j=1}^{\infty} f(x_{n_j})$ of the $f(x_j)$ possesses the same value as $\sum_{j=1}^{\infty} f(x_j)$.*

PROOF: If **(a)** is true, the assertion follows from Theorem 3.2 on p.58. If **(b)** is true, it follows from Theorem 3.1 on p.58. ■

Notation 3.1 (Notation for series that do not depend on the order of summation)**.** Assume that f, S, S^* are as in Theorem 3.3 and that f satisfies **(a)** or **(b)** of that theorem. Then

$$S^* \;=\; \{x \in S : f(x) \neq 0\}$$

is countable and thus, there are two cases.

(a) S^* is finite, i.e., $S^* = \{x_1, x_2, \ldots, x_n\}$ for some suitable n. Since $\sum_{j=1}^{n} f(x_{n_j}) = \sum_{j=1}^{n} f(x_j)$ for each rearrangement x_{n_j} of the x_j, we write $\sum_{x \in S^*} f(x)$ for this common value.

(b) S^* is countably infinite, i.e., $S^* = \{x_1, x_2, \ldots\}$. Since $\sum_{j=1}^{\infty} f(x_{n_j}) = \sum_{j=1}^{\infty} f(x_j)$ for each rearrangement x_{n_j} of the x_j, here too, we write $\sum_{x \in S^*} f(x)$ for this common value.

Since $f(x) = 0$ for $x \in (S^*)^{\complement}$, the complement of S^* in S and including additional terms of value zero into a series does not impact its value, we also write $\sum_{x \in S} f(x)$ for $\sum_{x \in S^*} f(x)$.

Likewise, assume that $I = \{i_1, i_2, \dots\}$ is a countable index set and $(a_i)_{i \in I}$ is a family of real numbers which is indexed by I. If $a_i \geq 0$ for all i or $a_{i_1} + a_{i_2} + \cdots$ is absolutely convergent (or both), then rearranging the indices does not alter the value of the series and we can denote it by $\sum_{i \in I} a_i$.

To summarize,

(1) If the value of a series does not depend on the order of summation, there is no need to indicate a specific order by writing, e.g., $\sum_{i=1}^{\infty} \cdots$ or $\sum_{i=1}^{n} \cdots$. Under these circumstances, we also use notation such as $\sum_{x \in \dots} \cdots$ or $\sum_{i \in \dots} \cdots$. □

Theorem 3.4.

Assume that $J_1, J_2, \dots$ is a countable collection of disjoint subsets of $\mathbb{N}$. and $J := J_1 \uplus J_2 \uplus \cdots$. Let $\sum_{j \in J_1} a_j, \sum_{j \in J_2} a_j, \dots$ be a corresponding collection of series such that

- *$a_j \geq 0$, for all $j \in J$ or*
- *$\sum_{j \in J} a_j$ is absolutely convergent.*

Then

$$\sum_{j \in J_1} a_j + \sum_{j \in J_2} a_j + \cdots = \sum_{j \in J} a_j \,.$$

PROOF: Will not be given here. ■

3.2 Integration – The Riemann Integral

Integration is of high importance in probability theory, because in many important cases the probability of an event is computed as an area $\int_a^b f(y)dy$ enclosed by the graph of a function $u = f(y)$, the horizontal y axis and the vertical lines $u = a$ and $u = b$. More accurately, this is the case when this event is associated with a "continuous random variable". [29]

A quick word about symbol names. Writing $u = f(y)$, i.e., representing the argument by x and the function value by y, is the standard notation of the WMS text. For now we will go back to the more familiar notation $y = f(x)$.

Introduction 3.1. Here is a quick overview of the definition and geometric meaning of the Riemann Integral, the type of integral that you are familiar with from calculus. Integration will be discussed in greater detail after this introduction, starting with Section 3.2.1 (The Riemann Integral of a Step Function).

[29] see Chapter 10 (Continuous Random Variables).

(A) Integrating a function $y = f(x)$ of a single variable x:

An integral $\int_\alpha^\beta f(x)dx$ was defined as the limit of **Riemann sums**. [30] Those are areas

$$\sum_{j=1}^{n} f(a_j)\,(\beta_j - \alpha_j)\,, \tag{3.1}$$

obtained when one partitions an interval $[a, b]$ into subintervals

$$a \;=\; \alpha_0 \;<\; \beta_0 = \alpha_1 \;<\; \beta_1 = \alpha_2 \;<\; \cdots \;<\; \beta_{n-1} = \alpha_n \;<\; \beta_n \;=\; b\,,$$

picks arguments $a_j \in [\alpha_j, \beta_j]$ and replaces the integrand $x \mapsto f(x)$ with a **step function** [31]

$$x \;\mapsto\; \sum_{j=1}^{n} f(a_j)\,\mathbf{1}_{]\alpha_j,\beta_j]}(x)\,, \quad \alpha_j \le a_j \le \beta_j\,. \tag{3.2}$$

Here,

$$x \mapsto \mathbf{1}_{]\alpha_j,\beta_j]}(x) \;=\; \begin{cases} 1 & \text{if } x \in]\alpha_j, \beta_j], \text{ i.e., } \alpha_j < x \le \beta_j, \\ 0 & \text{else}\,, \end{cases}$$

is the indicator function [32] of the subinterval $]\alpha_j, \beta_j]$ of the interval $]a, b]$.

In other words, f is approximated by the constant value $f(a_j)$ on $]\alpha_j, \beta_j]$, and the area $\int_{\alpha_j}^{\beta_j} f(x)dx$ of f belonging to $]\alpha_j, \beta_j]$ is replaced by the area of a rectangle of width $\beta_j - \alpha_j$ and height $f(a_j)$.

(B) Now, consider the case of a real–valued function $y = f(\vec{x})$ which accepts $\mathbb{R}^2$–valued "random vectors" $\vec{x} = (x_1, x_2)$ as arguments.

The onedimensional interval $[\alpha, \beta]$ is replaced with a 2–dimensional rectangle $A = [\alpha, \beta] \times [\gamma, \delta]$ which is partitioned by horizontal and vertical grid lines into a finite number of subrectangles, let us call them $A_1, A_2, \ldots, A_k$. A point $\vec{a}_j = (x_j, y_j)$ in the plane is chosen from each $A_j = [\alpha_j, \beta_j] \times [\gamma_j, \delta_j]$. The integral

$$\iint_A f(\vec{x})\,d\vec{x} \;=\; \int_\alpha^\beta \int_\gamma^\delta f(x_1, x_2)\,dx_1 dx_2 \tag{3.3}$$

is approximated by the Riemann sum

$$\sum_{j=1}^{n} f(\vec{a}_j)\,(\beta_j - \alpha_j)(\delta_j - \gamma_j)\,, \tag{3.4}$$

obtained by replacing the integrand $\vec{x} \mapsto f(\vec{x})$ with the step function

$$\vec{x} \;\mapsto\; \sum_{j=1}^{n} f(\vec{a}_j)\,\mathbf{1}_{A_j}(\vec{x})\,, \quad \vec{a}_j \in A_j \;=\; [\alpha_j, \beta_j] \times [\gamma_j, \delta_j]\,, \tag{3.5}$$

whch is equal to the constant $f(\vec{a}_j)$ on all of A_j.

The geometric meaning is this: For each j, he volume of the slab between A and the graph of f is approximated by the volume of the quad formed by A_j at the bottom, the corresponding rectangle $\{\big(x, y, f(\vec{a}_j)\big) : (x, y) \in A_j\}$ at the top, and the vertical rectangles that connect the two.

[30] Riemann sums will be defined and treated in more detail in section 3.2.2 (The Riemann Integral as the Limit of Riemann Sums).

[31] Step functions will be defined and treated in more detail in section 3.2.1 (The Riemann Integral of a Step Function).

[32] see Definition 2.32 (indicator function for a set) on p.53.

(C) The case of a real–valued function $y = f(\vec{x})$ where $\vec{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ is quite similar.

Now, the domain is a quad $A = [\alpha, \beta] \times [\gamma, \delta] \times [\eta, \zeta]$. It is partitioned by "grid planes" parallel to the planes $\{(x_1, x_2, x_3) : x_1 = 0\}$, $\{(x_1, x_2, x_3) : x_2 = 0\}$ and $\{(x_1, x_2, x_3) : x_3 = 0\}$, into a finite number of subquads, $A_1, A_2, \ldots, A_k$. A point $\vec{a}_j = (x_j, y_j, z_j)$ in $\mathbb{R}^3$ is chosen from each

$$A_j \;=\; [\alpha_j, \beta_j] \times [\gamma_j, \delta_j] \times [\eta_j, \zeta_j]\,.$$

This time, the integral

$$\iiint\limits_A f(\vec{x})d\vec{x} \;=\; \int_\alpha^\beta \int_\gamma^\delta \int_\eta^\zeta f(x_1, x_2, x_3)dx_1dx_2dx_3 \tag{3.6}$$

is approximated by the Riemann sums

$$\sum_{j=1}^{n} f(\vec{a}_j)\,(\beta_j - \alpha_j)\,(\delta_j - \gamma_j)\,(\zeta_j - \eta_j)\,, \tag{3.7}$$

which one obtains by replacing $\vec{x} \mapsto f(\vec{x})$ with step functions $\vec{x} \mapsto \sum\limits_{j=1}^{n} f(\vec{a}_j)\mathbf{1}_{A_j}(\vec{x})$.

(D) The above can be generalized to functions defined on rectangles of arbitrary dimension d.

Since vectors of dimension $d > 3$ are beyond the scope of what is taught in a standard calculus sequence, only the following is expected of you.

> Try to recognize that and how the familiar cases $d = 1, 2, 3$ are special cases of what is now explained for an arbitrary dimension d. Do not worry about anything else.

We assume that $f : A \to \mathbb{R}$ is defined on a d–dimensional rectangle $A = [\alpha_1, \beta_1] \times \cdots \times [\alpha_d, \beta_d]$. Since $A \subseteq \mathbb{R}^d$, the arguments of f have the form $\vec{x} = (x_1, x_2, \ldots, x_d)$. The (Riemann) integral of f over A,

$$\iint\limits_A \cdots \int f(\vec{x})d\vec{x} \;=\; \int_{\alpha_1}^{\beta_1} \int_{\alpha_2}^{\beta_2} \cdots \int_{\alpha_d}^{\beta_d} f(x_1, x_2, \ldots, x_d)\,dx_1dx_2\cdots dx_d\,, \tag{3.8}$$

is again defined as the suitable limit of Riemann sums. Those are constructed as follows.

A is partitioned into a finite collection of d–dimensional subrectangles, $A_1, A_2, \ldots, A_k$. They are parallel to the $\{\vec{x} : x_1 = 0\}$, $\{\vec{x} : x_2 = 0\}, \ldots,$ $\{\vec{x} : x_d = 0\}$, "hyperplanes" and thus of the form

$$A_j \;=\; [\alpha_j^{(1)}, \beta_j^{(1)}] \times [\alpha_j^{(2)}, \beta_j^{(2)}] \times \cdots \times [\alpha_j^{(d)}, \beta_j^{(d)}]\,,$$

for suitable real numbers $\alpha_j^{(j)}, \beta_j^{(j)}$ such that $\alpha_j^{(j)} < \beta_j^{(j)}$ for each $j = 1, \ldots, d$.

Beware the notation! As you can see, we do the following: When we need to keep track of both the index $j = 1, \ldots, k$ of the subrectangle A_j and the coordinate $i = 1, \ldots, d$, then the latter is written as a superscript!

For each A_j, we choose a point $\vec{a}_j \;=\; \left(x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(d)}\right) \in A_j$ and approximate the integral by the Riemann sum

$$\sum_{j=1}^{n} f(\vec{a}_j)\,(\beta_j^{(1)} - \alpha_j^{(1)})(\beta_j^{(2)} - \alpha_j^{(2)})\cdots(\beta_j^{(d)} - \alpha_j^{(d)})\,. \tag{3.9}$$

It is obtained by replacing $\vec{x} \mapsto f(\vec{x})$ with the step function $\vec{x} \mapsto \sum\limits_{j=1}^{n} f(\vec{a}_j)\mathbf{1}_{A_j}(\vec{x})$. $\square$

Remark 3.2 (Stewart's notation for multiple integrals)**.** [12] Stewart, J: Multivariable Calculus. uses notation different from these lecture notes for double and triple integrals:

	These Lecture Notes	Stewart's book
$d=2$	$\iint\limits_D f(\vec{x})\,d\vec{x} \;=\; \iint\limits_D f(x_1,x_2)\,d(x_1,x_2)$	$\iint\limits_D f(x,y)\,dA$
$d=3$	$\iiint\limits_E f(\vec{x})d\vec{x} \;=\; \iiint\limits_E f(x_1,x_2,x_3)\,d(x_1,x_2,x_3)$	$\iiint\limits_E f(x,y)\,dV$

Concerning the region of integration, you will see both $\iiint\limits_E \cdots$ and $\iiint_E \cdots$, in this document. □

3.2.1 The Riemann Integral of a Step Function

When the Riemann integral is introduced as a means to compute the area under the graph of a function, this first done for a step function, where this area is that of a finite list of rectangles.

Rectangles of arbitrary dimension d were already introduced in Example 2.22 on p.52.

Definition 3.2 (d dimensional rectangles)**.** For $a,b\in\mathbb{R}$, $a\prec b$ here denotes either $a<b$ or $a\le b$.

Let $a_1\le b_1,\ a_2\le b_2,\ \ldots,\ a_d\le b_d,$ be d pairs of numbers ($d\in\mathbb{N}$). We call the set

$$\{\vec{x}=(x_1,\ldots,x_d)\in\mathbb{R}^d:\ a_1\prec x_1\prec b_1,\ a_2\prec x_2\prec b_2,\ldots,\ a_d\prec x_d\prec b_d\}$$

a d**–dimensional rectangle** (simply **rectangles**, if there is no confusion about d).

Special cases are

- **(a)** $]a_1,b_1[\,\times\cdots\times\,]a_d,b_d[$ ($a_j\prec x_j\prec b_j$ means $a_j<x_j<b_j$ for all j: **open rectangles**),
- **(b)** $]a_1,b_1\,]\,\times\cdots\times\,]a_d,b_d\,]$ ($a_j\prec x_j\prec b_j$ means $a_j<x_j\le b_j$ for all j: **half open rectangles**, also called **half closed rectangles**),
- **(c)** $[\,a_1,b_1[\,\times\cdots\times\,[\,a_d,b_d[$ ($a_j\prec x_j\prec b_j$ means $a_j\le x_j<b_j$ for all j: **half open rectangles**, also called **half closed rectangles**),
- **(d)** $[\,a_1,b_1\,]\,\times\cdots\times\,[\,a_d,b_d\,]$ ($a_j\prec x_j\prec b_j$ means $a_j\le x_j\le b_j$ for all j: **closed rectangles**).

Usually, onedimensional rectangles are called **intervals** and 3 dimensional rectangles are called **quads** or **boxes**. □

Example 3.2. As usual, we "identify" $\mathbb{R}^1$ with the real numbers line $\mathbb{R}$.

Rectangles in $\mathbb{R}^d$ were defined in Example2.22 on p.52:

- Rectangles in $\mathbb{R}^1=\mathbb{R}$ are intervals, e.g., $A=[a,b]$, where $a\le b$.
- Rectangles in $\mathbb{R}^2$ are, e.g., $A=]a_1,b_1]\times]a_2,b_2]$, where $a_1\le b_1$ and $a_2\le b_2$.
- Rectangles in $\mathbb{R}^3$ are quads, e.g., $A=[a_1,b_1]\times]a_2,b_2]\times]a_3,b_3[$, where $a_j\le b_j$, for $j=1,2,3$.

The last example demonstrates that "$<$" and "$\le$" need not be employed the same way for different coordinates j: The rectangular braces face different directions for $j=1,2,3$. □

The natural measure of an interval (a onedimensional rectangle) I with end points $a<b$, is it's length, $b-a$. Thus, if we write λ^1 for this measure, then

$$\lambda^1(I)\;=\;b-a\,.$$

The natural measure of (2 dimensional) rectangles, such as $R = [a_1, b_1] \times [a_2, b_2]$ and $R' =]a_1, b_1[\times]a_2, b_2[$, is their area, $(b_1 - a_1)(b_2 - a_2)$. Thus, if we write λ^2 for this measure, then

$$\lambda^2(R) \;=\; \lambda^2(R') \;=\; (b_1 - a_1)(b_2 - a_2)\,.$$

The natural measure of a 3 dimensional rectangle, e.g., $Q =]a_1, b_1]\times\,]a_2, b_2]\times\,]a_3, b_3]$, is its volume. Thus, if we write λ^3 for this measure, then

$$\lambda^3(Q) \;=\; (b_1 - a_1)\,(b_2 - a_2)\,(b_3 - a_3)\,.$$

those observations lead us to the following

Definition 3.3 (Lebesgue measure [33] of d dimensional rectangles). ⋆

Let $a \prec b$ again stand for either $a < b$ or $a \le b$.

Let $d \in \mathbb{N}$ and $a_j, b_j \in \mathbb{R}$ such that $a_j \le b_j$, for $j = 1, 2, \ldots, d$. Let

$$R \;:=\; \{\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d :\; a_1 \prec x_1 \prec b_1,\; a_2 \prec x_2 \prec b_2, \ldots,\; a_d \prec x_d \prec b_d\}$$

be a d–dimensional rectangle. We call

$$\lambda^d(R) \;:=\; (b_1 - a_1)\,(b_2 - a_2)\ldots(b_d - a_d) \tag{3.10}$$

the **d–dimensional Lebesgue measure** of R. We also simply speak of the **Lebesgue measure** of R, if there is no confusion about d).

We extend λ^d as follows.

- If $a_j < b_j$ for all j and $a_j = -\infty$ and/or $b_j = \infty$ for at least one j, then $\lambda^d(R) := \infty$.
- If $a_j = b_j$ for at least one j, then $\lambda^d(R) := 0$, even if not all a_j and b_j are finite.
- $\lambda^d(\emptyset) := 0$.
- If $R_1, R_2, \ldots$ is a finite or infinite sequence of disjoint rectangles, i.e., $R_i \cap R_j = \emptyset$ for $i \ne j$, then we define the **Lebesgue measure** of the union by "**σ–addititivity**" as follows:

$$\lambda^d\left(R_1 \uplus R_2 \uplus \cdots\right) \;:=\; \lambda^d(R_1) \,+\, \lambda^d(R_2) \,+\cdots \;\square \tag{3.11}$$

Remark 3.3.

(a) Note that $\lambda^d(R) = 0$ if and only if $a_j = b_j$ for at least one j.

(b) Be careful when viewing a subset of $\mathbb{R}^d$ as one of $\mathbb{R}^m$, where $d < m$. If, for example, one "identifies" $I := [0, 1] \subseteq \mathbb{R}$ with $I' := \{(x, y) \in \mathbb{R}^2 : 0 \le x \le 1,\; y = 0\} \subseteq \mathbb{R}^2$, then $\lambda^1(I) = 1$, but $\lambda^2(I') = 0$. (I' has area zero.) Moreover, $\lambda^2(I)$ and $\lambda^1(I')$ are nonsense expressions as far as mathematics goes, since λ^d only is defined for subsets of $\mathbb{R}^d$.

(c) You may not have peviously encountered (3.11), since it involves an infinite sequence of sets and an infinite series. However, you are, for $d = 1, 2, 3$, familiar with the additivity of measures,

[33] Named after the French mathematician Henri Léon Lebesgue (1875 – 1941)

(3.12) $$\lambda^d\left(R_1 \uplus R_2 \uplus \cdots \uplus R_k\right) \;:=\; \lambda^d(R_1) \,+\, \lambda^d(R_2) \,+\cdots+\, \lambda^d(R_k)\,.$$

This formula merely states that the combined length/area/volume of a finite collection of items equals the sum of the lengths/areas/volumes of the individual items. □

(d) Of course, it needs proof that one can indeed extend the definition of Lebesgue measure from a single rectangle to an arbitrary, infinite sequence of rectangles in such a manner, that σ–additivity holds true.

Rectangles and their Lebesgue measure are at the basis of the theory of integration. You may want to review Introduction 3.1 of this Chapter (Integration – The Riemann Integral) on p.61 while studying the following material on integration.

Definition 3.4.

A function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is called a **step function** if there is $n \in \mathbb{N}$, a list of d–dimensional rectangles $A_1, \ldots, A_n$, and a list of real numbers $c_1, \ldots, c_n$, such that

(3.13) $$\varphi(\vec{x}) \;=\; \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\vec{x}).$$

We call

(3.14) $$\int \varphi(\vec{x})\, d\vec{x} \;:=\; \int_{\mathbb{R}^d} \varphi(\vec{x})\, d\vec{x} \;:=\; \iint\cdots\int_{\mathbb{R}^d} f(\vec{x})\, d\vec{x} \;:=\; \sum_{j=1}^{n} c_j\, \lambda^d(A_j)$$

the (d dimensional) **Riemann integral** of the step function φ.

Here,

$$\vec{x} \mapsto \mathbf{1}_{A_j}(\vec{x}) \;=\; \begin{cases} 1 & \text{if } \vec{x} \in A_j, \\ 0 & \text{else}\,, \end{cases}$$

is the indicator function [34] of the subset A_j of $\mathbb{R}^d$. □

Remark 3.4. Fix $k \in [1, n]_{\mathbb{Z}}$. Note that the subset $[0, c_k] \times A_k$ of $\mathbb{R}^{d+1}$ is a $d+1$ dimensional rectangle and thus has $d+1$ dimensional Lebesgue measure

$$\lambda^{d+1}([0, c_k] \times A_k) \;=\; c_k \cdot \lambda^d(A_k)\,.$$

Definition 3.4 of the Riemann integral of a step function is consistent with the depiction of the Riemann integral given in Introduction 3.1 on p.61, since we can match up (3.14) with the Riemann sums in all four cases ($d = 1, d = 2, d = 3$, general d) of the introduction.

(A) The case $d = 1$:

If $a = \alpha_0 < \beta_0 = \alpha_1 < \beta_1 = \alpha_2 < \cdots < \beta_{n-1} = \alpha_n < \beta_n = b$ partitions of an interval $[a, b]$, into subintervals $A_k = [\alpha_k, \beta_k]$ for $k = 0, \ldots, n$, and if one defines $c_k = f(a_k)$ for some $\alpha_k \le a_k < \beta_k$, then (3.13) matches (3.2) on p.62, and the right hand side of (3.14) matches (3.1) on p.62.

(B) The case $d = 2$:

If $y = f(\vec{x})$, where $\vec{x} = (x_1, x_2)$ is a function of two variables, $A = [\alpha, \beta] \times [\gamma, \delta]$ is partitioned into a grid of subrectangles, $A_j = [\alpha_j, \beta_j] \times [\gamma_j, \delta_j]$, where $j = 1, \ldots, k$, and if we set $c_j = f(\vec{a}_j)$, for some point $\vec{a}_j \in A_j$, then then (3.13) matches (3.5) on p.62, and the right hand side of (3.14) matches (3.4) on p.62.

[34] see Definition 2.32 (indicator function for a set) on p.53.

(C) The case $d = 3$:
Assume that $y = f(\vec{x})$, where $\vec{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, and that we integrate over a quad $A = [\alpha, \beta] \times [\gamma, \delta] \times [\eta, \zeta]$, which is partitioned into a finite number of subquads, $A_1, A_2, \ldots, A_k$ of the form

$$A_j \;=\; [\alpha_j, \beta_j] \times [\gamma_j, \delta_j] \times [\eta_j, \zeta_j]\,.$$

Further, assume that point $\vec{a}_j = (x_j, y_j, z_j)$ in $\mathbb{R}^3$ is chosen from each A_j. Then the right hand side of (3.14) matches the Riemann sum (3.7) on p.63 and the step functions are identical.

(D) The case of general dimension d:
Finally, one also sees that the Riemann sum (3.9) on p.63 matches the right hand side of (3.14). □

Example 3.3. Here are two examples for $d = 1$.
(a) Let $n \in \mathbb{N}$. Both

$$g(x) \;:=\; \sum_{j=1}^{n} j\,\mathbf{1}_{A_j}(x)\,, \qquad A_j \;:=\; [1 - 1/2^{j-1}, 1 - 1/2^j]\,.$$
$$h(x) \;:=\; \sum_{j=1}^{n} j\,\mathbf{1}_{B_j}(x)\,, \qquad B_j \;:=\;]1 - 1/2^{j-1}, 1 - 1/2^j]\,.$$

are step functions. The Lebesgue measures of A_j and B_j occur in the computation of $\int_{-\infty}^{\infty} g(x)\,dx$ and $\int_{-\infty}^{\infty} h(x)\,dx$:

$$\lambda^1(A_j) \;=\; \lambda^1(B_j) \;=\; \left(1 - \frac{1}{2^j}\right) - \left(1 - \frac{1}{2^{j-1}}\right) \;=\; \frac{2}{2^j} - \frac{1}{2^j} \;=\; \frac{1}{2^j}\,.$$

Thus,

$$\int_{-\infty}^{\infty} g(x)\,dx \;=\; \sum_{j=1}^{n} j \cdot \lambda^1(A_j) \;=\; \sum_{j=1}^{n} \frac{j}{2^j} \;=\; \sum_{j=1}^{n} j \cdot \lambda^1(B_j) \;=\; \int_{-\infty}^{\infty} h(x)\,dx$$

(b) Let $\psi(x) \;:=\; \sum_{j=1}^{\infty} \mathbf{1}_{A_j}(x)$, with A_j as above. Since we have replaced finite sums with an infinite series, ψ is not a step function. However, it is known from calculus that the integral of ψ can be computed in the same fashion as that of g and h:

$$\int_{-\infty}^{\infty} \psi(x)\,dx \;=\; \sum_{j=1}^{\infty} 1 \cdot \lambda^1(A_j) \;=\; \frac{1}{2} \sum_{j=0}^{n} \frac{1}{2^j} \;=\; \frac{1}{2} \cdot \frac{1}{1 - 1/2} \;=\; \frac{1/2}{1/2} \;=\; 1\,.\ \square$$

Example 3.4. ($d = 2$ example.) Let $n \in \mathbb{N}$. For $i, j = 1, 2, \ldots, n$, let $c_{i,j}$ be real numbers and

$$g(\vec{x}) \;=\; g(x_1, x_2) \;:=\; \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j}\,\mathbf{1}_{A_{i,j}}(x_1, x_2)\,, \qquad A_{i,j} \;:=\; \left[\frac{i-1}{n}, \frac{i}{n}\right] \times \left[\frac{j-1}{n}, \frac{j}{n}\right].$$

Then g is a step function in $\mathbb{R}^2$. The Lebesgue measures of the A_j occur in the computation of the integral of g:

$$\lambda^2(A_j) \;=\; \left(\frac{i}{n} - \frac{i-1}{n}\right) \cdot \left(\frac{j}{n} - \frac{j-1}{n}\right) \;=\; \frac{1}{n^2}$$

Thus,

$$\iint_{\mathbb{R}^2} g(\vec{x})\,d\vec{x} \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2)\,dx_1 dx_2 \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} \cdot \lambda^2(A_{i,j}) \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{c_{i,j}}{n^2}\,.\ \square$$

Example 3.5. ($d = 3$ example.) Let $n_1, n_2, n_3 \in \mathbb{N}$. For $i = 1, \dots, n_1$, $j = 1, \dots, n_2, k = 1, \dots, n_3$, let $c_{i,j,k}$ be real numbers and

$$g(\vec{x}) \;=\; g(x_1, x_2, x_3) \;:=\; \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} c_{i,j,k}\, \mathbf{1}_{A_{i,j,k}}(x_1, x_2, x_3)\,, \qquad A_{i,j,k} \;:=\;]a_i^{(1)}, b_i^{(1)}]\times]a_i^{(2)}, b_i^{(2)}]\times]a_i^{(3)}, b_i^{(3)}]\,.$$

Thus, if we denote the 3 dimensional volume measure by λ^3, then $A_{i,j,k}$ is a quad with volume

$$\lambda^3(A_{i,j,k}) \;=\; (b_i^{(1)} - a_i^{(1)})\,(b_i^{(2)} - a_i^{(2)})\,(b_i^{(3)} - a_i^{(3)})\,.$$

g is a step function and its integral is

$$\iiint_{\mathbb{R}^3} g(\vec{x})\, d\vec{x} \;=\; \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} c_{i,j,k}\cdot \lambda^3(A_{i,j,k})\,. \;\square$$

3.2.2 The Riemann Integral as the Limit of Riemann Sums

3.2.2.1 The Riemann Integral in Dimension 1 ⋆

Given are $a, b \in \mathbb{R}$ such that $a < b$ and a list of real numbers

$$\Pi \;:=\; \big(\, y_0,\, y_1,\, \dots,\, y_n;\, u_1^*,\, u_2^*,\, \dots,\, u_n^* \,\big), \tag{3.15}$$

such that

$$a \;=\; y_0 \;<\; y_1, \;<\; \cdots \;<\; y_n \;=\; b\,, \qquad \text{and} \qquad y_{j-1} \;\le\; u_j^* \;\le\; y_j\,, \;\; \text{for each } j = 1, 2, \dots, n.$$

The lengths $y_j - y_{j-1}$ are not assumed to be of equal size. We call

$$\|\Pi\| \;:=\; \max\{y_{j+1} - y_j \;:\; j = 0, \dots, n-1\} \tag{3.16}$$

the **mesh** of Π. Note that $\|\Pi\|$ only depends on the endpoints of the subintervals $]y_{j-1}, y_j]$ but not on the "sample points" [35] u_j. Also note that the "subintervals" $]y_{j-1}, y_j]$, $j = 1, \dots, n$ are a partition of the interval $]a, b]$ in the sense of Definition 2.10 on p.25.

You are familiar with the above and the next definition from your single variable calculus class.

Definition 3.5.

Let Π be defined as in (3.15), and let $f : [a, b] \to \mathbb{R}$ be a function on $[a, b]$. We call

$$\mathcal{RS}(f; \Pi) \;:=\; \sum_{j=1}^{n} f(u_j)(y_j - y_{j-1})$$

the **Riemann sum** of f with respect to Π, and we call

$$\int_a^b f(x)dx \;:=\; \lim_{\|\Pi\|\to 0} \mathcal{RS}(f; \Pi)$$

the **Riemann integral** of f on $[a, b]$, provided that this limit exists.

[35] It is generally accepted terminology to refer to u_j as a sample point. We use quotes around this term in this chapter on integration, because it is reserved in a course on probability for the elements of a probability space, also referred to as a sample space. See Remark 1.2 on p.12

It is very instructive to work through the following example of a function for which $\int_a^b f(x)dx$ does NOT exist.

Example 3.6. Let $A := \mathbb{Q} \cap [0,1]$ be the set of all rational numbers in the unit interval, and

$$f(y) \;:=\; \mathbf{1}_A(y) \;=\; \begin{cases} 1, & \text{if } y \in A, \\ 0, & \text{else}\,. \end{cases}$$

Then $\int_0^1 f(x)dx$ does not exist for the following reasons. Let $y_1, \ldots, y_n$ be an arbitrary list of numbers such that $0 = y_0 < y_1 < y_2 \cdots < y_{n-1} < y_n = 1$. Since any interval with bounds $\alpha < \beta$ contains both rational and irrational numbers, there are rational q_j and irrational i_j such that $y_{j-1} < q_j < y_j$ and $y_{j-1} < i_j < y_j$. Consider Π^*, Π_*, and their corresponding Riemann sums, defined as follows.

$$\Pi^* \;:=\; \big(\, y_0, \ldots, y_n;\, q_1, \ldots, q_n \,\big), \qquad \mathcal{RS}(\mathbf{1}_A; \Pi^*) \;=\; \sum_{j=1}^n \mathbf{1}_A(q_j)(y_j - y_{j-1})\,,$$

$$\Pi_* \;:=\; \big(\, y_0, \ldots, y_n;\, i_1, \ldots, i_n \,\big), \qquad \mathcal{RS}(\mathbf{1}_A; \Pi_*) \;=\; \sum_{j=1}^n \mathbf{1}_A(i_j)(y_j - y_{j-1})\,.$$

Since $q_j \in A$ and $i_j \notin A$ for all j, $\mathbf{1}_A(q_j) = 1$ and $\mathbf{1}_A(i_j) = 0$ for all j. From this we obtain

$$\mathcal{RS}(\mathbf{1}_A; \Pi^*) \;=\; 1 \qquad \text{and} \qquad \mathcal{RS}(\mathbf{1}_A; \Pi_*) \;=\; 0\,. \tag{3.17}$$

Since all this is true for any $n \in \mathbb{N}$ and sets of real numbers $0 = y_0 < \cdots < y_n = 1$, one can build partitions Π^* and Pi_* such that $\|\Pi^*\|$ and $\|\Pi_*\|$ both are arbitrarily close to zero.
For example, if $y_j = j/n$ for $j = 0, 1, \ldots, n$, then $\|\Pi^*\| = \|\Pi_*\| = 1/n$. One sees from (3.17) that

$$\int_a^b \mathbf{1}_A(x)dx \;=\; \lim_{\|\Pi\|\to 0} \mathcal{RS}(\mathbf{1}_A; \Pi) \quad \text{does not exist.} \;\square$$

Remark 3.5. Let Π be defined as in (3.15) on p.68, and let $f : [a,b] \to \mathbb{R}$. Consider

$$\varphi_\Pi(x) \;:=\; \sum_{j=1}^n f(u_j)\mathbf{1}_{[y_{j-1}, y_j]}\,.$$

Then φ_Π is a step function in the sense of Definition 3.4 (for dimension $d = 1$), with integral

$$\int_{\mathbb{R}} \varphi_\Pi(x) \;=\; \sum_{j=1}^n f(u_j)(y_j - y_{j-1})\,.$$

See (3.14) on p.66. Observe that the equations

$$\mathcal{RS}(f; \Pi) \;=\; \sum_{j=1}^n f(u_j)(y_j - y_{j-1}) \;=\; \int_{\mathbb{R}} \varphi_\Pi(x) \qquad \text{and} \qquad \int_a^b f(x)dx \;=\; \lim_{\|\Pi\|\to 0} \mathcal{RS}(f; \Pi)$$

imply that

$$\int_a^b f(x)dx \;=\; \lim_{\|\Pi\|\to 0} \int_{\mathbb{R}} \varphi_\Pi(x)\,. \;\square \tag{3.18}$$

3.2.2.2 The Riemann Integral in Dimension 2 ⋆

The notation becomes more complex if we integrate a function of two variables over a rectangle. We write $\vec{y} = (y^{(1)}, y^{(2)})$ for a vector $\vec{y} \in \mathbb{R}^2$, so its coordinates are written as superscripts.
Let $\vec{a}, \vec{b} \in \mathbb{R}^2$ such that $a^{(1)} < b^{(1)}$ and $a^{(2)} < b^{(2)}$. Let

$$\Pi := \big(\vec{y}_0, \vec{y}_1, \ldots, \vec{y}_n; \ \vec{u}(1,1), \vec{u}(1,2), \vec{u}(2,1), \vec{u}(2,2), \vec{u}(3,1), \ldots, \vec{u}(n-1,n), \vec{u}(n,n) \big) \tag{3.19}$$

be a list of vectors $\vec{y}(j), \vec{u}(j_1, j_2) \in \mathbb{R}^2$ with the following properties.

- The vectors $\vec{y}_0, \ldots, \vec{y}_n$ satisfy

$$\begin{aligned} a^{(1)} &= y_0^{(1)} < y_1^{(1)} < \cdots < y_n^{(1)} = b^{(1)} \\ a^{(2)} &= y_0^{(2)} < y_1^{(2)} < \cdots < y_n^{(2)} = b^{(2)} . \end{aligned} \tag{3.20}$$

- The $n+1$ vectors $\vec{y}_j$ generate, for each selection of indices j_1, j_2 such that $1 \le j_1 \le n$ and $1 \le j_2 \le n$, the edges of a rectangle

$$\begin{aligned} R(j_1, j_2) &:= \,]y_{j_1-1}^{(1)}, y_{j_1}^{(1)}\,] \times]y_{j_2-1}^{(2)}, y_{j_2}^{(2)}\,] , \\ \text{with area } A(j_1, j_2) &= (y_{j_1}^{(1)} - y_{j_1-1}^{(1)}) \cdot (y_{j_2}^{(2)} - y_{j_2-1}^{(2)}) . \end{aligned} \tag{3.21}$$

  The side lengths $y_j^{(i)} - y_{j-1}^{(i)}$ are not assumed to be of equal size for any i and j.

- Each vector $\vec{u}(j_1, j_2),\ 1 \le j_1 \le n,\ 1 \le j_2 \le n,$ satisfies

$$\vec{u}(j_1, j_2) \in R(j_1, j_2) . \tag{3.22}$$

In other words, if $\vec{u}(j_1, j_2) = \big(u(j_1, j_2)^{(1)}, u(j_1, j_2)^{(2)} \big)$, then its coordinates satisfy

$$\begin{aligned} y_{j_1-1}^{(1)} &\le u(j_1, j_2)^{(1)} \le y_{j_1}^{(1)} , \\ y_{j_2-1}^{(2)} &\le u(j_1, j_2)^{(2)} \le y_{j_2}^{(2)} . \end{aligned} \tag{3.23}$$

We measure the fineness of Π by the following two dimensional analogue of (3.16) on p.68.

$$\|\Pi\| := \max \left\{ y_{j_1}^{(1)} - y_{j_1-1}^{(1)} , y_{j_2}^{(2)} - y_{j_2-1}^{(2)} \ : \ j_1, j_2 = 1, \ldots, n \right\} \tag{3.24}$$

Note the following.

- $\|\Pi\|$ only depends on the side lengths of the subrectangles $R(j_1, j_2)$ of (3.21), but not on the "sample points" $\vec{u}(j_1, j_2)$.
- Those rectangles $R(j_1, j_2)$ are a partition of the rectangle $\,]a^{(1)}, b^{(1)}\,] \times]a^{(2)}, b^{(2)}\,]$ in the sense of Definition 2.10 on p.25.
- $\|\Pi\| \to 0$ requires that both their horizontal and vertical lengths must approach 0.

You know the above and the next definition from multivariable calculus.

Definition 3.6. Let Π be defined as in (3.19), and let R denote the rectangle $\,[\,a^{(1)}, b^{(1)}\,] \times [\,a^{(2)}, b^{(2)}\,]$.

Let $f : R \to \mathbb{R};\ \vec{y} \mapsto f(\vec{y})$, be a real–valued function on R. We call

$$\mathcal{RS}(f;\Pi) := \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} f(\vec{u}(j_1,j_2))\,(y^{(1)}_{j_1} - y^{(1)}_{j_1-1})\cdot(y^{(2)}_{j_2} - y^{(2)}_{j_2-1}) \tag{3.25}$$

the **Riemann sum** of f with respect to Π, and we call

$$\iint_R f(\vec{y})\,d\vec{y} := \lim_{\|\Pi\|\to 0} \mathcal{RS}(f;\Pi) \tag{3.26}$$

the **Riemann integral** of f on R, provided that this limit exists. □

Remark 3.6. Note that

$$\varphi_\Pi(\vec{y}) = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} f(\vec{u}(j_1,j_2))\,\mathbf{1}_{[y^{(1)}_{j_1-1},y^{(1)}_{j_1}]\times[y^{(2)}_{j_2-1},y^{(2)}_{j_2}]}(\vec{y})$$

is, for $d = 2$, a step function in the sense of Definition 3.4, with integral

$$\int_{\mathbb{R}^2} \varphi_\Pi(\vec{y})\,d\vec{y} = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} f(\vec{u}(j_1,j_2))\,(y^{(1)}_{j_1} - y^{(1)}_{j_1-1})\cdot(y^{(2)}_{j_2} - y^{(2)}_{j_2-1})\,,$$

and that the equations (3.25) and (3.26) imply that

$$\iint_R f(\vec{y})\,d\vec{y} = \lim_{\|\Pi\|\to 0}\int_{\mathbb{R}^2} \varphi_\Pi(\vec{y})\,d\vec{y}\,. \tag{3.27}$$ □

Example 3.7. Here is an example of a two dimensional partition of a rectangle into $n^2 = 9$ subrectangles (i.e., $n = 3$).

We write the vectors as column vector and square braces rather than parentheses as delimiters. Let

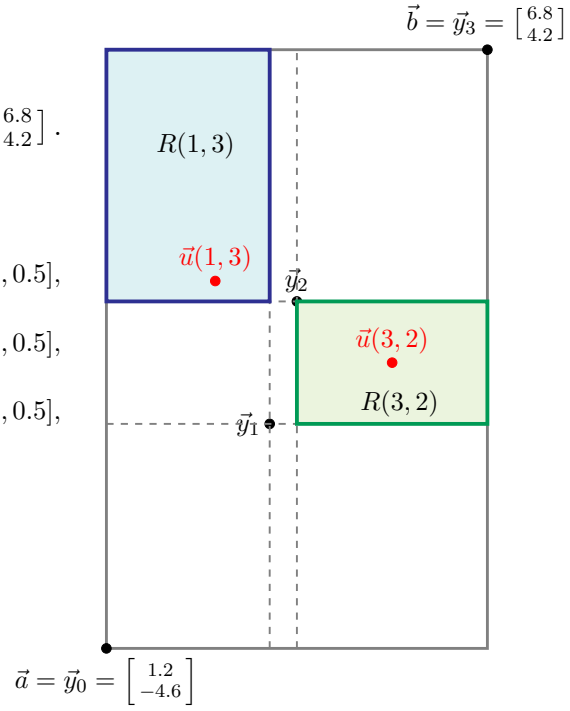
$$\vec{a} = \vec{y}_0 = \begin{bmatrix} 1.2 \\ -4.6 \end{bmatrix}, \ \vec{y}_1 = \begin{bmatrix} 3.6 \\ -1.3 \end{bmatrix}, \ \vec{y}_2 = \begin{bmatrix} 4.0 \\ 0.5 \end{bmatrix}, \ \vec{b} = \vec{y}_3 = \begin{bmatrix} 6.8 \\ 4.2 \end{bmatrix}.$$

These four vectors partition the rectangle $]1.2, 6.8\,]\times\,]-4.6, 4.2]$ into a 3×3 grid of subrectangles

$R(1,1) =]1.2, 3.6\,]\times\,]-4.6, -1.3]$, $R(1,2) =]1.2, 3.6\,]\times\,]-1.3, 0.5]$,
$R(1,3) =]1.2, 3.6\,]\times\,]0.5, 4.2]$,
$R(2,1) =]3.6, 4.0\,]\times\,]-4.6, -1.3]$, $R(2,2) =]3.6, 4.0\,]\times\,]-1.3, 0.5]$,
$R(2,3) =]3.6, 4.0\,]\times\,]0.5, 4.2]$,
$R(3,1) =]4.0, 6.8\,]\times\,]-4.6, -1.3]$, $R(3,2) =]4.0, 6.8\,]\times\,]-1.3, 0.5]$,
$R(3,3) =]4.0, 6.8\,]\times\,]0.5, 4.2]$.

Possible choices for the vectors $\vec{u}(j_1, j_2)$ are, e.g.,

$$\vec{u}(1,1) = \begin{bmatrix} 2.1 \\ -3.4 \end{bmatrix}, \ \vec{u}(1,2) = \begin{bmatrix} 2.0 \\ -1.1 \end{bmatrix}, \ \vec{u}(1,3) = \begin{bmatrix} 2.8 \\ 0.8 \end{bmatrix},$$
$$\vec{u}(2,1) = \begin{bmatrix} 3.8 \\ -2.2 \end{bmatrix}, \ \vec{u}(2,2) = \begin{bmatrix} 3.7 \\ -0.5 \end{bmatrix}, \ \vec{u}(2,3) = \begin{bmatrix} 3.8 \\ 2.8 \end{bmatrix},$$
$$\vec{u}(3,1) = \begin{bmatrix} 5.2 \\ -4.0 \end{bmatrix}, \ \vec{u}(3,2) = \begin{bmatrix} 5.4 \\ -0.4 \end{bmatrix}, \ \vec{u}(3,3) = \begin{bmatrix} 4.1 \\ 4.1 \end{bmatrix}, \quad \square$$



3.1 (Figure). **2–dim Riemann sum.**

3.2.2.3 The Riemann Integral in d Dimensions ⋆ We do not discuss separately the Riemann integral in $d = 3$ dimension and directly discuss the case of general d. We write

$$\vec{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(d)}\right),$$

so the d coordinates of the vector are written as superscripts.

Let $\vec{a}, \vec{b} \in \mathbb{R}^d$ such that $a^{(i)} < b^{(i)}$ for $i = 1, 2, \ldots, d$. Let

$$\Pi := \left(\vec{y}_0, \vec{y}_1, \ldots, \vec{y}_n; \ \left(\vec{u}(j_1, \ldots, j_d)\right)_{(j_1,\ldots,j_d)\in J}\right) \tag{3.28}$$

be a list of vectors $\vec{y}(j) \in \mathbb{R}^d$ and $\vec{u}(j_1, \ldots, j_d) \in \mathbb{R}^d$ as follows. [36]

- The vectors $\vec{y}_0, \ldots, \vec{y}_n$ satisfy

$$a^{(i)} = y_0^{(i)} < y_1^{(i)} < \cdots < y_n^{(i)} = b^{(i)}, \qquad \text{for each coordinate } \ i = 1, 2, \ldots, d. \tag{3.29}$$

- J is the set of all "composite indices" $(j_1, \ldots, j_d)$ that satisfy
  - ⊡ $j_1, j_2, \ldots, j_d \in \mathbb{N}$ (hence, $(j_1, \ldots, j_d) \in \mathbb{N}^d$)
  - ⊡ $1 \le j_k \le n$ for each $k = 1, 2, \ldots, d$. (Thus, Π contains d^n vectors $\vec{u}(j_1, \ldots, j_d)$.)
- The $n + 1$ vectors $\vec{y}_0, \ldots, \vec{y}_d$ generate, for each selection of indices $(j_1, j_2, \ldots, j_d) \in J$, the edges of a d–dimensional rectangle [37]

[36] ⋆ We have not given the order in which the vectors $\vec{u}(j_1, \ldots, j_d)$ are listed

[37] See Example 2.22 on p.52.

$$R(j_1, j_2, \ldots, j_d) \;:=\;]y^{(1)}_{j_1-1}, y^{(1)}_{j_1}] \times]y^{(2)}_{j_2-1}, y^{(2)}_{j_2}] \times \cdots \times]y^{(d)}_{j_d-1}, y^{(d)}_{j_d}] \,. \tag{3.30}$$

The side lengths $y^{(i)}_j - y^{(i)}_{j-1}$ are not assumed to be of equal size for any i and j.

- For each $(j_1, j_2, \ldots, j_d) \in J$, the vector $\vec{u}(j_1, j_2, \ldots, j_d)$ satisfies

$$\vec{u}(j_1, j_2, \ldots, j_d) \;\in\; R(j_1, j_2, \ldots, j_d) \,. \tag{3.31}$$

In other words, if $\vec{u}(j_1, j_2, \ldots, j_d) = \big(\, u(j_1, j_2, \ldots, j_d)^{(1)}, u(j_1, j_2, \ldots, j_d)^{(2)}, \ldots, u(j_1, j_2, \ldots, j_d)^{(d)} \,\big)$, then its d coordinates, $u(j_1, j_2, \ldots, j_d)^{(k)}$, satisfy

$$y^{(k)}_{j_k-1} \;\le\; u(j_1, j_2, \ldots, j_d)^{(k)} \;\le\; y^{(k)}_{j_k} \,, \qquad \text{for } k = 1, 2, \ldots, d. \tag{3.32}$$

The fineness of Π has the following d–dimensional analogue of (3.16) on p.68. and of (3.24) on p.70.

$$\|\Pi\| \;:=\; \max\Big\{\, y^{(1)}_{j_1} - y^{(1)}_{j_1-1} \,,\; y^{(2)}_{j_2} - y^{(2)}_{j_2-1} \,,\; \ldots \,,\; y^{(d)}_{j_d} - y^{(d)}_{j_d-1} \;:\; (j_1, j_2, \ldots, j_d) \in J \Big\} \,. \tag{3.33}$$

Note the following.

- $\|\Pi\|$ only depends on the side lengths $y^{(k)}_{j_k} - y^{(k)}_{j_k-1}$ of the subrectangles $R(j_1, j_2, \ldots, j_d)$ of (3.30), but not on the "sample points" $\vec{u}(j_1, j_2, \ldots, j_d)$.
- Those rectangles $R(j_1, j_2, \ldots, j_d)$ are a partition, in the sense of Definition 2.10 on p.25, of the rectangle $]a^{(1)}, b^{(1)}] \times]a^{(2)}, b^{(2)}] \times \cdots \times]a^{(d)}, b^{(d)}]$ on p.25.
- $\|\Pi\| \to 0$ requires that the side lengths $y^{(k)}_{j_k} - y^{(k)}_{j_k-1}$ must approach 0, for each coordinate $k = 1, 2, \ldots, d$.

For dimension $d = 3$, you should be familiar with the above and the next definition from multivariable calculus and it is strongly suggested that you write on paper this definition and the subsequent Remark 3.7 for $d = 3$.

Definition 3.7. Let Π be as in (3.28) and $R := [\,a^{(1)}, b^{(1)}\,] \times [\,a^{(2)}, b^{(2)} \times \cdots \times [\,a^{(d)}, b^{(d)}\,]$.

Let $f : R \to \mathbb{R}\,;\; \vec{y} \mapsto f(\vec{y})$, be a real–valued function on R. We call

$$\mathcal{RS}(f; \Pi) \;:=\; \sum_{j_1, \ldots, j_d = 1}^{n} f(\vec{u}(j_1, j_2, \ldots, j_d)) \, (y^{(1)}_{j_1} - y^{(1)}_{j_1-1}) \cdot (y^{(2)}_{j_2} - y^{(2)}_{j_2-1}) \cdots (y^{(d)}_{j_d} - y^{(d)}_{j_d-1}) \tag{3.34}$$

the **Riemann sum** of f with respect to Π, and we call

$$\iint\!\cdots\!\int_R f(\vec{y}) \, d\vec{y} \;:=\; \lim_{\|\Pi\| \to 0} \mathcal{RS}(f; \Pi) \tag{3.35}$$

the **Riemann integral** aka **proper Riemann integral** of f on R, provided that this limit exists.

In (3.34), $\sum_{j_1, \ldots, j_d = 1}^{n}$ indicates that each summation variable $j_1, j_2, \ldots, j_d$ takes each value $1, 2, \ldots, n$. We have introduced the notion of a proper integral here, because later on we will also define improper Riemann integrals. [38] □

[38] See Definition 3.8 (Improper Riemann integral) on p.75.

Remark 3.7. Note that

$$\varphi_\Pi(\vec{y}) \;:=\; \sum_{j_1,\dots,j_d=1}^{n} f(\vec{u}(j_1,j_2,\dots,j_d))\,\mathbf{1}_{[y^{(1)}_{j_1-1},y^{(1)}_{j_1}]\times[y^{(2)}_{j_2-1},y^{(2)}_{j_2}]\times\cdots\times[y^{(d)}_{j_d-1},y^{(d)}_{j_d}]}(\vec{y})$$

is, for general d, a step function in the sense of Definition 3.4, with integral

$$\int_{\mathbb{R}^d}\varphi_\Pi(\vec{y})\,d\vec{y} \;=\; \sum_{j_1,\dots,j_d=1}^{n} f(\vec{u}(j_1,j_2,\dots,j_d))\,(y^{(1)}_{j_1}-y^{(1)}_{j_1-1})\cdot(y^{(2)}_{j_2}-y^{(2)}_{j_2-1})\cdots(y^{(d)}_{j_d}-y^{(d)}_{j_d-1})\,,$$

and that the equations (3.34) and (3.35) imply that

$$\iint\!\cdots\!\int_R f(\vec{y})\,d\vec{y} \;=\; \lim_{\|\Pi\|\to 0}\int_{\mathbb{R}^d}\varphi_\Pi(\vec{y})\,d\vec{y}\,. \quad \square \tag{3.36}$$

Example 3.8. Here is an example of a Riemann sum for $d=3$ dimensions and $n=4$. We write the vectors as column vector and square braces rather than parentheses as delimiters. Let

$$\vec{a}=\vec{y}_0=\begin{bmatrix}1.2\\-4.6\\3.0\end{bmatrix},\quad \vec{y}_1=\begin{bmatrix}3.6\\-1.3\\4.2\end{bmatrix},\quad \vec{y}_2=\begin{bmatrix}4.0\\0.5\\5.6\end{bmatrix},\quad \vec{y}_3=\begin{bmatrix}6.8\\4.2\\6.0\end{bmatrix},\quad \vec{b}=\vec{y}_4=\begin{bmatrix}8.7\\9.2\\7.7\end{bmatrix}.$$

These five vectors partition the rectangle $\;]1.2,6.8\,]\times\,]-4.6,4.2]$ into $n^d=4^3=64$ subrectangles,

$$R(1,1,1),\ R(1,1,2),\ R(1,1,3),\ R(1,1,4),\ R(1,2,1),\ \dots,\ R(4,4,3),\ R(4,4,4)\,.$$

Possible choices for the vectors $\vec{u}(j_1,j_2,j_3)$ are, e.g.,

$$\vec{u}(1,1,1)=\begin{bmatrix}2.1\\-3.4\\3.8\end{bmatrix},\quad \vec{u}(1,1,2)=\begin{bmatrix}2.0\\-1.1\\5.0\end{bmatrix},\quad \vec{u}(1,1,3)=\begin{bmatrix}2.8\\-2.1\\5.7\end{bmatrix},\quad \vec{u}(1,1,4)=\begin{bmatrix}1.4\\-2.2\\6.9\end{bmatrix},$$

$$\vec{u}(1,2,1)=\begin{bmatrix}1.4\\-0.5\\3.9\end{bmatrix},\quad \cdots\cdots\cdots,\quad \vec{u}(4,4,3)=\begin{bmatrix}7.5\\8.2\\5.7\end{bmatrix},\quad \vec{u}(4,4,4)=\begin{bmatrix}7.0\\6.3\\6.3\end{bmatrix}.\quad \square$$

3.3 Improper Integrals and Integrals Over Subsets

We defined separately, for dimensions $d=1$, $d=2$ and for general d, the Riemann integral $\int\limits_R f(\vec{x})d\vec{x}$ of a function f on a d dimensional rectangle $R\subseteq\mathbb{R}^d$. See Definitions 3.5 on p.68, 3.6 on p.70 and 3.7 on p.73. We did so for didactic reasons since, strictly speaking, Definition 3.7 also includes the cases $d=1$ (functions of a real variable x) and $d=2$. We will state the definition of Riemann integrability only once, for general d.

But first, a quick reminder concerning improper integrals. That definition we only give for the onedimensional case. [39]

And now, the definition of Riemann integrability.

[39] ⋆ For multiple dimensions, $d>1$, the definition of the **improper Riemann integral** (over all of $\mathbb{R}^d$) is that

$$\iint\!\cdots\!\int_{\mathbb{R}^d} f(\vec{x})\,d\vec{x} \;:=\; \lim_{a\to\infty}\iint\!\cdots\!\int_{[-a,a]^d} f(\vec{x})\,d\vec{x}\,, \tag{3.37}$$

provided that this limit exists.

Definition 3.8 (Improper Riemann integral)**.**

Let $f : [a, \infty[\to \mathbb{R}$, $g :]-\infty, b] \to \mathbb{R}$, $h :]-\infty, \infty[\to \mathbb{R}$.
Their **improper Riemann integrals** are defined as follows:

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx\,, \tag{3.38}$$

$$\int_{-\infty}^b f(x)\,dx = \lim_{a\to-\infty} \int_a^b f(x)\,dx\,.$$

$$\int_{-\infty}^\infty f(x)\,dx = \lim_{a\to-\infty} \lim_{b\to\infty} \int_a^b f(x)\,dx\,. \quad \square$$

And now, the definition of Riemann integrability.

Definition 3.9 (Riemann integrability)**.**

(a) Let $A \subseteq \mathbb{R}^d$ be a d dimensional rectangle and $\varphi : A \to \mathbb{R}$, a real–valued function on A.
We say that φ is **Riemann integrable**, if its proper Riemann integral, as specified (for general d) in Definition 3.7 on p.73, exists and is finite.

(b) Let ψ be one of the functions f, g, h specified in Definition 3.8 (Improper Riemann integral) above. We say that ψ is **Riemann integrable**, if its improper integral, as specified in Definition 3.8 above, exists and is finite.

(c) If φ is as above and α, its proper Riemann integral exists, then we call α the (proper) Riemann integral, even if $\alpha = \pm\infty$ (and thus, φ is not Riemann integrable).

(c) If ψ is as above and β, its improper integral exists, then we call β the improper Riemann integral of ψ, even if $\beta = \pm\infty$ (and thus, ψ is not Riemann integrable). $\square$

Remark 3.8. ⋆ The distinction between a function having a Riemann integral and being Riemann integrable matches how one handles sequences x_n of real numbers and infinite series $\sum a_n$.

- Recall that, e.g., the sequence $x_n = -n$ does not converge to $-\infty$. Rather, it diverges, even though we say that it has the limit $\lim_{n\to\infty} x_n = -\infty$.
- For another example, consider the series $\sum n^{-1}$. We say that its limit is $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and that it diverges. We do not say that it converges to ∞. $\square$

Integration of functions over a subset utilizes indicator functions. See Definition 2.32 on p.53.

Definition 3.10.

(A): Let $R \subseteq \mathbb{R}^d$ be a d dimensional rectangle, $d \in \mathbb{N}$, and $\emptyset \neq A \subseteq R$. Let $f : A \to \mathbb{R}$ be a function on A such that the function

$$\mathbf{1}_A f : R \longrightarrow \mathbb{R} \;\; \vec{x} \mapsto \mathbf{1}_A(\vec{x}) f(\vec{x}) \;\; = \begin{cases} f(\vec{x}) & \text{if } \vec{x} \in A, \\ 0, & \text{else}, \end{cases} \tag{3.39}$$

possesses a Riemann integral. Then we call

$$\iint\limits_A \cdots \int f(\vec{x})\, d\vec{x} \;:=\; \iint\limits_R \cdots \int \mathbf{1}_A(\vec{x})\, f(\vec{x})\, d\vec{x} \tag{3.40}$$

the **Riemann integral of** f **on** (also, **over,**) **the subset** A.

We are not yet completely done with the case $d = 1$, since we also must consider improper integrals of functions of a single variable. We do that now.

(B): Let $I \subseteq \mathbb{R}$ be an interval of infinite length, i.e., I is one of $[a, \infty[, \;]-\infty, b], \;]-\infty, \infty[$, for suitable $a, b \in \mathbb{R}$. Let $\emptyset \neq A \subseteq I$ and $f : A \to \mathbb{R}$ a function on A such that the function

$$\mathbf{1}_A f : I \longrightarrow \mathbb{R} \;\; x \mapsto \mathbf{1}_A(x) f(x) \;\; = \begin{cases} f(x) & \text{if } x \in A, \\ 0, & \text{else}, \end{cases} \tag{3.41}$$

possesses an improper Riemann integral. Then we call

$$\int_A f(x)\, dx \;:=\; \int_I \mathbf{1}_A(x)\, f(x)\, dx \tag{3.42}$$

the **Riemann integral of** f **on** (also, **over,**) **the subset** A. □

Remark 3.9. We often use the following simplified notation for multivariable integrals:

- We also write $\int_A f(\vec{x})\, d\vec{x}$ for $\iint\limits_A \cdots \int f(\vec{x})\, d\vec{x}$. □

Remark 3.10. ⋆

Here is a fine point which may have escaped your attention. Per se, both (3.39) and (3.41) depend on the containing rectangle, R. However, one can show that the number $\int_R \mathbf{1}_A(\vec{x}) f(\vec{x}) d\vec{x}$ does not depend on $R \supseteq A$, and that the number $\int_I \mathbf{1}_A(x) f(x) dx$ does not depend on $I \supseteq A$. □

Remark 3.11. Consider the formula (3.42) for the special case, that $f(\vec{x}) = 1$, for all $\vec{x}$. You will find in [12] Stewart, J: Multivariable Calculus. the following formulas that relate 2 dimensional integrals of the constant function 1 to areas and 3 dimensional integrals of the constant function 1 to olume:

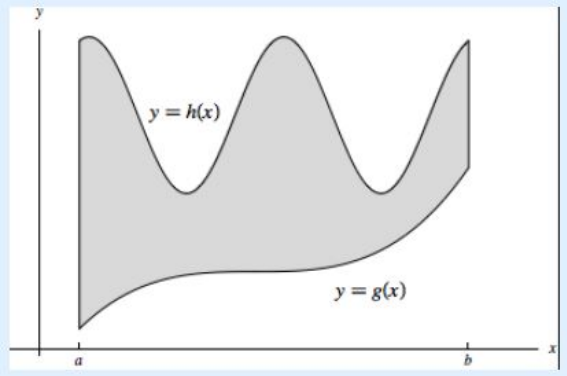
$$\iint\limits_{A_2} d\vec{x} \;=\; \text{area of } A_2 \qquad \text{and} \qquad \iiint\limits_{A_3} d\vec{x} \;=\; \text{volume of } A_3 \tag{3.43}$$

Those integrals exist for very general sets A_2 and A_3. For example, A_2 can be a a type 1 or type 2 region, as shown in the pictures below. [40] For more detail, see your multivariable calculus book.

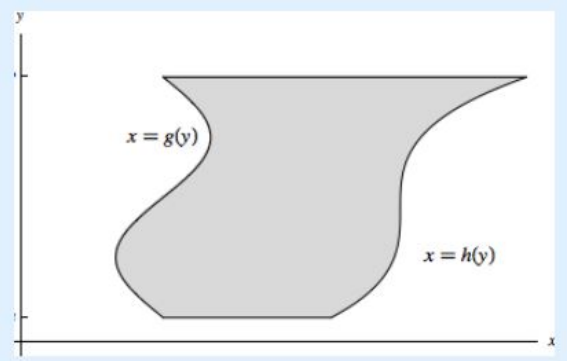
[40] Source: University of Texas. The type 2 region picture does not extend far enough to the left. Otherwise one could see that the region extends vertically from $y = c$ to $y = d$.

The areas of such type 1 and type 2 regions are computed according to the formulas

$$\text{area of Type 1 region} \;=\; \int_{x=a}^{b}\int_{y=g(x)}^{h(x)} dy\,dx \qquad \text{area of Type 2 region} \;=\; \int_{y=c}^{d}\int_{x=g(y)}^{h(y)} dx\,dy$$



3.2 (Figure). **Type 1 region in $\mathbb{R}^2$.**



3.3 (Figure). **Type 2 region in $\mathbb{R}^2$.** □

The next theorem is one of the many reasons why integration is such an important tool in probability theory and statistics.

Theorem 3.5.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real–valued, nonnegative, and Riemann–integrable function on $\mathbb{R}^d$. Let

$$\mathscr{R} \;:=\; \{A \subseteq \mathbb{R}^d : \mathbf{1}_A \text{ is Riemann integrable }\}.$$

If $\int_{\mathbb{R}^d} f(\vec{x})\,d\vec{x} \;=\; 1$, then the set function $P(A) \;:=\; \int_A f(\vec{x})\,d\vec{x}$ satisfies Definition 1.2 on p.13 of a Probability measure on $\mathscr{R}$, in the following sense:

- $P(\emptyset) = 0$ • $P(\mathbb{R}^d) = 1$ • $0 \le P(A) \le 1$, *for all* $A \in \mathscr{R}$.
- *σ–additivity: If $A_n \in \mathscr{R}$ are disjoint and $A := \biguplus_{n\in\mathbb{N}} A_n \in \mathscr{R}$, then $P(A) \;=\; \sum_{n\in\mathbb{N}} P(A_n)$.*

PROOF: Will not be given here. We just mention that you will see the assertion of this theorem restated in Corollary 4.2 on p.98 for Lebesgue integrals instead of Riemann integrals. ■

Remark 3.12. A lot more will be said in later chapters about the following:

- It is not always possible to define a probability for all subsets of the probability space.
- This issue will mostly be of no concern to us. □

3.4 Series and Integrals as Tools to Compute Probabilities

3.4.1 Series and Sums

We repeat in the next remark the most important results of Section 3.1 (Absolute Convergence of Series).

Remark 3.13. In the next theorem we consider countable probability spaces (Ω, P). Thus,

- either Ω is finite and can be written $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ for some suitable $k \in \mathbb{N}$,
- or Ω is countably infinite and can be written $\Omega = \{\omega_j : j \in \mathbb{N}\}$ for some suitable $k \in \mathbb{N}$.

For what follows, recall Theorem 3.3 on p.60 and the subsequent Notation 3.1 (Notation for series that do not depend on the order of summation):

Let $a_1, a_2, \dots \in \mathbb{R}$ such that $a_j \geq 0$ for all j. Then

- ANY rearrangement $\sum_{j=1}^{\infty} a_{n_j}$ of the a_j possesses the same value as $\sum_{j=1}^{\infty} a_j$.
- We are allowed to write $\sum_{j\in\mathbb{N}} a_j$ instead of $\sum_{j=1}^{\infty} a_j$.

We apply this as follows. Assume that Ω is countable and f is a nonnegative function on Ω. If our aim is to compute $\sum_{j=1}^{\infty} f(\omega_j)$, then it does not matter in what order Ω has been arranged as a sequence $\omega_1, \omega_2, \dots$. The value of $\sum_{j=1}^{\infty} f(\omega_j)$ is the same for any such sequencing of Ω and we can write $\sum_{\omega\in\Omega} f(\omega)$ rather than $\sum_{j=1}^{\infty} f(\omega_j)$ for that common value.

Since all subsets of Ω are countable, all of the above remains true for $A \subseteq \Omega$ in place of Ω.

Since finitely many terms can be summed in any order, all of the above also applies to finite Ω.

Thus, for finite $A = \{a_1, \dots, a_n\} \subseteq \Omega$ or countably infinite $A' = \{a'_1, a'_2, \dots\} \subseteq \Omega$, we can write

$$\sum_{\omega\in A} f(\omega) \;=\; \sum_{j=1}^{n} f(a_j)\,, \qquad \sum_{\omega\in A'} f(\omega) \;=\; \sum_{j=1}^{\infty} f(a'_j)\,. \tag{3.44}$$

Independence of the order in which a finite or infinite sequence of nonnegative is used in the formulation of the next theorem, a simplified version (no "σ–algebra") of Corollary 5.1**(b)** on p.111. □

Theorem 3.6.

Let Ω be an arbitrary, nonempty, countable set. Let $p : \Omega \longrightarrow \mathbb{R}$ be a function on Ω which satisfies

$$\bullet\ p(\omega) \geq 0 \text{ for all } \omega \in \Omega, \qquad \bullet\ \sum_{\omega\in\Omega} p(\omega) \;=\; 1\,. \tag{3.45}$$

Then, $\omega \mapsto p(\omega)$ defines a probability measure P on Ω as follows.

$$P(\emptyset) \;:=\; 0\,; \qquad P(A) \;:=\; \sum_{\omega\in A} p(\omega) \tag{3.46}$$

PROOF: $P(\emptyset) = 0$ is true by (3.46), and $P(\Omega) = 1$ follows from the second assumption of (3.45). σ–additivity can be shown by employing Theorem 3.4 on p.61. ■

Remark 3.14. It follows from (3.46) that $P(\{\omega\}) = p(\omega)$. Thus, for general $\emptyset \neq A \subseteq \Omega$,

$$P(A) \;=\; \sum_{\omega\in A} P(\{\omega\})\,. \tag{3.47}$$ □

All probability spaces that were discussed in Section 1.2 (A First Look at Probability) were finite. For example, the outcomes of rolling three dice were modeled as the set $\Omega = [1, 2, \dots, 6]^3$, a set of size $6^3 = 216$, with equiprobable outcomes: $P(\omega) = 1/216$. here are some examples of countably infinite probability spaces.

Example 3.9. Let

$$p : \mathbb{N} \to [0, \infty[\,; \qquad j \mapsto p(j) := \left(\frac{2}{3}\right)^{j-1} \left(\frac{1}{3}\right) .$$

Thus, $p(1) = \frac{1}{3}$, $p(2) = \frac{2}{3} \cdot \frac{1}{3} = \frac{2}{9}$, $p(3) = \frac{4}{9} \cdot \frac{1}{3} = \frac{4}{27}, \dots$

Certainly, $p(j) \geq 0$ for all $j \in \mathbb{N}$. If we can show that $\sum\limits_{j=1}^{\infty} p(j) = 1$, then (3.46) defines a probability measure on $\mathbb{N}$. For convenience, let $a := 2/3$, $b := 1/3$. Then $p(j) = ba^{j-1}$. Since $a + b = 1$ and $\sum\limits_{j=0}^{\infty} a^j = \dfrac{1}{1-a}$, 'we obtain

$$\sum_{j=1}^{\infty} p(j) = b \sum_{j=1}^{\infty} a^{j-1} = b \sum_{j=0}^{\infty} a^j = b \cdot \frac{1}{1-a} = \frac{b}{b} = 1 .$$

We have shown that $P(A) = \sum\limits_{j \in A} \left(\frac{2}{3}\right)^{j-1} \left(\frac{1}{3}\right)$ defines a probability measure on $\mathbb{N}$. We will learn in Section 9.3 (Geometric + Negative Binomial + Hypergeometric Distributions) that P is a geometric distribution with parameter $\frac{1}{3}$. This distribution is used, for example, to model the probabilities pertaining to the number of times one must roll a die until a 5 or a 6 shows up for the first time. □

Example 3.10. Let $a_1, a_2, \dots$ a sequence of nonnegative numbers such that $c := \sum\limits_{j=1}^{\infty} a_j < \infty$. Let Ω be some countably infinite set which has been arranged into the (specific) sequence $\Omega = \{\omega_1, \omega_2, \dots\}$.

Let $f : \Omega \to [0, \infty[$ be defined by $f(\omega_j) := a_j$. Unless $\sum\limits_{j=0}^{\infty} a_j = 1$, the conditions of Theorem 3.6 are not met and f does not define a probability measure on Ω. However, $p(\omega_j) := f(\omega_j)/c$ satisfies

$$\sum_{j \in \mathbb{N}} p(\omega_j) = \frac{1}{c} \cdot \sum_{j \in \mathbb{N}} f(\omega_j) = \frac{1}{c} \cdot \sum_{j \in \mathbb{N}} a_j = 1 .$$

Thus, $P(A) = \sum\limits_{\omega \in A} p(\omega)$ defines a probability measure on Ω. □

Example 3.11. Let $c \in \mathbb{R}$ and

$$p : [0, \infty[_{\mathbb{Z}} \to [0, \infty[\,; \qquad j \mapsto p(j) := c \cdot \frac{1}{4^j\, j!} .$$

is there a value of c that makes $A \mapsto P(A) := \sum_{j \in A} p(j)$ a probability measure on $[0, \infty[_{\mathbb{Z}}$?
Since j is a nonnegative integer, $p(j) > 0$ for $c > 0$. Thus we are done if we can find $c > 0$ such that

$$\sum_{j=0}^{\infty} \frac{c}{4^j\, j!} = 1 . \tag{3.48}$$

We find c as follows. c must satisfy (3.48). Thus

$$\frac{1}{c} = \sum_{j=0}^{\infty} \frac{1}{4^j\, j!} = \sum_{j=0}^{\infty} \frac{(1/4)^j}{j!} = e^{1/4}\,.$$

It follows that if $c = e^{-1/4}$, then $P(A) = e^{-1/4} \cdot \sum_{j\in A} \frac{1}{4^j\, j!}$ defines a probability measure on the nonnegative integers. It will be defined in Section 9.4 (The Poisson Distribution) as a Poisson distribution with parameter $\frac{1}{4}$. This distribution is used, for example, to model the probabilities pertaining to the number of occurences of a rather rare item within a unit. An example would be the number of car accidents in a town (the rare occurrences) during a day (the unit). □

Example 3.12. A sample of the eye colors of 75 persons is taken. The frequencies are as follows.

brown	25
blue	15
black	20
green	5
other	10

Thus, the corresponding relative frequencies are obtained by dividing by the sample size.

brown	1/3
blue	1/5
black	4/15
green	1/15
other	2/15

Let $\Omega := \{$ brown, blue, black, green, other $\}$. Then $p(\text{brown}) := 1/3$, $p(\text{blue}) := 1/5$, ..., $p(\text{other}) := 2/15$ satisfies (3.45); thus (3.46) defines a probability measure P on Ω.

Observe that this probablity measure is not about the true distribution of eye colors in the population from which the sample was taken. It only tells us about the apportionment of eye colors in the particular sample of 75 persons that we have taken.

For example, $P\{$ blue or green $\} = 1/5 + 1/15 = 4//15$ is the probability that a random pick from the sample has blue or green eyes. The corresponding probability for a random pick from the population could be different.

Evidently the procedure just described can be applied to any finite collection of frequencies. Note however, that statisticians will not refer to the relative frequencies of sample data as probabilities. [41] They reserve that term for probability measures that defined for the model of reality they study. They compare the relative frequencies of the sample to the corresponding probabilities of that model and make a decision whether that model is or is not appropriate. □

We now switch focus from series to integrals as a tool to define probability measures.

[41] The major exception is if those sample data are used to define empirical probabilities. See Example 1.1 (Empirical probability) on p.6.

3.4.2 Integrals

Introduction 3.2. According to Theorem 3.5 on p.77, nonnegative, Riemann integrable functions $f(\vec{x})$ on $\mathbb{R}^d$ that satisfy $\int_{\mathbb{R}^d} f(\vec{x})\, d\vec{x} = 1$, define probability measures by means of

$$P(A) := \int_A f(\vec{x})\, d\vec{x}\,. \tag{3.49}$$

Here, we assume that A is Riemann integrable (i.e., its indicator function $\mathbf{1}_A(\vec{x})$ is Riemann integrable). We study some examples in this section. □

- Throughout this section, "P is a probability measure on $\mathbb{R}^d$" does not imply that $A \mapsto P(A)$ is defined for all $A \subseteq \mathbb{R}^d$. Rather, it suffices that $P(A)$ is defined for Riemann integrable A.

Example 3.13. Show that

$$f(t) := 7 \cdot \mathbf{1}_{[0,\infty[}(x) e^{-7t}\,.$$

makes the assignment $P(A) = \int_A f(t)dt$ a probability measure.

Solution: We compute the Riemann integral

$$\int_{-\infty}^{\infty} f(t)\, dt = 7\int_0^{\infty} e^{-7t}\, dt = 7\frac{-1}{7}e^{-7t}\Big]_0^{\infty} = \frac{-7}{7}(0-1) = \frac{7}{7} = 1\,.$$

This shows that $f(x) = 3 \cdot \mathbf{1}_[0,\infty[(x)\, e^{-3x}$ defines a probability measure via (3.2). □

Example 3.14. Let $c \in \mathbb{R}$ and

$$h(y) := c \cdot \mathbf{1}_{[0,\pi]}(y) \sin(y)\,.$$

(a) What value of c makes the assignment $P(A) = \int_A h(y)dy$ a probability measure?
(b) Compute $P(]-10\pi, \pi/2[)$.

Solution for (a): c must be chosen such that $\int_{\mathbb{R}} h(y)dy = 1$. We compute the Riemann integral

$$\int_{-\infty}^{\infty} h(y)\, dy = c\int_0^{\pi} \sin(y)\, dy = (-c)\cos(y)\Big]_0^{\pi} = (-c)(-1+1) = 2c\,.$$

This expression equals 1 for $c = \frac{1}{2}$. Thus, $f(y) = \mathbf{1}_[0,\infty[(y)\ \sin(y)/2$ defines a probability measure via (3.2).

Solution for (b):

$$P(]-10\pi, \pi[) = \int_{-10\pi}^{\pi/2} h(y)\, dy = \frac{1}{2}\int_0^{\pi/2} \sin(y)\, dy = \left(-\frac{1}{2}\right)\cos(y)\Big]_0^{\pi/2} = \frac{1}{2}\,.$$ □

Example 3.15. Show that

$$g(x) \;:=\; \mathbf{1}_{[0,\infty[}(x)\cdot xe^{-x}\,.$$

makes the assignment $P(A)=\int_A g(x)dx$ a probability measure.

Solution: We compute the Riemann integral

$$\int_{-\infty}^{\infty} g(x)\,dx \;=\; \int_0^{\infty} xe^{-x}\,dx\,.$$

Since $\lim\limits_{x\to\infty} xe^{-x}=0$, integration by parts yields

$$\int_0^{\infty} xe^{-x}\,dx \;=\; x\Big(-e^{-x}\Big]_0^{\infty}\Big) - (-1)\int_0^{\infty} e^{-x}\,dx \;=\; 0+\int_0^{\infty} e^{-x}\,dx \;=\; 1\,.$$

This shows that $g(x) \;:=\; \mathbf{1}_{[0,\infty[}(x)\cdot xe^{-x}$ defines a probability measure via (3.2). □

Example 3.16. Let $a\in\mathbb{R}$. Let $B \;:=\; [0,\infty[\times[0,\pi]\times[0,\infty[$ and

$$f(\vec{y}) \;:=\; f(y_1,y_2,y_3) \;:=\; a\cdot\mathbf{1}_B(\vec{y})\cdot(7/2)\cdot e^{-7y_1}\cdot\sin(y_2)\cdot y_3e^{-y_3}\,.$$

What value of a makes the assignment $P(A)=\int_A f(\vec{y})d\vec{y}$ a probability measure?

Solution: This example is easy if you have worked through the previous three examples. Let $g(\vec{y}):=f(\vec{y})/a$. (Note that we may assume $a\neq 0$. Otherwise $f(\vec{y})\equiv 0$, and thus, $\int_{\mathbb{R}^3} f(\vec{y})=0\neq 1$.)

$$\int\limits_{\mathbb{R}^3} g(\vec{y})\,d\vec{y} \;=\; \iiint\limits_{[0,\infty[\times[0,\pi]\times[0,\infty[} 7e^{-7y_1}\cdot\sin(y_2)/2\cdot y_3e^{-y_3}\,d\vec{y}$$

We apply Fubini's Theorem (applied iteration) and obtain

$$\int_{\mathbb{R}^3} g(\vec{y})\,d\vec{y} \;=\; \int_{[0,\infty[} 7e^{-7y_1}\left[\int_{[0,\pi]}\frac{\sin(y_2)}{2}\left(\int_{[0,\infty[} y_3e^{-y_3}\,dy_3\right)dy_2\right]dy_1\,.$$

By Example 3.15, this simplifies to

$$\int_{\mathbb{R}^3} g(\vec{y})\,d\vec{y} \;=\; \int_{[0,\infty[} 7e^{-7y_1}\left[\int_{[0,\pi]}\frac{\sin(y_2)}{2}\cdot 1\,dy_2\right]dy_1\,.$$

By Example 3.14, this simplifies to

$$\int_{\mathbb{R}^3} g(\vec{y})\,d\vec{y} \;=\; \int_{[0,\infty[} 7e^{-7y_1}\cdot 1\,dy_1\,.$$

By Example 3.13, this simplifies to

$$\int_{\mathbb{R}^3} g(\vec{y})\,d\vec{y} \;=\; 1.$$

Thus, g itself is the function we are looking for! Since $g(\vec{y}):=f(\vec{y})/a$, We must set $a:=1$. □

This concludes our review of Riemann integration. In the next chapter we will extend the Riemann integral to a larger set of functions.

4 Calculus Extensions

You will see the following advice repeated more than once in this document.

- Many results are formulated for general dimension d. If you find dealing with this level of generality difficult, we suggest that you formulate the assertions for dimensions $1, 2, 3$ and see to it that you understand those special cases.

Introduction 4.1. We had announced at the end of the previous chapter that we will extend the Riemann integral to a larger set of functions. Before embarking on this endeavor, let us review some of the core properties of the Riemann integral that we would like to maintain for most if not all members of this enlarged set of integrands.

(a) For step functions $\varphi(\vec{x}) = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\vec{x})$, we defined $\int \varphi(\vec{x})\, d\vec{x} = \sum_{j=1}^{n} c_j\, \lambda^d(A_j)$. See Definition 3.4 on p.66.

(b) We defined $\int f(\vec{x})\, d\vec{x}$ for a general function f as the limit of step function integrals. (Those step functions were Riemann sums. See Definition 3.7 on p.73.

(c) For subsets $A \subseteq \mathbb{R}^d$, we defined $\int_A f(\vec{x})\, d\vec{x} = \int \mathbf{1}_A(\vec{x}) f(\vec{x})\, d\vec{x}$ by use of the indicator function $\mathbf{1}_A$ See Definition 3.10 on p.75.

Note that the integrals $\int f(\vec{x})\, d\vec{x}$ obtained in **(b)** are proper Riemann integrals. Improper Riemann integral are defined by means of additional limits.

The proper Riemann integral satisfies the following:

(d)	$\int_A 0\, d\vec{x} = 0$, and $f \geq 0 \;\Rightarrow\; \int_A f(\vec{x})\, d\vec{x} \geq 0$	(positivity)
(e)	$f \leq g$ on $A \;\Rightarrow\; \int_A f(\vec{x})\, d\vec{x} \leq \int_A g(\vec{x})\, d\vec{x}$	(monotonicity)
(f)	$\int_A \big(c_1 f(\vec{x}) + c_2 g(\vec{x})\big)\, d\vec{x} = c_1 \int_A f(\vec{x})\, d\vec{x} + c_2 \int_A g(\vec{x})\, d\vec{x}$	(linearity)

Assume that $R = [\alpha_1, \beta_1] \times \cdots \times [\alpha_d, \beta_d]$ is a d dimensional, closed and bounded rectangle, and that the function f is defined on R. Under certain conditions, [42] the integral $\int_R f(\vec{x})\, d\vec{x}$ can be computed as an iterated integral, and the order of integration is unimportant: This is Fubini's Theorem. [43]

(g)
$$\int_R f(\vec{x})\, d\vec{x} = \int_{\alpha_1}^{\beta_1} \left(\int_{\alpha_2}^{\beta_2} \left(\cdots \int_{\alpha_d}^{\beta_d} f(\vec{x})\, dx_d \cdots \right) dx_2 \right) dx_1$$
$$= \int_{\alpha_{j_1}}^{\beta_{j_1}} \left(\int_{\alpha_{j_2}}^{\beta_{j_2}} \left(\cdots \int_{\alpha_{j_d}}^{\beta_{j_d}} f(\vec{x})\, dx_{j_d} \cdots \right) dx_{j_2} \right) dx_{j_1} \qquad \text{(Fubini)}$$

holds true for any rearrangement $j_1, j_2, \ldots, j_d$ of $1, 2, \ldots, d$. If we think of the innermost integral as being evaluated first and the outermost integral as being evaluated last, **(g)** states that the order of integration can be switched from $dx_d dx_{d-1} \cdots dx_2 dx_1$ to $dx_{j_d} dx_{j_{d-1}} \cdots dx_{j_2} dx_{j_1}$.

Be sure to understand this formula for $d = 2$ and $d = 3$. If $d = 2$, then there is only one rearrangement of $1, 2$ different from $1, 2$, and that is $2, 1$. Thus, **(g)** simplifies to

[42] f is bounded and has at most finitely many points of discontinuity

[43] Named after the Italian mathematician Guido Fubini (1879 – 1943)

(h) $\displaystyle\int_R f(x_1,x_2)\,d(x_1,x_2) \;=\; \int_{\alpha_1}^{\beta_1}\left(\int_{\alpha_2}^{\beta_2} f(x_1,x_2)\,dx_2\right)dx_1 \;=\; \int_{\alpha_2}^{\beta_2}\left(\int_{\alpha_1}^{\beta_1} f(x_1,x_2)\,dx_1\right)dx_2$

If $d = 3$, then there already are five different ways to rearrange $1,2,3$, giving you six ways to compute $\int_R f(x_1,x_2,x_3)\,d(x_1,x_2,x_3)$ as an iterated integral. (What are they?)

Another important property of the Riemann integral is the following.

(i) If f is Riemann integrable and $g(x) = f(x)$ except for finitely many arguments x, then $\int f(x)dx = \int g(x)dx$.

Also, in certain situations, one can interchange the order of integration and taking limits. For example, the sequence of functions $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = x^n$, has as limit the function $f(x) = \mathbf{1}_1(x)$ which equals 1 if $x = 1$ and 0, else. Note that

$$\lim_{n\to\infty}\int_0^1 f_n(x)dx \;=\; \lim_{n\to\infty}\frac{x_{n+1}}{n+1}\Big]_{x=0}^{1} \;=\; 0 \quad\text{and}\quad \int_0^1 \mathbf{1}_1(x)dx \;=\; \int_1^1 dx \;=\; 0\,.$$

In other words, it is true in this particular case, that

(j) $$\lim_{n\to\infty}\int_0^1 f_n(x)dx \;=\; \int_0^1\left(\lim_{n\to\infty} f_n(x)\right)dx$$

Unfortunately, this is an area where the Riemann is seriously lacking. It is even possible that

- the sequence $f_n(x)$ converges to a function $f(x) = \lim_{n\to\infty} f_n(x)$ on some interval $[a,b]$.
- $\int_a^b f_n(x)dx$ exists for all n.
- Not only is $\int_a^b f(x)dx = \lim_{n\to\infty}\int_a^b f_n(x)dx$ false, but $\int_a^b f(x)dx$ does not even exist.

Here is an example. It is known that the set $\mathbb{Q}$ of all rational numbers is countable, i.e., it can be enumerated as a sequence. Thus, the subset $A := \mathbb{Q}\cap[0,1]$ also is countable and we can write $A = \{q_j : j\in\mathbb{N}\}$, for suitable rational numbers $q_1, q_2, \ldots$. Define

$$f_n(x) \;:=\; \mathbf{1}_{\{q_1,\ldots,q_n\}} \;=\; \begin{cases} 1, & \text{if } x\in\{q_1,\ldots,q_n\},\\ 0, & \text{else}\,,\end{cases} \qquad f(x) \;:=\; \mathbf{1}_A(x) \;=\; \begin{cases} 1, & \text{if } y\in\{q_1,q_2,\ldots\},\\ 0, & \text{else}\,.\end{cases}$$

Clearly, $f(x) = \lim_{n\to\infty} f_n(x)$. Moreover, $\int_0^1 f_n(x)dx = \int_0^1 0dx = 0$. See **(i)**. However, we have seen in Example 3.6 on p.69 that $\int_0^1 f(x)dx$ does not exist.

We are now going to create an extension of the Riemann integral. It is called the Lebesgue integral [44] and we will see that

- its construction shows some parallels to steps **(a)** – **(c)**;
- it possesses the very desirable properties **(d)** – **(i)**;
- it will be much better behaved as far as **(j)** is concerned. □

4.1 Extension of Lebesgue Measure to the Borel sets of $\mathbb{R}^d$

First, we extend Lebesgue measure which is, for dimensions $d = 1,2,3$, how we measure length, area, volume, to larger collections of subsets of $\mathbb{R}^d$. So far, λ^d only is defined for d dimensional rectangles and, by σ–additivity, to countable, disjoint unions of such rectangles. See Definition

[44] That is again Henri Lebesgue, the mathematician after whom the Lebesgue measure is named.

3.3 (Lebesgue measure) on p.65. Formulas (3.40) and (3.43) of Remark 3.11 on p.76 make the next definition seem very natural.

Definition 4.1.

Let $A \subseteq \mathbb{R}^d$. If it exists, we call the Riemann integral of the constant function 1 over A,

$$\lambda^d(A) := \idotsint_A d\vec{x} = \idotsint_R \mathbf{1}_A(\vec{x})\, d\vec{x}, \quad (R \text{ is a rectangle that contains } A), \tag{4.1}$$

the d **dimensional Lebesgue measure** of A. □

The next theorem shows that Lebesgue measure can be extended beyond the sets of Definition 4.1 to an even larger collection of sets.

Theorem 4.1. ⋆ *There exists a set of subsets of $\mathbb{R}^d$, we denote it $\mathfrak{B}^d$, and a function*

$$\lambda^d := \mathfrak{B}^d \longrightarrow \mathbb{R} \cup \{\infty\}; \qquad A \mapsto \lambda^d(A), \tag{4.2}$$

in the abstract sense of Definition 2.17 (Function) on p.33, such that

(A) *$\mathfrak{B}^d$ satisfies the following:*

$$\text{If } \idotsint_A d\vec{x} \text{ exists, then } A \in \mathfrak{B}^d, \quad \text{and} \quad \lambda^d(A) = \idotsint_A d\vec{x}, \tag{4.3}$$

$$\emptyset \in \mathfrak{B}^d, \quad \text{and} \quad \mathbb{R}^d \in \mathfrak{B}^d, \tag{4.4}$$

$$A \in \mathfrak{B}^d \quad \Rightarrow \quad A^{\complement} \in \mathfrak{B}^d, \tag{4.5}$$

$$A_n \in \mathfrak{B}^d \text{ for all } n \in \mathbb{N} \quad \Rightarrow \quad \bigcup_{n\in\mathbb{N}} A_n \in \mathfrak{B}^d, \quad \text{and} \quad \bigcap_{n\in\mathbb{N}} A_n \in \mathfrak{B}^d. \tag{4.6}$$

(B) *λ^d satisfies the following:*

$$A \in \mathfrak{B}^d \quad \Rightarrow \quad \lambda^d(A) \geq 0, \qquad \textit{(positivity)} \tag{4.7}$$

$$\lambda^d(\emptyset) = 0, \tag{4.8}$$

$$A, B \in \mathfrak{B}^d \text{ and } A \subseteq B \quad \Rightarrow \quad \lambda^d(A) \leq \lambda^d(B), \qquad \textit{(monotony)} \tag{4.9}$$

$$(A_n)_{n\in\mathbb{N}} \in \mathfrak{B}^d \text{ disjoint} \quad \Rightarrow \quad \lambda^d\Big(\biguplus_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \lambda^d(A_n). \qquad \textit{(σ–additivity)} \tag{4.10}$$

PROOF: Beyond the scope of this class. ■

Definition 4.2 (Borel sets [45]). ⋆

We call the elements of $\mathfrak{B}^d$ the **Borel sets** of $\mathbb{R}^d$. We also simply say that they **are Borel**. We call $B \in \mathfrak{B}^d$ **Lebesgue Null**, also, λ^d **Null**, if $\lambda^d(B) = 0$. □

[45] Named after the French mathematician and politician Émile Borel (full name: Félix Édouard Justin Émile Borel) (1871 – 1956)

Remark 4.1. When we introduce σ–algebras in Section 5.1 (Probability Spaces), $\mathfrak{B}^d$ turns out to be the σ–algebra which is generated by the d–dimensional rectangles. See Definition 5.6 on p.114. □

Example 4.1. The following shows how to work with some of the formulas of Theorem 4.1.

(a) (4.4) states that $\mathbb{R}^d \in \mathfrak{B}^d$. We could have omitted this part from Theorem 4.1, because it follows from $\emptyset^{\complement} = \mathbb{R}^d$ and ⊡ (4.4) $\emptyset \in \mathfrak{B}^d$ and ⊡ (4.5) $A \in \mathfrak{B}^d \;\Rightarrow\; A^{\complement} \in \mathfrak{B}^d$

(b) Alternatively, $\mathbb{R}^d \in \mathfrak{B}^d$ follows from (4.6), since $A_n := [-n, n]^d$ is a rectangle, thus Borel, and $\bigcup [A_n : n \in \mathbb{N}] = \mathbb{R}^d$.

(c) If $\vec{a} = (a_1, \dots, a_d) \in \mathbb{R}^d$, then the singleton $\{\vec{a}\}$ is Borel, and $\lambda^d\{\vec{a}\} = 0$:

$\{\vec{a}\} \in \mathfrak{B}^d$, since $\{\vec{a}\} = [a_1, a_1] \times [a_2, a_2] \times \cdots \times [a_d, a_d]$ is a rectangle and thus, Borel.

If that seems like cheating, one could also have expressed $\{\vec{a}\}$ as an intersection

$$\{\vec{a}\} \;=\; \bigcap_{n\in\mathbb{N}} A_n\,, \qquad \text{where} \qquad \Big]\, a_1 - \frac{1}{n},\, a_1 + \frac{1}{n}\Big[\times \cdots \times \Big]\, a_d - \frac{1}{n},\, a_d + \frac{1}{n}\Big[\tag{4.11}$$

of "proper" rectangles $A_n \in \mathfrak{B}^d$; thus, by (4.6), $\{\vec{a}\} \in \mathfrak{B}^d$. This proof is not as short, but (4.11) gives a quick way to prove that $\lambda^d\{\vec{a}\} = 0$:

By $\lambda^d(A_n) = 1/(2n)^d$ and $\emptyset \subseteq \{\vec{a}\} \subseteq A_n$ and (4.9), we have $0 = \lambda^d(\emptyset) \le \lambda^d\{\vec{a}\} \le 1/(2n)^d$ for all n. Since $1/(2n)^d \downarrow 0$ as $n \to \infty$, $\lambda^d\{\vec{a}\} = 0$. □

Theorem 4.2. ⋆

All countable subsets of $\mathbb{R}^d$ are Lebesgue Null. In particular, they are Borel sets.

PROOF: Let $B \subseteq \mathbb{R}^d$ be countable. Then

$$B \;=\; \{\vec{b}_1, \vec{b}_2, \dots\} \;=\; \{\vec{b}_1\} \uplus \{\vec{b}_2\} \uplus \cdots$$

for some finite or infinite sequence $\vec{b}_j$. We have seen in Example 4.1**(c)** that the singletons are Lebesgue Null sets. It follows from (4.6) that $\{\vec{b}_1\} \uplus \{\vec{b}_2\} \uplus \cdots$ is Borel and from (4.10) that it is Lebesgue Null. ■

Corollary 4.1. ⋆

(a) *All finite subsets of $\mathbb{R}^d$. In particular, all singleton sets $\{\vec{x}\}$ ($\vec{x} \in \mathbb{R}^d$), are Borel.*
(b) *adding and/or removing countably many points to/from a Borel set results in a Borel set.*

PROOF of **(a)**: Follows from Theorem 4.2 because finite sets are countable.

PROOF of **(b)**: Let $B \in \mathfrak{B}^d, U \subseteq \mathbb{R}^d$ countable. Then $U \in \mathfrak{B}^d$ by Theorem 4.2, because finite sets are countable It follows from (4.6) that $B \cup U \in \mathfrak{B}^d$ and $B \cap U \in \mathfrak{B}^d$. ■

Remark 4.2. ⋆ Only for this remark, let $\mathfrak{Rect}^d$ denote the set of all rectangles of $\mathbb{R}^d$, and let $\mathfrak{RiemInt}^d$ denote the set of all sets A in $\mathbb{R}^d$ such that $\mathbf{1}_A$ is Riemann integrable.

(a) Note that $\mathfrak{Rect}^d \subseteq \mathfrak{RiemInt}^d \subseteq \mathfrak{B}^d \subseteq 2^{\mathbb{R}^d}$:

- Rectangles in $\mathbb{R}^d$ are elements of $\mathfrak{RiemInt}^d$: Apply Definition 3.4 on p.66 with $\varphi = \mathbf{1}_A$.
- Elements of $\mathfrak{RiemInt}^d$ are Borel sets: That is the assertion of (4.3) in Theorem 4.1.
- $\mathfrak{B}^d \subseteq 2^{\mathbb{R}^d}$: Borel sets are subsets of $\mathbb{R}^d$, and $2^{\mathbb{R}^d}$ is the set of all subsets of $\mathbb{R}^d$. [46]

(b) (4.3) in Theorem 4.1 expresses that the extension of λ^d from $\mathfrak{Rect}^d$ to B^d is consistent with formula (4.1) of Definition 4.1 on p.85, which extends λ^d from $\mathfrak{Rect}^d$ (only) to $\mathfrak{RiemInt}^d$.

(c) There are Borel sets with infinite Lebesgue measure. For example, $\mathbb{R}^d \in \mathfrak{B}^d$, and $\lambda^d(\mathbb{R}^d) = \infty$.

(d) All set inclusions in **(a)** are strict, i.e., we have $\mathfrak{Rect}^d \subsetneq \mathfrak{RiemInt}^d \subsetneq \mathfrak{B}^d \subsetneq 2^{\mathbb{R}^d}$:

- $\mathfrak{Rect}^d \subsetneq \mathfrak{RiemInt}^d$ is true, because, e.g., the union of two disjoint rectangles R_1 and R_2, has Riemann integral $\iint\cdots\int_{R_1\cup R_2} d\vec{x} = \iint\cdots\int_{R_1} d\vec{x} + = \iint\cdots\int_{R_2} d\vec{x}$.
- For $\mathfrak{RiemInt}^d \subsetneq \mathfrak{B}^d$, consider the set $A := \mathbb{Q} \cap [0,1]$. Since $A \subseteq \mathbb{Q}$ is countable, A is Borel by Theorem 4.2 on p.86. On the other hand, we have seen in Example 3.6 on p.69 that the Riemann sums for $\mathbf{1}_A$ do not have a limit $\lim_{\|\Pi\|\to 0} \mathcal{RS}(\mathbf{1}_A;\Pi)$. Hence, A is not Riemann integrable.
- The proof that $\mathfrak{B}^d \subsetneq 2^{\mathbb{R}^d}$, i.e., Lebesgue measure cannot be reasonably defined for all subsets of $\mathbb{R}^d$, is very sophisticated and cannot be given here. All sets of interest for this course are Borel. This justifies the following:

> Unless something different is explicitly stated, all sets $B \subseteq \mathbb{R}^d$ we deal that with in this course may be assumed to be Borel Thus, $\lambda^d(B)$ exists (but might be infinite).

Only completely weird and useless subsets of $\mathbb{R}^d$ are not Borel. □

4.2 The Lebesgue Integral

Definition 4.3 (Simple Function on $\mathbb{R}^d$)**.**

Let $d, n \in \mathbb{N}$. Let $A_1, \ldots, A_n$ be Borel sets of $\mathbb{R}^d$. (Thus, $\lambda^d(A_j)$ is defined for all A_j.) Further, let $c_1, c_2, \ldots, c_n$ be a corresponding set of non–negative real numbers. Let

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}\,; \qquad \vec{x} \mapsto f(\vec{x}) := \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\vec{x}) \tag{4.12}$$

Then we call f a **simple function**. □

Proposition 4.1. ⋆

(a) *All step functions are simple functions.*
(b) *Not all simple functions are step functions.*
(c) *Not all simple functions possess a Riemann integral.*

PROOF of **(a)**: This is trivially true, since rectangles in $\mathbb{R}^d$ are Borel. See Remark 4.2**(a)**.

[46] Recall Definition 2.9 (power set) on p.24.

PROOF of **(b)**: The set $A := \mathbb{Q} \cap [0,1]$ obviously cannot be written as a finite union of onedimensional rectangles (intervals). Thus, $x \mapsto \mathbf{1}_A(x)$ is not a step function. On the other hand, A is Borel as a countable set. See Theorem 4.2. We set $n = 1, c_1 = 1,\ A_1 = A$ and see that

$$\mathbf{1}_A(x) \;=\; 1 \cdot \mathbf{1}_A(x) \;=\; \sum_{j=1}^{1} c_j \cdot \mathbf{1}_{A_j}(x),$$

is a simple function.

PROOF of **(c)**: Again, let $A := \mathbb{Q} \cap [0,1]$. We just have established that $f := \mathbf{1}_A$ is a simple function. We also have seen in Example 3.6 on p.69 that the Riemann integral

$$\int_a^b f(x)dx \;=\; \lim_{\|\Pi\|\to 0} \mathcal{RS}(f;\Pi)$$

does not exist for this function. ■

The next definition is very important and you must remember it.

Definition 4.4.

Let $f(\vec{x}) \;=\; \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\vec{x})$ be a simple function such that $c_j \geq 0$ for all j. Then we call

(4.13) $$\int f\, d\lambda^d \;:=\; \int f(\vec{x}) d\lambda^d(\vec{x}) \;:=\; \int f(\vec{x}) \lambda^d(d\vec{x}) \;:=\; \sum_{j=1}^{n} c_j \lambda^d(A_j)\,.$$

the **Lebesgue integral** of the simple function f. □

Remark 4.3 (Construction of the Lebesgue integral)**.** Compare the following to the construction of the Riemann integral.

(a) All step functions are simple functions. (See Proposition 4.1**(a)** on p.87.)
(b) Lebesgue integral and Riemann integral are identical for step functions. (Compare (3.14) on p.66 with (4.13) above. That bodes well for making them both identical for at least all those functions which possess a proper Riemann integral. □

Remark 4.4. ⋆ We just mentioned that Definition 4.4 mirrors Definition 3.4 on p.66 of the Riemann integral of a step function. But note the following differences.

(a) The rectangles that appear in a step function have finite Lebesgue measure, whereas the Borel sets of a simple function are allowed to have infinite Lebesgue measure.
That is precisely the reason for requiring in Definition 4.4 that $c_j \geq 0$ for all j: This condition ensures that there is no occurrence of $\infty - \infty$ on the right side of $\int f d\lambda^d \;=\; \sum_{j=1}^{n} c_j \lambda^d(A_j)$.

(b) Since the Borel sets of a simple function need not be disjoint, there can be different choices of n, c_j, A_j that yield the same simple function $f(\vec{x}) \;=\; \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\vec{x})$. It can be shown that they all result in the same number $\sum_{j=1}^{n} c_j \lambda^d(A_j)$. Thus, the expression for $\int f d\lambda^d$ given in (4.13) is well defined. □

Remark 4.5. ⋆ We obtained the Riemann integral for general functions from that of step functions as limits

$$\iint\cdots\int_R f(\vec{y})\,d\vec{y} \;:=\; \lim_{\|\Pi\|\to 0} \mathcal{RS}(f;\Pi)\,,$$

where the Riemann sums (3.35) (see p.73) are the Riemann integrals of step functions (defined on d–dimensional rectangles, R). Those limits were obtained by **dividing the domain** into finer and finer partitions.

We create the Lebesgue integral for more general functions $f \geq 0$, by **subdividing the codomain** rather than the domain into finer and finer partitions. We then approximate f by a sequence $f_n \uparrow f$ (i.e., $f_n(\vec{x}) \uparrow f(\vec{x})$ for all $\vec{x}$), of simple functions f_n with Lebesgue integral $\int f_n d\lambda^d$, given by (4.13).

This procedure for creating the functions f_n is surprisingly simple: Fix $n \in \mathbb{N}$, and define, for $k \in \mathbb{N}$,

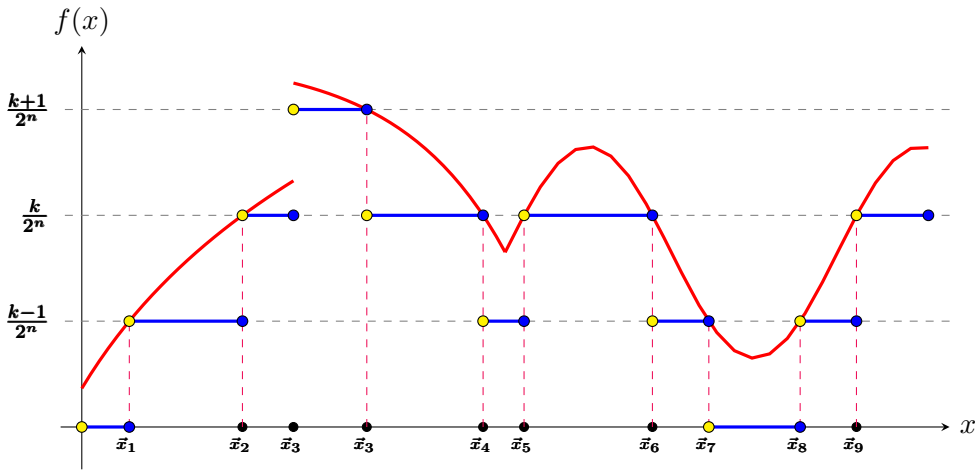
$$I_{k,n} \;:=\; \left]\, \frac{k-1}{2^n}\,,\, \frac{k}{2^n} \,\right] .$$

Note that $[0,\infty[= \{0\} \uplus \left(\biguplus[\, I_{k,n} : k \in \mathbb{Z}\,]\right)$ partitions the codomain into small intervals. Let

$$A_{k,n} \;:=\; \left\{ \vec{x} \in \mathbb{R}^d : \frac{k-1}{2^n} \;<\; f(\vec{x}) \;\leq\; \frac{k}{2^n} \right\} \quad (k = 1,\ldots,4^n),$$

Note that $\vec{x} \in A_{k,n} \;\Leftrightarrow\; (k-1)/2^n < f(\vec{x}) \leq k/2^n$. Next, we define

$$f_n(\vec{x}) \;:=\; \sum_{k=1}^{4^n} \frac{k-1}{2^n} \cdot \mathbf{1}_{A_{k,n}}(\vec{x}). \tag{4.14}$$



The picture above demonstrates how the simple functions $f_n \uparrow f$ are constructed. Observe that

$$f_n(\vec{x}) \;=\; \frac{k-1}{2^n} \;\text{ on } A_{k,n} \;=\; \left\{ \vec{x} \in \mathbb{R}^d : \frac{k-1}{2^n} \;<\; f(\vec{x}) \;\leq\; \frac{k}{2^n} \right\}.$$

(Here, $A_{k,n} =]\vec{x}_1,\vec{x}_2] \cup]\vec{x}_4,\vec{x}_5] \cup]\vec{x}_6,\vec{x}_7] \cup]\vec{x}_8,\vec{x}_9]$.) Further, $0 \leq f(\vec{x}) - f_n(\vec{x}) \leq \frac{1}{2^n}$, for $\vec{x} \in A_{k,n}$. Let $A_0 := \{\vec{x} \in \mathbb{R}^d : f(\vec{x}) = 0\}$. Since $f \geq 0$, (4.14) implies that $f_n(\vec{x}) = f(\vec{x}) = 0$ on A_0, we see that

$$0 \;\leq\; f(\vec{x}) - f_n(\vec{x}) \;\leq\; \frac{1}{2^n}\,, \quad \text{for} \quad \vec{x} \in A_0 \cup A_{1,n} \cup A_{2,n} \cup \cdots \cup A_{4^n,n}\,.$$

Since $1 \leq k \leq 4^n$ is equivalent to $0 \leq (k-1)/2^n < k/2^n \leq 4^n/2^n = 2^n$, we obtain

$$0 \le f(\vec{x}) - f_n(\vec{x}) \le \frac{1}{2^n}, \quad \text{for} \quad f(\vec{x}) \le 2^n.$$

Finally, since $f(\vec{x}) < \infty$ for all $\vec{x} \in \mathbb{R}^d$ and $2^{-n} \to 0$ and $2^n \to \infty$ as $n \to \infty$, we conclude that

$$f_n(\vec{x}) \uparrow f(\vec{x}), \quad \text{for} \quad \vec{x} \in \mathbb{R}^d.$$

It is not difficult to show for two simple functions $0 \le \varphi \le \psi$, that $\int \varphi d\lambda^d \le \int \psi d\lambda^d$. Accordingly, the sequence $\int f_n d\lambda^d$ (those are real numbers!) is nondecreasing. Thus,

$$\int f_n \, d\lambda^d \uparrow \lim_{n\to\infty} \int f_n \, d\lambda^d = \sup_{n\in\mathbb{N}} \int f_n \, d\lambda^d .^{47} \qquad \text{(Not guaranteed to be finite.)}$$

One can prove the following.

Let $f, f_n, \tilde{f}_n : \mathbb{R}_d \to [0, \infty[$ as follows. f_n and $\tilde{f}_n$ are two sequences of simple functions both of which satisfy $f_n \uparrow f$ and $\tilde{f}_n \uparrow f$. Then we have equal limits,

$$\lim_{n\to\infty} \int f_n \, d\lambda^d = \lim_{n\to\infty} \int \tilde{f}_n \, d\lambda^d .$$

This makes part **(a)** of the next definition possible. □

Recall Definition 2.15 (Absolute value, positive and negative part) on p.29 and the subsequent Remark 2.9: Any real–valued function f (with arbitrary domain) can be written as the difference

$$f(x) = f^+(x) - f^-(x)$$

of the nonnegative functions

$$f^+(x) = \max\big(f(x), 0\big), \qquad f^-(x) = -\min\big(-f(x), 0\big).$$

Definition 4.5 (Lebesgue integral). ⋆

(a) Either let $f : \mathbb{R}_d \to [0, \infty[$ be a nonnegative function on $\mathbb{R}_d$, such that

- there is a nondecreasing sequence of simple functions, $f_n \ge 0$, satisfying $f_n \uparrow f$;

Or let $f : \mathbb{R}_d \to]-\infty, 0]$ be a nonpositive function on $\mathbb{R}_d$, such that

- there is a nonincreasing sequence of simple functions, $f_n \le 0$, satisfying $f_n \downarrow f$.

We define the **Lebesgue integral** of that nonnegative or nonpositive function f as

$$\int f \, d\lambda^d := \lim_{n\to\infty} \int f_n \, d\lambda^d . \tag{4.15}$$

(b) Let $f : \mathbb{R}_d \to \mathbb{R}$ be a function on $\mathbb{R}_d$ such that

- both f^+ and f^- are limits of nondecreasing sequences of simple functions ≥ 0;
- at least one of $\int f^+ d\lambda^d$, $\int f^- d\lambda^d$ is finite. (According to **(a)**, those integrals exist, but neither of them was guaranteed to be finite.)

Then we define the **Lebesgue integral** of the function f as the expression

$$\int f \, d\lambda^d = \int (f^+ - f^-) \, d\lambda^d := \int f^+ \, d\lambda^d - \int f^- \, d\lambda^d . \tag{4.16}$$

[47] See Theorem 2.3 on p.50.

(c) We call a real–valued function f **Lebesgue integrable**, if $\int f\,d\lambda^d$ exists and is finite. □

Remark 4.6. ⋆

(a) We remind you that the sets $A_1, \ldots, A_n$ that belong to a simple function, $\sum_{j=1}^{n} c_j \mathbf{1}_{A_j}$ are not arbitrary subsets of $\mathbb{R}^n$. Rather, they must be Borel sets.

(b) It is not hard to see that sums and differences of simple functions are simple functions and that the following is true for real–valued functions f, g which are limits of simple functions $f_n \to f$, $g_n \to g$ on $\mathbb{R}^d$ (but **not necessarily** $f_n \uparrow f$ and/or $g_n \uparrow g$):

- $\lim_{n\to\infty} f_n = f$ and $\lim_{n\to\infty} g_n = g \;\Rightarrow\; \lim_{n\to\infty}(f_n \pm g_n) = f \pm g$.

(c) In particular, the functions f of Definition 4.5**(b)** are limits of simple functions, since we assumed so for f^+ and f^-, and $f = f^+ - f^-$.

(d) Thus, all functions f for which we have defined their Lebesgue integral are limits of sequences of simple functions.

(e) $\int f^+\,d\lambda^d = \infty$ (thus, $\int f^-\,d\lambda^d < \infty$) $\Rightarrow$ $\int f\,d\lambda^d = \infty$.
$\int f^-\,d\lambda^d = \infty$ (thus, $\int f^+\,d\lambda^d < \infty$) $\Rightarrow$ $\int f\,d\lambda^d = -\infty$.

(f) As far as integrability is concerned, we follow the same rule for the Lebesgue integral as for the Riemann integral: It is not sufficient that the integral exists. Moreover, it also must be finite. See Definition 3.9 (Riemann integrability) on p.75.

(g) The Lebesgue integral satisfies many important formulas. We will list them in Theorem 4.5 on p.94, after we have defined how to integrate over subsets of $\mathbb{R}^d$. □

Considering Remark 4.6**(d)**, limits of simple functions deserve a special name.

Definition 4.6. ⋆

- We call simple functions, and real–valued functions that are limits of sequences of simple functions, **Borel measurable functions** (or simply, **Borel functions**). □

Remark 4.7. ⋆ Let f_j be a sequence of simple functions and $f(x) := \lim_{n\to\infty} f_n(x)$. We mention in passing that this limit $f(x)$ is allowed to take values $\pm\infty$ for some or all x. We will generally gloss over the issues that this might entail. □

The next theorem asserts that about anything that can be done with a countable collection of Borel functions results again in a Borel function. Note that we have suppressed the arguments in the functions listed there. For example, $\max(f_1, f_2)$ is the function $\vec{x} \mapsto \max\big(f_1(\vec{x}),\, f_2(\vec{x})\big)$, and $\sum_{j=1}^{\infty} f_j$ is the function $\vec{x} \mapsto \sum_{j=1}^{\infty} f_j(\vec{x})$.

Theorem 4.3. ⋆ *Assume that $f_1, f_2, \ldots$ are Borel functions, $c_1, c_2, \cdots \in \mathbb{R}$, $B \in \mathfrak{B}^d$.*

Each of the following also is a Borel function:

- c_1 *(constant function)* • $c_1 f_1$ • $f_1 \pm f_2$ • $f_1 f_2$ • $\mathbf{1}_B f_1$ • f_1/f_2 *(if* $f_2 \neq 0$*)* • $\sum_{j=1}^{n} c_j f_j$
- $\min(f_1, f_2)$ • $\max(f_1, f_2)$ • $\min_{j=1,\ldots,n} f_j$ • $\max_{j=1,\ldots,n} f_j$ • $\inf_{j\in\mathbb{N}} f_j$ • $\sup_{j\in\mathbb{N}} f_j$ □

If they exist (see the subsequent remark), the following also are Borel functions:

- $\lim_{j\to\infty} f_j$ • $\sum_{j=1}^{\infty} f_j$ • $\min_{j\in\mathbb{N}} f_j$ • $\max_{j\in\mathbb{N}} f_j$

PROOF: ■

Remark 4.8. ⋆ Theorem 4.3 **(i)** asserts that $\lim_{j\to\infty} f_j$, $\sum_{j=1}^{\infty} f_j$, $\min_{j\in\mathbb{N}} f_j$, $\max_{j\in\mathbb{N}} f_j$ may not exist and **(ii)** does not raise an issue with $\inf_{j\in\mathbb{N}} f_j$ and $\sup_{j\in\mathbb{N}} f_j$. Let us take a look at both points.

(a) For $x \in \mathbb{R}$ and $j \in \mathbb{N}$, let $h_j(x) := (-1)^j x$. Let $f_n(x) := \sum_{j=1}^{n} h_j(x) = -x + x - x + x \ldots$. Thus, $\lim_{j\to\infty} f_j(x)$ does not exist for $x \neq 0$.

(b) For $x \in \mathbb{R}$ and $j \in \mathbb{N}$, let $f_j(x) := (-1)^j x$. Then $\sum_{j=1}^{\infty} f_j(x) = -x + x - x + x \ldots$. does not exist for $x \neq 0$.

(c) For $x \in \mathbb{R}$ and $j \in \mathbb{N}$, let $f_j(x) := 1/n$. (Each function f_j is constant in x.) Then $\inf_{j\in\mathbb{N}} f_j(x) = 0$, but $\min_{j\in\mathbb{N}} f_j(x)$ does not exist for any x.

(d) For $x \in \mathbb{R}$ and $j \in \mathbb{N}$, let $f_j(x) := 1 - 1/n$. Then $\sup_{j\in\mathbb{N}} f_j(x) = 1$, but $\max_{j\in\mathbb{N}} f_j(x)$ does not exist for any x.

Examples **(c)** and **(d)** also illustrate why inf and sup are not a concern: Any sequence of real numbers (and that's what we have for fixed x) has an inf (might be $-\infty$) and a sup (might be ∞). □

Remark 4.9. We stated in Remark 4.2**(d)** on p.86 that »Only completely weird and useless sets are not Borel« and that »All sets $B \subseteq \mathbb{R}^d$ we deal that with in this course may be assumed to be Borel.« The same can be said about the Borel functions of $\mathbb{R}^d$. This justifies the following.

Unless something different is explicitly stated, all real–valued functions defined on subsets of $\mathbb{R}^d$ that we deal that with in this course may be assumed to be Borel. □

For the next theorem, recall that the product of a Borel set and a Borel function is a Borel function. [48]

[48] see Theorem 4.3 on p.92

Theorem 4.4. ⋆ *Lebesgue integrals satisfy the following.*

Let $B \in \mathfrak{B}^d$ and assume that f is a Borel function. Then

(a) *If $\int f\, d\lambda^d$ exists, then $\int \mathbf{1}_B\, f\, d\lambda^d$ exists.*

(b) *If f is Lebesgue integrable, then $\mathbf{1}_B f$ is Lebesgue integrable.*

PROOF: ■

This last theorem allows us to make the following definition. (NOT optional!)

Definition 4.7.

Let $B \in \mathfrak{B}^d$ and assume that f is a Borel function on $\mathbb{R}^d$ for which the Lebesgue integral $\int f d\lambda^d$ exists. The **Lebesgue integral of f on B** or **over B** is defined by the expression

$$\int_B f\, d\lambda^d \;:=\; \int_B f(\vec{x}) d\,\lambda^d(\vec{x}) \;:=\; \int_B f(\vec{x}) \lambda^d(d\vec{x}) \;:=\; \int \mathbf{1}_B f\, d\lambda^d\,. \tag{4.17}$$

We say that **Lebesgue integrable on B**, if $\int_B f\, d\lambda^d$ exists and is finite. □

Fact 4.1. *Let $B \subseteq \mathbb{R}^d$ and $f : B \to \mathbb{R}$, such that f and B are of any relevance for this course.*

- *If the Riemann integral $\int_B f(\vec{x})\, d\vec{x}$ exists, then the Lebesgue integral $\int_B f\, d\lambda^d$ exists.*
- *Further, $\int_B f(\vec{x})\, d\vec{x} \;=\; \int_B f\, d\lambda^d$.*
- *Accordingly, all the techniques one has learned in calculus to evaluate the Riemann integral can be used to compute the Lebesgue integral.* □

Be sure to master the following trivial example.

Problem 4.1. Evaluate the following Lebesgue integrals.

(1) $\int_{[0,\infty[} e^{-3t} d\lambda^1$ **(2)** $\int_{[2,5]} 4x^2 y\, \lambda^1(dy)$ **(3)** $\int_{[1,2]\times[2,5]} 4x^2 y\, d\lambda^2$

Solution for **(1):** We compute the Riemann integral

$$\int_0^\infty e^{-3t}\, dt \;=\; \frac{-1}{3} e^{-3t}\Big]_0^\infty \;=\; \frac{-1}{3}(0-1) \;=\; \frac{1}{3}\,.$$

Solution for **(2):** Note how the notation $\int \cdots \lambda^1(dy)$ leaves no doubt that the integration variable is y. We compute the Riemann integral

$$\int_2^5 4x^2 y\, dy \;=\; \frac{4x^2}{2} \cdot y^2\Big]_{y=2}^5 \;=\; 2x^2 \cdot 21 \;=\; 42x^2\,.$$

Solution for **(3):** We compute the 2 dimensional Riemann integral

$$\int_{x=1}^2 \int_{y=2}^5 4x^2 y\, dy\, dx \;=\; \int_{x=1}^2 \frac{4x^2}{2} \cdot y^2\Big]_{y=2}^5 dx \;=\; \int_1^2 42x^2\, dx \;=\; \frac{42}{3} \cdot x^3\Big]_1^2 \;=\; 14 \cdot 7 \;=\; 98\,.$$ □

For the sake of completeness, we will give in Remark 4.11 on p.96 below an example of a function which has a finite (but improper) Riemann integral which does not possess a Lebesgue integral, since that one would be of the form $\infty - \infty$. This is related to the following proposition.

Proposition 4.2 (Integrability criterion). ⋆ *Let f be a Borel function and B a Borel set. Then*

$$f \text{ is integrable on } B \;\Leftrightarrow\; \int_B |f|\,d\lambda^d < \infty \;\Leftrightarrow\; \text{both } \int_B f^+\,d\lambda^d < \infty \text{ and } \int_B f^-\,d\lambda^d < \infty.$$

PROOF: ■

You are familiar with **(a)** and **(b)** of the next theorem from the Riemann integral. **(c)** and **(d)** are the properties that make the Lebesgue integral so much more powerful than the Riemann integral.

Theorem 4.5. *Assume that $f, g, f_1, f_2, \ldots$ are Borel functions, $c, c_1, c_2, \cdots \in \mathbb{R}$, and B is a Borel set. Then Lebesgue integrals on B satisfy the following.*

(a) Positivity: $\displaystyle\int_B 0\,d\lambda^d = 0; \qquad f \geq 0 \text{ on } B \;\Rightarrow\; \int_B f\,d\lambda^d \geq 0,$

(b) Monotonicity: $\lambda^d\{\vec{x} \in B : f(\vec{x}) > g(\vec{x})\} = 0 \;\Rightarrow\; \displaystyle\int_B f\,d\lambda^d \leq \int_B g\,d\lambda^d.$

In particular, $f \leq g \text{ on } B \;\Rightarrow\; \displaystyle\int_B f\,d\lambda^d \leq \int_B g\,d\lambda^d,$

and also, $\lambda^d\{\vec{x} \in B : f(\vec{x}) \neq g(\vec{x})\} = 0 \;\Rightarrow\; \displaystyle\int_B f\,d\lambda^d = \int_B g\,d\lambda^d.$

(c) Linearity I: *f, g integrable on B* $\Rightarrow \displaystyle\int_B (f \pm g)\,d\lambda^d = \int_B f\,d\lambda^d \pm \int_B g\,d\lambda^d$

and also, $\displaystyle\int_B (cf)\,d\lambda^d = c\int_B f\,d\lambda^d.$

Linearity II: *$f_1 \ldots, f_n$ integrable* $\Rightarrow \displaystyle\int_B \Big(\sum_{j=1}^n f_j\Big)\,d\lambda^d = \sum_{j=1}^n c_j \int_B f_j\,d\lambda^d.$

(d) Monotone Convergence: *Assume that* $0 \leq f_1 \leq f_2 \leq \cdots$, $0 \geq g_1 \geq g_2 \geq \cdots$.

Then $\displaystyle\int_B f_n\,d\lambda^d \uparrow \int_B \Big(\sup_{n\in\mathbb{N}} f_n\Big)\,d\lambda^d$ *and* $\displaystyle\int_B g_n\,d\lambda^d \downarrow \int_B \Big(\inf_{n\in\mathbb{N}} g_n\Big)\,d\lambda^d$ *as* $n \to \infty$.

(e) Dominated Convergence: *Assume that*

- $\lim_{n\to\infty} f_n$ *exists,*
- $|f_n| \leq g$ *for all* $n \in \mathbb{N}$,
- $\displaystyle\int_B g\,d\lambda^d < \infty.$

Then $\displaystyle\lim_{n\to\infty} \int_B f_n\,d\lambda^d = \int_B \Big(\lim_{n\to\infty} f_n\Big)\,d\lambda^d$ *as* $n \to \infty$.

PROOF: ■

Remark:

Remark 4.10.

(a) We will refer to Theorem 4.5**(d)** as the **monotone convergence theorem** and to Theorem 4.5**(e)** as the **dominated convergence theorem** for Lebesgue integrals.

(b) Clearly, the dominated convergence is about switching integrals and limits of a function sequence. Note that so is the monotone convergence theorem, since [49]

$$\bullet\ f_n \uparrow \Rightarrow \sup_{n\in\mathbb{N}} f_n = \lim_{n\to\infty} f_n, \quad \bullet\ g_n \downarrow \Rightarrow \inf_{n\in\mathbb{N}} g_n = \lim_{n\to\infty} g_n.$$

Thus, the monotone convergence formulas of Theorem 4.5 can be written

$$\int_B f_n\, d\lambda^d \uparrow \int_B \left(\lim_{n\to\infty} f_n\right) d\lambda^d; \text{ and } \int_B g_n\, d\lambda^d \downarrow \int_B \left(\lim_{n\to\infty} g_n\right) d\lambda^d, \text{ as } n\to\infty.$$

(c) Note for the dominated convergence theorem, that $|f_n| \le g$ implies $g > 0$. □

Theorem 4.6 (Fubini's theorem for Lebesgue integrals). ⋆ *Assume that $f_1, f_2, \dots$ are Borel functions, and B_1, B_2 are Borel sets. Then, for any rearrangement $j_1, j_2, \dots, j_d$ of $1, 2, \dots, d$,*

$$\begin{aligned}\int_{B_1\times B_2\times\cdots\times B_d} f\, d\lambda^d &= \int_{B_1}\left(\int_{B_2}\left(\cdots\int_{B_d} f\, d\lambda^1\cdots\right) d\lambda^1\right) d\lambda^1 \\ &= \int_{B_{j_1}}\left(\int_{B_{j_2}}\left(\cdots\int_{B_{j_d}} f\, d\lambda^1\cdots\right) d\lambda^1\right) d\lambda^1\end{aligned} \tag{4.18}$$

This formula is technically correct, but let us supply all arguments and write, [50] *e.g., $\lambda^1(dx_j)$ for $d\lambda^1$:*

$$\begin{aligned}\int_{B_1\times B_2\times\cdots\times B_d} f(\vec{x})\,\lambda^d(d\vec{x}) &= \int_{B_1}\left(\int_{B_2}\left(\cdots\int_{B_d} f(\vec{x})\,\lambda^1(dx_d)\cdots\right)\lambda^1(dx_2)\right)\lambda^1(dx_1) \\ &= \int_{B_{j_1}}\left(\int_{B_{j_2}}\left(\cdots\int_{B_{j_1}} f(\vec{x})\,\lambda^1(dx_{j_d})\cdots\right)\lambda^1(dx_{j_2})\right)\lambda^1(dx_{j_1})\,.\end{aligned} \tag{4.19}$$

*In particular, assume that each B_j is an interval $[\alpha_j, \beta_j]$ or $[\alpha_j, \beta_j]$ or $[\alpha_j, \beta_j]$ or $[\alpha_j, \beta_j]$, where $\alpha_j \le \beta_j$. If we adjust the notation to that of Riemann integrals and replace $\int_{B_j}$ with $\int_{\alpha_j}^{\beta_j}$, $\lambda^d(d\vec{x})$ with $d\vec{x}$, and $\lambda^1(dx_j)$ with dx_j, then (4.19) matches Fubini's formula (4.1)**(g)** for Riemann integrals (see p.83).*

Here is another version of Fubini's theorem. It features "only" two vector– valued components.

[49] by Theorem 2.3 on p.50

[50] Recall that (4.4) on p.88 and (4.7) on p.93 give us a choice of notation

$$\int_B f\, d\lambda^d = \int_B f(\vec{x})\, d\lambda^d(\vec{x}) = \int_B f(\vec{x})\lambda^d(d\vec{x})$$

Assume that $d, d_1, d_2 \in \mathbb{N}$, that $d_1 + d_2 = d$, that $f : \mathbb{R}^d \to \mathbb{R}$ is a nonnegative and/or λ^d–integrable Borel function, and that $B_1 \in \mathfrak{B}^{d_1}$ and $B_2 \in \mathfrak{B}^{d_2}$. For $\vec{x} = (x_1, x_2 \ldots, x_{d_1})$ and $\vec{y} = (y_1, y_2 \ldots, y_{d_2})$, let $(\vec{x}, \vec{y}) := (x_1, \ldots, x_{d_1}, y_1, \ldots, y_{d_2})$. Then

$$\begin{aligned} \int_{B_1 \times B_2} f(\vec{x}, \vec{y})\, \lambda^d\big(d(\vec{x}, \vec{y})\big) &= \int_{B_1} \left(\int_{B_2} f(\vec{x}, \vec{y})\, \lambda^{d_2}(d\vec{y}) \right) \lambda^{d_1}(d\vec{x}) \\ &= \int_{B_2} \left(\int_{B_1} f(\vec{x}, \vec{y})\, \lambda^{d_1}(d\vec{x}) \right) \lambda^{d_2}(d\vec{y})\,. \end{aligned} \tag{4.20}$$

Even though there only are two integrations $\lambda^{d_1}(d\vec{x})$ and $\lambda^{d_2}(d\vec{y})$, (4.20) is more general than (4.19), because the Borel sets B_1, B_2, and $B_1 \times B_2$ are no more cartesian products of onedimensional Borel sets.

PROOF: ■

Remark 4.11. ⋆ Here is a curiosity, an example of a function that a Riemann integral but not a Lebesgue integral. Let $f(x) := \mathbf{1}_{[0,\infty[}\dfrac{\sin x}{x}$. This function has the following properties.

(a) It has the following (improper) Riemann integrals:

$$\int_0^\infty f^+(x)\, dx \;=\; \int_0^\infty f^-(x)\, dx \;=\; \int_0^\infty |f(x)|\, dx \;=\; \infty\,.$$

(b) The Lebesgue integral $\displaystyle\int_{[0,\infty[} f\, d\lambda^1$ does not exist.

(c) It has the (improper) Riemann integral $\displaystyle\int_0^\infty \frac{\sin x}{x}\, dx \;=\; \frac{\pi}{2}\,.$

PROOF of **(a)**: [51] We will reference the following below:

(A) $(2j+1)\pi < x < 2j\pi \;\Rightarrow\; \sin x < 0 \;\Rightarrow\; f^+(x) = 0$,
(B) $2j\pi < x < (2j+1)\pi \;\Rightarrow\; \sin x > 0 \;\Rightarrow\; f^-(x) = 0$,
(C) $\sum 1/j \;=\; \infty$ (harmonic series).

$$\begin{aligned} \int_0^\infty f^+(x)\, dx \;&=\; \int_0^\infty \frac{\sin^+ x}{x}\, dx \overset{\text{(A)}}{=} \sum_{j=0}^\infty \int_{2j\pi}^{(2j+1)\pi} \frac{\sin x}{x}\, dx \;\geq\; \sum_{j=0}^\infty \int_{2j\pi}^{(2j+1)\pi} \frac{\sin x}{(2j+1)\pi}\, dx \\ &= \sum_{j=0}^\infty \frac{1}{(2j+1)\pi} \int_{2j\pi}^{(2j+1)\pi} \sin x\, dx \;=\; \sum_{j=0}^\infty \frac{1}{(2j+1)\pi} \big(-\cos x\big)\Big]_{2j\pi}^{(2j+1)\pi} \\ &= \sum_{j=0}^\infty \frac{2}{(2j+1)\pi} \;=\; \frac{2}{\pi} \sum_{j=1}^\infty \frac{1}{2j-1} \;\geq\; \frac{2}{\pi} \sum_{j=1}^\infty \frac{1}{2j} \;=\; \frac{1}{\pi} \sum_{j=1}^\infty \frac{1}{j} \overset{\text{(C)}}{=} \infty\,. \end{aligned}$$

[51] Source: Showing $\frac{\sin x}{x}$ is NOT Lebesgue integrable on $\mathbb{R}_{\geq 0}$.

Similarly,

$$\begin{aligned}\int_0^\infty f^-(x)\,dx &= \int_0^\infty \frac{\sin^- x}{x}\,dx \overset{\text{(B)}}{=} \sum_{j=1}^\infty \int_{(2j-1)\pi}^{2j\pi} \frac{\sin x}{x}\,dx \geq \sum_{j=1}^\infty \int_{(2j-1)\pi}^{2j\pi} \frac{\sin x}{2j\pi}\,dx \\ &= \sum_{j=1}^\infty \frac{1}{2j\pi} \int_{(2j-1)\pi}^{2j\pi} \sin x\,dx = \sum_{j=1}^\infty \frac{1}{2j\pi}\,(-\cos x)\Big]_{(2j-1)\pi}^{2j\pi} \\ &= \sum_{j=1}^\infty \frac{2}{2j\pi} = \frac{1}{\pi}\sum_{j=1}^\infty \frac{1}{j} \overset{\text{(C)}}{=} \infty\,.\end{aligned}$$

PROOF of **(b)**: Since $\int f^+ d\lambda^1 = \int f^- d\lambda^1 = \infty$, we see from **(a)** that

$$\int f d\lambda^1 = \int f^+ d\lambda^1 - \int f^- d\lambda^1 = \infty - \infty\,.$$

Thus, the Lebesgue integral $\int f d\lambda^1$ does not exist.

PROOF of **(c)**: Will not be given. [52] □

Remark 4.12. The monotone convergence and dominated convergence theorems are very powerful and you are encouraged to consider them when you want to compute the limit of a sequence of integrals, $\lim_{n\to\infty} \int_B f_n d\lambda^d$, or the integral of the limit of a function sequence, $\int_B \big(\lim_{n\to\infty} f_n\big) d\lambda^d$.

However, you must always check whether the conditions are met!

Monotone convergence:

- Is $f_n(x) \geq 0$ for all n and all $x \in B$?
- Is the sequence $\big(f_n(x)\big)_n$ nondecreasing for all $x \in B$?

Dominated convergence:

- Does $\lim_{n\to\infty} f_n(x)$ exist for all $x \in B$?
- Is there $x \to g(x)$ such that $\int_B g d\lambda^d < \infty$ and $|f_n(x)| \leq g(x)$ for all $x \in B$?

  Equivalently: Let $h(x) := sup_n |f_n(x)|$. Is $\int_B h d\lambda^d < \infty$?) □

Problem 4.2. Neither monotone convergence nor dominated convergence can be applied for the following sequences.

(a) Let $f_n(x) := \mathbf{1}_{[n,\infty[}(x)$. Note that $f_n \geq 0$ and $f_n \downarrow$ (rather than $f_n \uparrow$) on $\mathbb{R}$.
Compute $\lim_n \int f_n\,d\lambda^d$ and $\int (\lim_n f_n)\,d\lambda^d$.

(b) Let $B := [0,\infty[$ and $f_n(x) := \mathbf{1}_B(x)[e^{-x} + (1/n)e^{-x/n}]$. Clearly, $f_n \geq 0$ on B.
Is it true that f_n is nondecreasing? If $h(x) = \sup f_n(x)$, is $\int_B h d\lambda^d < \infty$?

[52] A proof can be found in Socratic Q&A: Integration of sinx/x from 0 to infinity?. It uses techniques from complex analysis and is beyond the scope of this course.

Solution for **(a):**

(1) First, observe that $f_n(x) \downarrow 0$ for all $x \in \mathbb{R}$:
This is obvious for $x < 0$, since then $f(x) = 0$ for all n.
Fix $x \geq 0$ and observe that $n > x \;\Rightarrow\; x \notin [n, \infty[\;\Rightarrow\; f_n(x) = 0$. Thus, $\lim\limits_{n\to\infty} f_n(x) = 0$.
So we have $\lim\limits_{n\to\infty} f_n(x) = 0$ on $\mathbb{R}$; thus, $\int \left(\lim\limits_{n\to\infty} f_n\right) d\lambda^1 = \int 0\, d\lambda^1 = 0$.

(2) On the other hand, $\int f_n\, d\lambda^d = \int_n^\infty dx = \infty$, for all n. Thus, $\lim\limits_{n\to\infty} \int f_n\, d\lambda^d = \infty$.

The morale is that monotone convergence may not work for $f_n \geq 0$ if $f_n \uparrow$ is replaced with $f_n \downarrow$.

Solution for **(b):**

(1) If it is true that $f_n \uparrow$, then the conditions for monotone convergence are met.
If it is true that $\int_B h d\lambda^d < \infty$, then the conditions for dominated convergence are met.

‘ ‘ Neither assertion can be true: We will show that ‘ $\int \left(\lim\limits_{n\to\infty} f_n\right) d\lambda^d \neq$ ‘ $\lim\limits_{n\to\infty} \int f_n\, d\lambda^d$. ‘

(2) $e^{-x/n} \leq 1$ on $B \;\Rightarrow\; \lim\limits_{n\to\infty} (1/n) e^{-x/n} = 0$ on $B \;\Rightarrow\; \lim\limits_{n\to\infty} f_n(x) = \mathbf{1}_B e^{-x}$.
thus, $\int \left(\lim\limits_{n\to\infty} f_n\right) d\lambda^1 = \int_0^\infty e^{-x}\, dx = 1$.

(3) Moreover, $n \in \mathbb{N} \;\Rightarrow\; \int f_n\, d\lambda^d = \int_0^\infty e^{-x}\, dx + \frac{1}{n}\int_0^\infty e^{-x/n}\, dx = 1 + 1 = 2$.

(4) We obtain from **(2)** and **(3)** that $\int \left(\lim\limits_{n\to\infty} f_n\right) d\lambda^d = 1 \neq 2 = \lim\limits_{n\to\infty} \int f_n\, d\lambda^d$.
It follows that neither of the two assertions made in **(1)** can be true. □

Theorem 4.7.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real–valued, Borel–measurable function on $\mathbb{R}^d$. If f is nonnegative or Lebesgue integrable (i.e., $\int |f| d\lambda^d < \infty$), then the set function

$$\Psi : \mathfrak{B}^d \longrightarrow [0, \infty], \qquad \Psi(A) := \int_A f\, d\lambda^d \tag{4.21}$$

is σ–additive.

PROOF: ⋆

Will not be given here. We just mention that the proof for nonnegative f is based on the monotone convergence theorem and that for integrable f is based on the dominated convergence theorem. ■

Corollary 4.2.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real–valued, nonnegative, and Borel–measurable function on $\mathbb{R}^d$.
If $\int f d\lambda^d = 1$, *then the set function*

$$P : \mathfrak{B}^d \longrightarrow [0,\infty], \qquad P(A) := \int_A f\, d\lambda^d \tag{4.22}$$

defines a probability measure on $\mathbb{R}^d$.

PROOF: Clearly,

$$P(\emptyset) = \int_\emptyset f\, d\lambda^d = \int 0\, d\lambda^d = 0\,.$$

By assumption, $\int f\, d\lambda^d = 1$. Finally, the σ–additivity of P follows from Theorem 4.7 ■

Definition 4.8 (Support of a real–valued function)**.** ⋆

Let Ω be some nonempty set and $f : \Omega \to [-\infty, \infty]$. We call

$$\operatorname{suppt}(f) := \{\, \omega \in \Omega : f(\omega) \neq 0 \,\} \tag{4.23}$$

the **support** of the function f. □

Remark 4.13. ⋆ Since it is true for any function $\varphi : \mathbb{R}^d \to [-\infty, \infty]$ and $A \subseteq \mathbb{R}^d$ that

$$\iint\!\cdots\!\int_A \varphi(\vec{x}) d\vec{x} = \iint\!\cdots\!\int_{A \cap \{\vec{x}\,:\,\varphi(\vec{x}) \neq 0\}} \varphi(\vec{x}) d\vec{x}\,,$$

we see by defining $\varphi(\vec{x}) := f(\vec{x})g(\vec{x})$ for two arbitrary functions $f, g : \mathbb{R}^d \to \mathbb{R}$, that

$$\iint\!\cdots\!\int_A f(\vec{x})g(\vec{x}) d\vec{x} = \iint\!\cdots\!\int_{A \cap \operatorname{suppt}(f)} f(\vec{x})g(\vec{x}) d\vec{x}\,.$$

This can be helpful since it means that g only needs to be "well" behaved on the support of f. □

Remark 4.14. ⋆ At this point we see the following when comparing the Lebesgue integral to the Riemann integral: [53]

- Both first assigned to functions $\varphi = \sum_{j=1}^n c_n \mathbf{1}_{B_n}$ the integral $\sum_{j=1}^n c_j \lambda^d(B_j)$:
  - ⊡ For Riemann integrals: step functions φ with d dimensional rectangles B_j.
  - ⊡ For Lebesgue integrals: simple functions φ (more general) with Borel sets B_j.
- For both, the integral for general functions was obtained by taking limits.
- For both, the integral $\int_B f \cdots$ over a subset B was obtained by integrating $\mathbf{1}_B f$ over $\mathbb{R}^d$.

[53] Concerning the Riemann integral, see Introduction 4.1 on p.83.

- Both satisfy positivity, monotonicity, linearity.
- Both satisfy Fubini's theorem (iterated integrals)
- The theorems for monotone and dominated convergence are the reason that the Lebesgue integral satisfies $\lim_n \int_B f_n d\lambda^d = \int_B (\lim_n f_n) d\lambda^d$ under extremely general conditions. $\square$

5 The Probability Model

5.1 Probability Spaces

Introduction 5.1. In Section 1.2 (A First Look at Probability) we had arrived at Definition 1.2 (Probability measure - Preliminary Definition, version II; see p.13) of a probability measure P: A function which assigns to events A (subsets of the probability space Ω) a probability $P(A)$ that satisfies

- $0 \le P(A) \le 1$ • $P(\emptyset) = 0$
- σ–**additivity**: For any finite or infinite sequence of disjoint events $\big(A_n\big)_{n\in\mathbb{N}}$,

$$P\Big(\biguplus_{j=1}^{\infty} A_j\Big) \;=\; \sum_{j=1}^{\infty} P(A_j)\,. \tag{5.1}$$

In this chapter we will provide a solid mathematical foundation of the issues that were discussed in Section 1.2 (A First Look at Probability). □

There is a catch to making σ–additivity a condition for probability measures. We had stated this in a footnote of Remark 1.4 on p.14. The next example elaborates on why σ–additivity might have to come with a trade–off.

Example 5.1. A point located somewhere at $]-\infty, 0[$ starts moving to the right at a constant velocity and is stopped at random somewhere in the unit interval $[0, 1]$ in the following sense: It is stopped just as likely in the left half, $[0, \frac{1}{2}]$, as in the right half, $[\frac{1}{2}, 1]$. More generally, for any $n \in \mathbb{N}$, it is stopped equally likely in each one of the intervals $[\frac{k-1}{n}, \frac{k}{n}]$ $(k = 1, 2, \ldots, n)$.

- It should be obvious that the only reasonable probability measure on $\Omega := [0, 1]$ is the Lebesgue measure λ^1 (considered only on subsets of the unit interval): [54]

$$P : [0,1] \to [0,1]\,; \qquad [\alpha,\beta] \;\mapsto\; P\big([\alpha,\beta]\big) := \lambda^1\big([\alpha,\beta]\big) \;=\; \beta - \alpha\,, \quad \text{where } 0 \le \alpha \le \beta \le 1, \tag{5.2}$$

since it is the only one that assigns probabilities proportionate to interval length (including $P\big([\alpha,\alpha]\big) = 0$ for intervals of length zero) and also satisfies $P(\Omega) = 1$.
Note that $P\big([\alpha,\beta]\big) = \lambda^1\big([\alpha,\beta]\big)$ implies the following: The probability measure P is Lebesgue measure (considered only on subsets of the unit interval).

- Unfortunately, it has been proven that no σ–additive function that satisfies those properties exists on the entire power set of $[0, 1]$. [55]

> The only way out of this dilemma without sacrificing σ–additivity is to relax the condition that $P(A)$ must exist for ALL $A \subseteq \Omega$ and define P only on a subset of 2^Ω. □

Remark 5.1. Example 5.1 above suggests that the definition of a probability measure $A \mapsto P(A)$ should be adjusted as follows: It must be a function

$$P : \mathfrak{F} \longrightarrow [0,1]\,, \qquad \text{where } \mathfrak{F} \text{ is a suitable subset of } 2^\Omega,$$

[54] see Definition 4.1 on p.85
[55] Since $P = \lambda^1$, this corresponds to not all subsets of $\mathbb{R}$ being Borel sets. See Remark 4.2**(d)** on p.86.

such that

$$(5.3) \qquad \bullet\ P(\emptyset) = 0 \quad \bullet\ P(\Omega) = 1 \quad \bullet\ P\left(\biguplus_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \quad \text{for disjoint } A_1, A_2, \cdots \in \mathfrak{F}.$$

Those probabilities only exist lf the underlying events belong to $\mathfrak{F}$. Accordingly, $\mathfrak{F}$ should satisfy

$$(5.4) \qquad \bullet\ \emptyset \in \mathfrak{F} \quad \bullet\ \Omega \in \mathfrak{F} \quad \bullet\ \biguplus_{k=1}^{\infty} A_k \in \mathfrak{F}, \ \text{ for all sequences of disjoint } A_1, A_2, \cdots \in \mathfrak{F}.$$

In addition, we would like to be able to assign a probability to the following events:

$$\begin{aligned} A_1 \cup A_2 \cup \cdots &= \text{the event that at least one of } A_1 \text{ or } A_2 \text{ or } \ldots \text{ happens,} \\ A_1 \cap A_2 \cap \cdots &= \text{the event that each one of } A_1 \text{ and } A_2 \text{ and } \ldots \text{ happens,} \\ A^{\complement} = \Omega \setminus A &= \text{the event that } A \text{ does not happen.} \end{aligned}$$

To have a probability $P(A)$, a subset A of Ω must belong to the domain of P. Thus, $\mathfrak{F}$ should satisfy

$$(5.5) \qquad A_1, A_2, \cdots \in \mathfrak{F} \ \Rightarrow A_1 \cup A_2 \cup \cdots \in \mathfrak{F},$$

$$(5.6) \qquad A_1, A_2, \cdots \in \mathfrak{F} \ \Rightarrow A_1 \cap A_2 \cap \cdots \in \mathfrak{F},$$

$$(5.7) \qquad A \in \mathfrak{F} \ \Rightarrow A^{\complement} \in \mathfrak{F}.$$

We have found an answer to the question what properties $\mathfrak{F}$ should have. It should satisfy (5.4), (5.5), (5.6) (5.7). We can remove some redundancies from this set of conditions as follows.

(A) We can remove (5.6) for the following reason:

Let $A_1, A_2, \cdots \in \mathfrak{F}$. It follows from (5.5) and (5.7) that $(A_1 \cup A_2 \cup \cdots)^{\complement} \in \mathfrak{F}$. It follows from De Morgan's laws (Theorem 2.1 on p.40), that $A_1 \cap A_2 \cap \cdots \in \mathfrak{F}$. We have obtained (5.6).

(B) Disjoint unions are unions. Thus, by (5.5), $A_1, A_2, \cdots \in \mathfrak{F} \ \Rightarrow \biguplus_j A_j = \bigcup_j A_j \in \mathfrak{F}$. Also, $\emptyset \in \mathfrak{F}$ implies with (5.7) that $\Omega \in \mathfrak{F}$. Hence, all we need to keep from (5.4) is $\emptyset \in \mathfrak{F}$.

To sum it up, the domain $\mathfrak{F}$ of a probability measure P should satisfy $\emptyset \in \mathfrak{F}$, (5.5), and (5.7). □

All the above leads to the definition of σ–algebras as suitable domains for probability measures.

Definition 5.1 (σ–algebra)**.** Let Ω be a nonempty set and $\mathfrak{F} \subseteq 2^{\Omega}$ [56] such that

(a) $A \in \mathfrak{F} \ \Rightarrow \ A^{\complement} \in \mathfrak{F}$.

(b) $A_n \in \mathfrak{F}$ arbitrary $\Rightarrow \ \bigcup_{j=1}^{\infty} A_j \in \mathfrak{F}$.

(c) $\emptyset \in \mathfrak{F}$.

Then we call $\mathfrak{F}$ a **σ–algebra** for Ω. (Also, a σ–algebra on Ω or asociated with Ω.)

$\mathfrak{F}$ is also called a **σ–field** for Ω, but that is considered old–fashioned terminology. □

[56] thus, $\mathfrak{F}$ is a collection of sets: $A \in \mathfrak{F} \ \Rightarrow \ A \subseteq \Omega$ (!)

Proposition 5.1.

σ–algebras $\mathfrak{F}$ satisfy the following.

(a) $\Omega \in \mathfrak{F}$.

(b) *Let $n \in \mathbb{N}$ and $A_1, \ldots, A_n \in \mathfrak{F}$. Then $A_1 \cup A_2 \cup \cdots \cup A_n \in \mathfrak{F}$. (finite union.)*

(c) *Let $n \in \mathbb{N}$ and $A_1, A_2, \cdots \in \mathfrak{F}$. Let $A = \bigcap_{k=1}^{n} A_k$ and $B = \bigcap_{k=1}^{\infty} A_k$.*
Then $A \in \mathfrak{F}$ and $B \in \mathfrak{F}$. □

PROOF: ⋆

PROOF of **(a)**: True, since $\Omega = \emptyset^{\complement}$ and complements of elements of $\mathfrak{F}$ belong to $\mathfrak{F}$ and $\emptyset \in \mathfrak{F}$.

PROOF of **(b)**: Since any finite list $A_1, \ldots, A_n$ can be written as an infinite sequence

$$B_1 = A_1,\ B_2 = A_2,\ \cdots, B_n = A_n,\ B_{n+1} = B_{n+2} = \cdot = \emptyset$$

and since $B_j \in \mathfrak{F}$ for each $j \in \mathbb{N}$, it follows from Def.5.1**(b)** that $\bigcup_{j=1}^{\infty} B_j \in \mathfrak{F}$. Since

$$\bigcup_{j=1}^{n} A_j \ = \ \bigcup_{j=1}^{n} A_j \cup \emptyset \cup \emptyset \cup \cdot \cup \emptyset \ = \ \bigcup_{j=1}^{\infty} B_j\,,$$

it follows that $\bigcup_{j=1}^{n} A_j \in \mathfrak{F}$. This proves **(b)**.

PROOF of **(c)**: According to De Morgan's laws, any countable intersection can be written as the (countable) union of its complements. Thus we automatically get from **(A)** and **(B)** that countable intersections of a sequence in $\mathfrak{F}$ belong to $\mathfrak{F}$.

Here is a detailed argument. For each j let $C_j := A_j^{\complement}$. Further, let $C := \bigcup_{j=1}^{n} C_j$ and $D := \bigcup_{j=1}^{\infty} C_j$.

Since each each C_j is the complement of a member of $\mathfrak{F}$, we have $C_j \in \mathfrak{F}$. Thus, $D \in \mathfrak{F}$ by the definition of $\mathfrak{F}$, and we have seen in part **(b)** of this proposition that $C \in \mathfrak{F}$

It follows from De Morgan's laws that $C^{\complement} = A$ and $D^{\complement} = B$.

Thus, both A, B belong to $\mathfrak{F}$ as complements of elements of $\mathfrak{F}$. We have shown **(c)**. ■

Example 5.2. Let $\Omega := \{a, b, c, d, e, f\}$. Let $A_1 := \{a, b\}$, $A_2 := \{c, d\}$, $A_3 := \{e, f\}$. Then

$$\mathfrak{F} \ := \ \{\,\text{all unions involving } A_1, A_2, A_3\,\}$$

is a σ–algebra.

To see that this is true, note the following.

(a) For convenience, let $J := \{1, 2, 3\}$ (the full set of indices j for the sets A_j)

(b) $\Omega = A_1 \cup A_2 \cup A_3 = \bigcup \big[A_j : j \in J\big] \in \mathfrak{F}$. Also, by (2.34), $\emptyset \in \mathfrak{F}$.

(c) Let $A \in \mathfrak{F}$. Then there is an index set $J_A \subseteq J$ such that $A = \bigcup \big[A_j : j \in J_A\big]$.
$J_* := J \setminus J_A \ \Rightarrow\ A^{\complement} = \Omega \setminus A = \big[\bigcup_{j\in J} A_j\big] \setminus \big[\bigcup_{j\in J_A} A_j\big] = \bigcup_{j\in J_*} A_j \in \mathfrak{F}$.

Examples are: ⊡ $A = \emptyset \Rightarrow J_A = \emptyset \Rightarrow J_* = J \Rightarrow \bigcup_{j\in J_*} A_j = \Omega \in \mathfrak{F}$. ⊡ $A = \{a, b, e, f\}$ $\Rightarrow J_A = \{1, 3\}$, $J_* = \{2\} \Rightarrow \bigcup_{j\in J_*} A_j = \bigcup_{j\in\{2\}} A_j = A_2 = A^{\complement}$, $\Rightarrow A_2 \in \mathfrak{F}$.

(d) Let $B_n \in \mathfrak{F}, n \in \mathbb{N}$. Let $B := \bigcup_{n\in\mathbb{N}} B_n$. Since $B_n \in \mathfrak{F}$, there is an index set $J_n \subseteq J$ such that $B_n = \bigcup_{j\in J_n} A_j$. Let $J^* := \bigcup_{n\in\mathbb{N}} J_n$. Then $J^* \subseteq J = \{1, 2, 3\}$(!!) and $B = \bigcup_{n\in\mathbb{N}} \big[\bigcup_{j\in J_n} A_j\big] = \bigcup_{j\in J^*} A_j$. Thus, B is a union of the sets A_1, A_2, A_3, thus, $B \in \mathfrak{F}$.

Examples: ⊡ $B_1 = B_3 = B_5 = \ldots, = A_2$; $B_2 = B_4 = \ldots, = A_3 \Rightarrow \bigcup_{n\in\mathbb{N}} B_n = A_2 \cup A_3 \in \mathfrak{F}$.
⊡ $B_1 = A_1 \cup A_2$; $B_2 = A_1$; $B_3 = A_1 \cup A_3$; $B_4 = A_1 \Rightarrow J^* = I \Rightarrow \bigcup_{n\in\mathbb{N}} B_n = \Omega \in \mathfrak{F}$.

It follows from **(d)**, **(d)**, and **(d)**, that $\mathfrak{F}$ is a σ–algebra. □

Example 5.3. ⋆

Let $A_1 := \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0\}$, $A_2 := \{(x, y) \in \mathbb{R}^2 : x > 0, y < 0\}$,
$A_3 := \{(x, y) \in \mathbb{R}^2 : x < 0, y > 0\}$, $A_4 := \{(x, y) \in \mathbb{R}^2 : x < 0, y < 0\}$,
$A_5 := \{(x, y) \in \mathbb{R}^2 : x = 0 \text{ or } y = 0\}$. Then

$$\mathfrak{F} := \{\text{ all finite unions involving } A_1, \ldots, A_5 \}$$

is a σ–algebra.

Note how similar this example is to Example 5.2.

- Here, $A_1, \ldots, A_5$ is a partition of $\mathbb{R}^2$. There, $A_1, \ldots, A_3$ is a partition of Ω.
- "all finite unions involving $A_1, \ldots, A_5$" means the same as "all unions involving $A_1, \ldots, A_5$", so both examples have matching definitions of $\mathfrak{F}$.
- Here, the full set of indices j for the sets A_j is $J := \{1, \ldots, 5\}$.
- (In Example 5.2, $J = \{1, \ldots, 3\}$.)
- We replace the set $J = \{1, \ldots, 3\}$ from Example 5.2 with $J := \{1, \ldots, 5\}$.

With those adjustments, the proof that $\mathfrak{F}$ is a σ–algebra is that of Example 5.2. □

Example 5.4. ⋆

For $n \in \mathbb{Z}$, Let $A_n :=]n - 1, n] = \{x \in \mathbb{R} : n - 1 < x \leq n\}$. Then

$$\mathfrak{F} := \{\text{ all countable unions involving } A_n,\ n \in \mathbb{Z} \}$$

is a σ–algebra for $\mathbb{R}$.

Again, note the similarity of this example to Example 5.2.

- Here, $(A_n)_{n\in\mathbb{Z}}$ is a (countable) partition of $\mathbb{R}$. There, $A_1, \ldots, A_3$ is a partition of Ω.
- "all countable unions involving A_n, $n \in \mathbb{Z}$" equals "all unions involving A_n, $n \in \mathbb{Z}$", since there only are countably many A_j. Thus, both examples have matching definitions of $\mathfrak{F}$.
- Here, the full set of indices j for the sets A_j is $J := \mathbb{Z}$. (In Example 5.2, $J = \{1, \ldots, 3\}$.)
- We replace the set $J = \{1, \ldots, 3\}$ from Example 5.2 with $J := \mathbb{Z}$.

We illustrate this by computing the complement of $A := \bigcup \big[A_{3n^2-18n} : n \in \mathbb{N}\big]$.

- Let $J_A := \{3n^2 - 18n : n \in \mathbb{Z}\}$, $J_* := \mathbb{Z} \setminus J_A$. Then $A = \bigcup_{j\in J_A} A_j$ and $A^{\complement} = \bigcup_{j\in J_*} A_j$. Thus, $A^{\complement} \in \mathfrak{F}$. $\square$

Now, the general case.

Proposition 5.2. ⋆

Assume that $(A_j)_{j\in J}$ is a countable partition of a nonempty set Ω. In other words, the sets A_j are mutually disjoint subsets of Ω, $\biguplus \big[A_j : j \in J\big] = \Omega$, and the index set J is countable. Then

(5.8) $\qquad \mathfrak{F} := \{\, \textit{all unions involving some or all of the } A_j \,\}$

is a σ–algebra for Ω.

PROOF:

(a) By definition of $\mathfrak{F}$, for each $A \subseteq \mathfrak{F}$ there is an index set $J_A \subseteq J$ such that $A = \bigcup_{j\in J_A} A_j$. Since $J_A \subseteq J$, J_A is countable. Thus, $\mathfrak{F} = \{$ all countable unions involving A_j, $j \in J \}$

(b) $\Omega = \bigcup_{j\in J} A_j$ is a countable union of elements of $\mathfrak{F}$. Thus, $\Omega \in \mathfrak{F}$.

(c) By convention (2.34), $\emptyset = \bigcup_{j\in\emptyset} A_j$ Thus, $\emptyset \in \mathfrak{F}$.

(d) Let $A \in \mathfrak{F}$. Then there is an index set $J_A \subseteq J$ such that $A = \bigcup_{j\in J_A} A_j$. Let $J_* := J \setminus J_A$. Since $J = J_A \uplus J_*$, $\Omega = \biguplus_{j\in J} A_j = \big[\biguplus_{j\in J_A} A_j\big] \uplus \big[\biguplus_{j\in J_*} A_j\big] = A \uplus \big[\biguplus_{j\in J_*} A_j\big]$. Thus, $A^{\complement}$ is a union of elements of $\mathfrak{F}$. Thus, $A^{\complement} \in \mathfrak{F}$.

(e) For $n \in \mathbb{N}$, let $B_n \in \mathfrak{F}$. Let $B := \bigcup_{n\in\mathbb{N}} B_n$. Since $B_n \in \mathfrak{F}$, there is $J_n \subseteq J$ s.t. $B_n = \bigcup_{j\in J_n} A_j$. Let $J^* := \bigcup_{n\in\mathbb{N}} J_n$. Then $J^* \subseteq J$ and $B = \bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} \big[\bigcup_{j\in J_n} A_j\big] = \bigcup_{j\in J^*} A_j$. Thus, B is a union of sets $A_j \in \mathfrak{F}$, thus, $B \in \mathfrak{F}$.

It follows from **(b)** – **(e)** that $\mathfrak{F}$ is a σ–algebra. ■

Part **(b)** of the next example provides a counterexample!

Example 5.5. ⋆

Assume that $(A_j)_{j\in J}$ is an uncountable partition of Ω such that $A_j \neq \emptyset$ for all j. (Thus, not only the index set J, but also Ω itself is uncountable.) Then

(a) $\mathfrak{F} := \{$ all unions involving A_j, $j \in J \}$ is a σ–algebra,

(b) $\mathcal{E} := \{$ all countable unions involvingA_j, $j \in J \}$ is not a σ–algebra.

Showing that **(a)** is not much different from, e.g., Example 5.2 on p.103 or the proof of Proposition 5.2 and left as an exercise.

Now we show **(b)**. By Fact 2.1**(c)** on p.37, countable unions of countable sets are countable.

Let $E \in \mathcal{E}$. By definition of $\mathcal{E}$, there is some countable $J_E \subseteq J$ such that $E = \biguplus_{j\in J_E} A_j$. Since J_E is countable and J is uncountable, $J_E \subsetneq J$. Thus, $J_* := J \setminus J_E \neq \emptyset$. Since none of the A_j are empty,

$E_* := \biguplus_{j\in J_*} A_j \neq \emptyset$. From $J = J_E \uplus J_*$ we obtain $\Omega = \biguplus_{j\in J} A_j = \Big[\biguplus_{j\in J_E} A_j\Big] \uplus \Big[\biguplus_{j\in J_E} A_j\Big] = E \uplus E_*$. We have seen that $E_* \neq \emptyset$. Thus, $E \neq \Omega$. All this has been obtained for an arbitrary $E \in \mathcal{E}$. Thus, $\Omega \notin \mathcal{E}$. We conclude that $\mathcal{E}$ is not a σ–algebra. □

Definition 5.2 (Probability measures and probability spaces)**.**

Given are a nonempty set Ω with a σ–algebra $\mathfrak{F} \subseteq 2^\Omega$ and a function

$$P : \mathfrak{F} \longrightarrow [0,1]\,; \quad A \mapsto P(A) \qquad \text{as follows.}$$

(5.9) $\quad P(\emptyset) = 0\,,$ (5.10) $\quad P(\Omega) = 1\,,$

(5.11) $\quad (A_n)_{n\in\mathbb{N}} \in \mathfrak{F}$ disjoint $\Rightarrow P\Big(\biguplus_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n\in\mathbb{N}} P(A_n)$. ($\sigma$–additivity)

- We call P a **probability measure** or simply a **probability**
- The triplet $(\Omega, \mathfrak{F}, P)$ is called a **probability space**.
- (Only) the elements of $\mathfrak{F}$ are called **events**.
- We often call disjoint events **mutually exclusive** events.
- An event A is a P **Null event**, also, **Null event**, if $P(A) = 0$.

We suggest to reserve the term "probability" for the function value $P(A)$ that belongs to a specific event A, and always refer to P, i.e., the function $A \mapsto P(A)$, as a "probability measure". □

Notation 5.1 (Sample spaces and sample points)**.**

- We also call a probability space a **sample space** and an outcome a **sample point**.
- We also call Ω by itself (as opposed to the triplet $(\Omega, \mathfrak{F}, P)$) a probability space or sample space. Sometimes we refer to Ω as the **carrier set** or **carrier** of $(\Omega, \mathfrak{F}, P)$.
- We like to write Ω for the carrier set, $\mathfrak{F}$ for the σ–algebra and P for the probability measure of a probability space, but different notation may be used. For example, there may be a probability space $(S, \mathscr{S}, Q)$ and outcomes s or x or $\vec{y}$ (vector notation).

Remark 5.2. We noted in Section 1.2 (A First Look at Probability), that "sample space" is the statistician's terminology for a probability space. We will mostly use the term "probability space", since we usually think of a sample as a list of items that that has been picked in some random fashion from an underlying "population". We will consider probability spaces in this lecture where it would require a huge stretch of the imagination to consider their elements as such samples. Note though that there are occasions where the term "sample space" is preferable terminology.

You, my students, may choose whatever notation you prefer.

And more good news: We have introduced σ–algebras to properly deal with the issue that was raised in Example 5.1 on p.101

It won't be long and we will on only few occasions deal with σ–algebras.
- Thus, we will usually refer to probability spaces (Ω, P) and (S, P).
- In particular, we will also revert to calling any subset of Ω an event. □

Remark 5.3. How do we interpret $P\Big(\biguplus_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n\in\mathbb{N}} P(A_n)$ (formula (5.11) for σ–additivity in the definition of a probability measure)? There are two issues.

(a) What is the meaning of $\biguplus_{n\in\mathbb{N}} A_n$ as opposed to $\biguplus_{n=1}^{\infty} A_n$?

(b) What is the meaning of $\sum_{n\in\mathbb{N}} P(A_n)$, as opposed to $\sum_{n=1}^{\infty} P(A_n)$? Does it really not matter in which order we add the terms of an inifinite series?

The answer to **(a)** is easy. Unions are defined without any reference to an order "first A_1, then A_2, then $A_3, \dots$", since the definition of $a \in \biguplus_{n\in\mathbb{N}} A_n$ is the existence of at least one index i_0 such that $a \in A_{i_0}$. No reference to an ordering is made. The only justification for the notation $\biguplus_{n=1}^{\infty} A_n$ is that it looks more familiar. By the way, what was said here about disjoint unions also applies to arbitrary unions and to intersections.

Now, to **(b)**. The series $\sum P(A_n)$ is absolutely convergent. [57] To see this, let $A := \biguplus_{n=1}^{\infty} A_n$.

Clearly, $P(A_n) \geq 0$ for all n. Moreover, by (σ–)additivity applied to $A \uplus A^{\complement} = \Omega$,

$$P\left(\biguplus_{n=1}^{\infty} A_n\right) = P(A) \leq P(A) + P\left(A^{\complement}\right) = P(\Omega) = 1 < \infty\,.$$

Since $\sum P(A_n)$ is absolutely convergent, it does indeed not matter how the terms A_n are arranged. See Theorem 3.2 on p.58. □

In Section 1.2 (A First Look at Probability), we used throws of a die to illustrate the concepts of random actions and their potential outomes. This motivated us to give a preliminary definition of a probability measure as a function. Now that we have the final definition of a probability measure, elt us study some more examples.

Example 5.6. We model k rolls of a fair die ($k \in \mathbb{N}$) as follows. Let

$$\Omega := \{1,2,3,4,5,6\}^k = \{(a_1, a_2, \dots, a_k) : a_j = 1,2,..,6, \text{ where } j = 1,2,\dots,k\}\,.$$

For example, let $k = 5$. then $\omega_1 = (2,6,2,1,4) \in \Omega$. On the other hand, $\omega_2 = (2,6,2,9,4) \notin \Omega$, since $a_j = 1,2,\dots,6$ is not true for $j = 4$ (because $a_4 = 9$).
Ω is a finite set, and you will learn later that its size is 6^k. Thus, $\Omega = \{\omega_1, \omega_2, \dots, \omega_{6^k}\}$ where, e.g.,

$$\omega_1 = (1,1,\dots,1,1),\ \omega_2 = (1,1,\dots,1,2),\ \dots,\ \omega_{6^k-1} = (6,6,\dots,6,5),\ \omega_{6^k} = (6,6,\dots,6,6).$$

[57] See Definition 3.1 (Absolute Convergence) on p.58.

Since the die is fair, each one of those 6^k elements of Ω should have the same probability $p := P(\{\omega\})$ for all $\omega \in \Omega$. Since $P(\Omega) = 1$ and

$$\Omega \;=\; \biguplus \big[\{\omega\} : \omega \in \Omega\big] \;=\; \biguplus_{j=1}^{\infty} \{\omega_j\}\,.$$

is a union of a sequence of disjoint sets, we obtain from the σ–additivity of $P(\cdot)$ the following:

$$1 \;=\; P(\Omega) \;=\; \sum_{j=1}^{6^k} P\{\omega_j\} \;=\; 6^k p \quad\Rightarrow\quad p \;=\; \frac{1}{6^k}\,.$$

- So then, how does one define a probability measure $P : \mathfrak{F} \to [0,1]$?
- And what is that σ–algebra $\mathfrak{F}$ going to be?

To answer those questions, we define the function $P : 2^\Omega \to \mathbb{R}$ as follows.

$$P(A) \;:=\; \frac{|A|}{|\Omega|} \;=\; \frac{|A|}{6^k}\,. \tag{5.12}$$

Observe the following.

(1) $A \subseteq \Omega \;\Rightarrow\; 0 \le |A| \le |\Omega| = 6^k \;\Rightarrow\; 0 \le P(A) \le 1$.

(2) The empty set has size $|\emptyset| = 0$ and Ω has size $|\Omega| = 6^k$. Thus, $P(\emptyset) = 0$ and $P(\Omega) = 1$.

(3) Assume that $A_1, A_2, \ldots$ are disjoint subsets of Ω. Since Ω is finite, only finitely many A_j are not empty. (THINK!)

(4) We rearrange that sequence such that its nonempty members will be $A_1, A_2, \ldots, A_m$, for some suitable m.

(5) Then, $A = A_1 \uplus A_2 \uplus \cdots \uplus A_m$ is a finite union. Disjointness of the A_j implies that

$$|A| \;=\; |A_1| \,+\, |A_2| \,+\cdots+\, |A_m|\,.$$

(6) By σ–additivity, $P(A) \;=\; |A|/6^k \;=\; \sum_{j=1}^{m} (|A_j|/6^k) \;=\; \sum_{j=1}^{m} P(A_j) \;=\; \sum_{\text{all } j} P(A_j)$

For the last equation, observe that the omitted sets $A_{m+1}, A_{m+2}, \ldots$ were empty; thus, $P(A_j) = 0/6^k = 0$ for those j.

We obtain from **(1) – (6)** that $P(A) = |A|/6^k$ is a probability measure on 2^Ω. $\square$

Example 5.7. One easily sees the generalization of the last example to arbitrary finite sets: Let Ω be a finite set of size $N := |\Omega| < \infty$. Let the function $P : 2^\Omega \to \mathbb{R}$ be given as

$$P(A) \;:=\; \frac{|A|}{|\Omega|} \;=\; \frac{|A|}{N}\,. \tag{5.13}$$

Then everything stated in **(1) – (6)** of **(a)** remains valid if we replace 6^k with N. This shows that P is a probability measure on 2^Ω. $\square$

Definition 5.3 (Equiprobability).

Let (Ω, P) be a finite probability space, i.e., $|\Omega| < \infty$. Let $n := |\Omega|$. We say that P has **equiprobable** outcomes and also, that P **satisfies equiprobability**, if

$$P\big(\{\omega\}\big) = \frac{1}{|\Omega|} \quad \text{(since then } P\{\omega\} \text{ is constant for all } \omega \in \Omega). \tag{5.14}$$ □

Remark 5.4. The finiteness of Ω was crucial in the last two examples for the following reason.

If Ω is infinite and countable, then $\Omega = \{\omega_1, \omega_2, \dots\}$ can be written as an infinite sequence of distinct(!) members. It is not possible to define a "uniform" probability measure on Ω as we did in parts **(a)** and **(b)**, i.e., a number p such that $P(\omega_j) = p$, for all $j \in \mathbb{N}$. How so?

(1) p would have to be strictly positive: Otherwise, $P(\Omega) = \sum_j P(\omega_j) = p + p + \dots \leq 0$, but we require $P(\Omega) = 1$.

(2) Thus, $p > 0$. Thus, $P(\Omega) = \sum_j P(\omega_j) = p + p + \dots = \infty$. However, we require $P(\Omega) = 1$. □

Remark 5.5. We will see that the most important probability measures on the uncountable set $\mathbb{R}$ satisfy $P(x) = 0$ for all $x \in \mathbb{R}$. [58] That is no contradiction to σ–additivity and $P(\mathbb{R}) = 1$, since one cannot write the real numbers as a countable union $\mathbb{R} = \{x_1\} \uplus \{x_1\} \uplus \{x_2\} \uplus \cdots$. Obviously, it is no more possible in those cases to determine a probability measure on $\mathbb{R}$ by only listing the probabilities $P(x)$ of the atomar events $\{x\}$ for all $x \in \mathbb{R}$. Rather, P often is characterized by integrals $P([a,b]) = \int_a^b \varphi(t)dt$. (And if this is the case, we obtain indeed $P(x) = \int_x^x \varphi(t)dt = 0$ for all x.) □

Recall for the next theorem that we denote by $A_n \uparrow$ a nondecreasing sequence of events: $i < J \Rightarrow A_i \subseteq A_j$ and by $B_n \downarrow$ a nondecreasing sequence of events: $i < J \Rightarrow B_i \supseteq B_j$. (See Definition 2.23 on p.36.)

Theorem 5.1 (Continuity property of probability measures). ⋆

Let $(\Omega, \mathfrak{F}, P)$ be a probability space. If $A_n, B_n \in \mathfrak{F}$, then the following is true:

$$A_n \uparrow \Rightarrow P(A_n) \uparrow P\left(\bigcup_{n\in\mathbb{N}} A_n\right), \tag{5.15}$$

$$B_n \downarrow \Rightarrow P(B_n) \downarrow P\left(\bigcap_{n\in\mathbb{N}} B_n\right). \tag{5.16}$$

PROOF: We prove (5.15) as follows: Let $A := \bigcup_{j=1}^{\infty} A_j$ and

$$C_1 := A_1, \quad C_{n+1} := A_{n+1} \setminus A_n \ (n \in \mathbb{N}).$$

[58] Those probability measures are the so-called distributions of continuous random variables.

Note that $A_n \uparrow \Rightarrow A_n = \bigcup_{j=1}^{n} A_j$ and thus, $C_{n+1} := A_{n+1} \setminus \left(\bigcup_{j=1}^{n} A_j \right)$.
According to Proposition 2.6 (Rewrite unions as disjoint unions) on p.41, the sets C_j form a partition of A and we have

$$A_n = \biguplus_{j=1}^{n} C_j\,, \qquad A = \biguplus_{j=1}^{\infty} C_j\,,$$

It follows from the σ–additivity of P that

$$P(A) = P\left(\biguplus_{j=1}^{\infty} C_j \right) = \sum_{j=1}^{\infty} P(C_j) = \lim_{n\to\infty} \sum_{j=1}^{n} P(C_j) = \lim_{n\to\infty} P\left(\biguplus_{j=1}^{n} C_j \right) = \lim_{n\to\infty} P(A_n)\,.$$

This proves (5.15). We use this result to prove (5.16) as follows.
Let $B := \bigcap_{j=1}^{\infty} B_j$. For $n \in \mathbb{N}$, let $A_n := B_n^{\complement}$. Further, let $A := \bigcup_{j=1}^{\infty} A_j$. Then $A_n \uparrow$ and it follows from De Morgan that

$$A^{\complement} = \left(\bigcup_{j=1}^{\infty} A_j \right)^{\complement} = \bigcap_{j=1}^{\infty} A_j^{\complement} = \bigcap_{j=1}^{\infty} B_j = B\,.$$

We apply (5.15) and obtain

$$1 - P(B_n) = P(A_n) \uparrow P\left(\bigcup_{n\in\mathbb{N}} A_n \right) = 1 - P\left[\left(\bigcup_{n\in\mathbb{N}} A_n \right)^{\complement} \right] = 1 - P(B)\,.$$

Thus, $P(B_n) \downarrow P(B)$ and this proves (5.16). ■

Definition 5.4 (Discrete probability space)**.**

Assume that the probability space $(\Omega, \mathfrak{F}, P)$ satisfies the following:
(a) $P(\{\omega\})$ is defined for all $\omega \in \Omega$. In other words, we ask that $\{\omega\} \in \mathfrak{F}$ for all $\omega \in \Omega$.
(b) There exists a countable subset A^* of Ω such that $\sum_{\omega\in A^*} P\{\omega\} = 1$

Then we call $(\Omega, \mathfrak{F}, P)$ a **discrete probability space**. □

We will later on talk about discrete and continuous random variables, but note that there is no such thing as a "continuous probability space".

Remark 5.6. For the interpretation of the summation $\sum_{\omega\in A^*} P\{\omega\}$ we note the following.

(a) Either A^* is finite and can be written $A^* = \{\omega_1, \omega_2, \ldots, \omega_n\}$ for some suitable n.
Then $\sum_{\omega\in A^*} P\{\omega\} = \sum_{j=1}^{n} P\{\omega_j\}$.

(b) Or A^* is infinite and can be written $A^* = \{\omega_j : j \in \mathbb{N}\}$. We reason as in Remark 5.3 on p.107 with $\{\omega_j\}$ in place of A_j and see that the series $\sum P\{\omega_j\}$ is absolutely convergent. Thus, the value of $\sum_{j=1}^{n} P\{\omega_j\}$ does not depend on how the elements of A^* were sequenced and we can write $\sum_{\omega\in A^*} P\{\omega\}$ for that common value. □

In the next theorem we intentionally deviate from the standard notation $(\Omega, \mathfrak{F}, P)$ for a probability space, because it is typically applied to the codomain (rather than domain) of a random element.

Theorem 5.2.

Let $(\Omega', \mathfrak{F}', P')$ be a discrete probability space and $A^* \in \mathfrak{F}'$ countable such that $\sum_{\omega'\in A^*} P'\{\omega'\} = 1$. Then,

- ***(a)*** $A^* \in \mathfrak{F}'$.
- ***(b)*** $P'(A^*) = 1$ *and thus,* $P'\left((A^*)^{\complement}\right) = 0$.
- ***(c)*** $P'(A) = P'(A \cap A^*)$ *for all* $A \in \mathfrak{F}'$.
- ***(d)*** $P'(A) = \sum_{\omega'\in A\cap A^*} P'\{\omega'\}$ *for all* $A \in \mathfrak{F}'$.
- ***(e)*** ⋆ *The formula* $\widetilde{P}(B) := P'(B \cap A^*)$ *"extends"* P' *to a probability measure* $\widetilde{P}$ *on the entire power set* $2^{\Omega'}$.

PROOF: ⋆

PROOF of **(a)**: This is true, because $\{\omega'\} \in \mathfrak{F}'$ for all ω' and $A^* = \biguplus_{\omega'\in A^*} \{\omega'\}$ is a countable union of elements of $\mathfrak{F}'$.

PROOF of **(b)**: By definition, $\sum_{\omega'\in A^*} P'\{\omega'\} = 1$. Since $A^* = \biguplus_{\omega'\in A^*} \{\omega'\}$, we obtain $P'(A^*) = 1$.

Further, $\Omega' = A \uplus (A^*)^{\complement} \Rightarrow 1 = P'(A^*) + P'\left((A^*)^{\complement}\right) = 1 + P'\left((A^*)^{\complement}\right)$. Thus, $P'\left((A^*)^{\complement}\right) = 0$.

PROOF of **(c)**: From $0 \le P'\left(A \cap (A^*)^{\complement}\right) \le P'\left((A^*)^{\complement}\right) = 0$, we obtain $P'\left(A \cap (A^*)^{\complement}\right) = 0$.

From $A = [A \cap A^*] \uplus \left[A \cap (A^*)^{\complement}\right]$, we obtain $P'(A) = P'(A \cap A^*) + P'\left(A \cap (A^*)^{\complement}\right) = P'(A \cap A^*)$.

PROOF of **(d)**: $A \cap A^*$ is a subset of A^*, hence, countable. Thus, $P'(A \cap A^*) = \sum_{\omega'\in A\cap A^*} P'\{\omega'\}$. We obtain from **(c)** that $P'(A) = \sum_{\omega'\in A\cap A^*} P'\{\omega'\}$.

PROOF of **(e)**: Tedious but easy, if one uses **(c)** and distributivity $A^* \cap \biguplus_j A_j = \biguplus_j (A^* \cap A_j)$. ■

Corollary 5.1.

(a) *If $(\Omega', \mathfrak{F}', P')$ be a discrete probability space, then P' is characterized by the probabilities $P'\{\omega'\}$ of the outcomes ω'.*

(b) *Let Ω' be some arbitrary, nonempty set. Assume that $\big(p_j\big)_j$ is a finite or infinite sequence of real numbers that satisfies*

- *$p_j \geq 0$ for all j and $\sum_j p_j = 1$*

Further, assume that $\big(\omega'_j\big)_j$ is a corresponding sequence of distinct elements of Ω', then $\big(p_j\big)_j$ defines a discrete probability space $(\Omega', 2^{\Omega'}, P')$ as follows.

- $P'(\emptyset) := 0, \quad P'(A) := \sum\limits_{j\,:\,\omega'_j \in A} p_j$, *for $A \neq \emptyset$.* $\square$

PROOF: ⋆ This follows from Theorem 5.2. The details are left to the reader. ■

Remark: We mentioned in Remark 1.7 on p.17 the following for a random element $X : (\Omega, \mathfrak{F}, P) \to (\Omega', \mathfrak{F}')$: The formula

$$P_X(A') \;=\; P\{X \in A'\}, \quad (A' \in \mathfrak{F}')$$

defines a probability measure on the subsets of Ω' (on $\mathfrak{F}'$, to be precise), which we referred to as the distribution of X under P. [59]

Assume that Ω^* has been sequenced as $\Omega^* = \{\omega'_1, \omega'_2, \ldots$. Let

$$p_j \;:=\; P_X\{\omega'_j\} \;=\; P\{X = \omega'_j\}$$

Since

$$\sum_{j\in\mathbb{N}} p_j \;=\; \sum_{j\in\mathbb{N}} P_X\{\omega'_j\} \;=\; P_X(\omega^*) \;=\; 1,$$

The assignment [60] $\omega' \;\mapsto\; p_X(\omega') \;=\; \begin{cases} p_j, & \text{if } \omega' = \omega'_j\,, \\ 0, & \text{otherwise,} \end{cases}$

uniquely determines the distribution P_X $\square$

Remark 5.7. The probability spaces $(\Omega, \mathfrak{F}, P)$ we will be faced with when doing computations for practical applications belong to one of the following categories:

(a) $(\Omega, \mathfrak{F}, P)$ is a discrete probability space. According to Theorem 5.2**(e)** on p.111, we may choose $\mathfrak{F} = 2^{\Omega}$).

(b) $\Omega = \mathbb{R}$ and $P(A)$ is known (at a minimum) for intervals such as $[a, b]$ or $]a, b]$ or $[a, b[$ or $]a, b[$.

(c) $\Omega = \mathbb{R}^n$ and $P(A)$ is known (at a minimum) for n–dimensional rectangles such as $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$ (cartesian products of onedimensional intervals!)

It is important that we can assign probabilities to Intervals in **(c)** and n–dimensional rectangles in **(d)**, for the following reason.

[59] The precise definition of a probability distribution will be given in in Definition 5.13 (Probability Distribution) on p.124.

[60] later on referred to as the probability mass function (PMF) of X

(c′) the most important probabilities P defined for sets in $\mathbb{R}$ come with a so called **probability density function** $f : \mathbb{R} \to [0, \infty[$ which assigns to an interval $]a, b]$ the probability

$$P\big(]a, b]\big) \;=\; \int_a^b f(u)\, du\,.$$

It seems plausible that the σ–algebra $\mathfrak{B}$ for such P should contain all intervals $]a, b]$.

(d′) Likewise, the most important probabilities P defined for sets in $\mathbb{R}^n$ come with a probability density function $f : \mathbb{R}^n \to [0, \infty[$ which assigns to an n–dimensional rectangle $]a_1, b_1]\times]a_2, b_2] \times \cdots \times]a_n, b_n]$ the probability

$$P\big(]a_1, b_1]\times]a_2, b_2] \times \cdots \times]a_n, b_n]\big) \;=\; \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_1}^{b_1} f(\vec{u})\, d\vec{u}$$
$$= \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_1}^{b_1} f(u_1, \ldots, u_n)\, du_1\, du_2 \cdots du_{n-1}\, du_n\,.$$

Accordingly, it is desirable that the σ–algebra $\mathfrak{B}^n$ for such P contains all rectangles

$$]a_1, b_1] \times]a_2, b_2\,] \;\times\; \cdots \;\times\;]a_n, b_n\,]\,.$$

You may have Noticed that we could have worked with either of $]a_j, b_j[$, $[a_j, b_j[$, $[a_j, b_j]$ instead of $]a_j, b_j]$, since $\int_a^a \ldots da$ is always zero. However, it is more convenient to work with intervals that are open on the left and closed on the right. We will see that when we deal with the so-called cumulative distribution functions on $\mathbb{R}$ and $\mathbb{R}^n$. □

Theorem 5.3. ⋆ *Let Ω be some arbitrary set and $\big(\mathfrak{F}_i\big)_{i\in I}$ a family of σ–algebras on Ω, i.e., $\mathfrak{F}_i \subseteq 2^\Omega$ for each $i \in I$. No assumption is made about the index set other than $I \neq \emptyset$. Thus, this family may consist of finitely many σ–algebras or of entire sequence or even uncountably many σ–algebras.*

- *Let $\mathfrak{F} := \bigcap_{i\in I} \mathfrak{F}_i$, i.e., $\mathfrak{F} = \{A \subseteq \Omega : A \in \mathfrak{F}_i$ for each index $i\}$. Then $\mathfrak{F}$ is a σ–algebra.*

This can also be stated as follows. *Any intersection of σ–algebras results in a σ–algebra.*

PROOF: We show that **(b)** of Definition 5.1 (σ–algebra) on p.102 holds:

- $A_n \in \mathfrak{F}$ for all $n \;\Rightarrow\; \bigcup_{j=n}^{\infty} A_n \in \mathfrak{F}$.

So let $A_n \in \mathfrak{F}$ for all n and let $A := \bigcup_{n\in\mathbb{N}} A_n$. Let $i \in I$. Since $\mathfrak{F} \subseteq \mathfrak{F}_i$, $A_n \in \mathfrak{F}_i$ for all n.

Since $\mathfrak{F}_i$ is a σ–algebra, $A \in \mathfrak{F}_i$. Since this is true for an arbitrary $i \in I$, $A \in \bigcap i \in I\mathfrak{F}_i$, i.e., $A \in \mathfrak{F}$.

The proofs of $A \in \mathfrak{F} \;\Rightarrow\; A^{\complement} \in \mathfrak{F}$ and of $\emptyset \in \mathfrak{F}$ follow the same template and are left to the reader. ■

Theorem 5.4. ⋆

Let Ω be some arbitrary set and $\mathscr{A} \subseteq 2^\Omega$. In other words, each element of $\mathscr{A}$ is a subset of Ω.

- *There exists a minimal (i.e., smallest) σ–algebra that contains $\mathscr{A}$.*
- *Further, this σ–algebra is uniquely determined by $\mathscr{A}$. This allows us to name it $\sigma\{\mathscr{A}\}$.*

PROOF: We obtain $\sigma\{\mathscr{A}\}$ as the intersection of all σ–algebras that contain $\mathscr{A}$. According to Theorem 5.3, this intersection is a σ–algebra. ■

Definition 5.5 (σ–algebra generated by a collection of sets). ⋆ Let Ω be some nonempty set.

(a) Let $\mathscr{A} \subseteq 2^\Omega$, i.e., the elements of $\mathscr{A}$ are subsets of Ω.
We call $\sigma\{\mathscr{A}\}$ the **σ–algebra generated by $\mathscr{A}$**. If $\mathscr{A}$ is of the form $\mathscr{A} = \{...\}$, we also write $\sigma\{...\}$ for $\sigma\{\{...\}\}$.

(b) Assume in addition that $\mathfrak{F}$ is a σ–algebra for Ω and $\mathscr{A} \subseteq \mathfrak{F}$. If $\sigma\{\mathscr{A}\} = \mathfrak{F}$, we call $\mathscr{A}$ a **generator for** $\mathfrak{F}$ a.k.a. generator of $\mathfrak{F}$, and we say that $\mathscr{A}$ **generates** $\mathfrak{F}$.

Concerning notation:

- One also can write $\sigma(\mathscr{A})$ or $\sigma[\mathscr{A}]$ for $\sigma\{\mathscr{A}\}$.
- Given a family of subsets $A_i \subseteq \Omega$, $(i \in I)$, $\sigma\{A_i : i \in I\}$ can also be written as $\sigma\{A_i : i \in I\} = \sigma\big((A_i)_{i\in I}\big) = \sigma\big[(A_i)_{i\in I}\big] = \sigma\big\{(A_i)_{i\in I}\big\}$.
As usual, it is OK to omit the "$i \in I$" part if the meaning of I is unambiguous. □

Example 5.8. ⋆ The simplest example possible is the computation of $\sigma\{A\}$, for some $A \subseteq \Omega$.

- Since $\emptyset$ and Ω belong to any σ–algebra on Ω, $\emptyset \in \sigma\{A\}$ and $\Omega \in \sigma\{A\}$.
- $\boldsymbol{\mathcal{E}} \subseteq \sigma\{\boldsymbol{\mathcal{E}}\}$ is true for any $\boldsymbol{\mathcal{E}} \subseteq 2^\Omega$. Thus, $\{A\} \subseteq \sigma\{A\}$, i.e., $A \in \sigma\{A\}$.
- If A belongs to a σ–algebra, so does $A^\complement$. Since $A \in \sigma\{A\}$, we also have $A^\complement \in \sigma\{A\}$.
- Thus, if $\boldsymbol{\mathcal{A}} = \{\emptyset, A, A^\complement\Omega\}$, then $\{A\} \subseteq \boldsymbol{\mathcal{A}} \subseteq \sigma\{A\}$. Since $\boldsymbol{\mathcal{A}}$ is a σ–algebra that contains $\{A\}$, and $\sigma\{A\}$ is minimal among those, we also have $\boldsymbol{\mathcal{A}} \supseteq \sigma\{A\}$. Thus, $\boldsymbol{\mathcal{A}} = \sigma\{A\}$. □

The next definition is marked optional, but note that Borel sets will be mentioned frequently during lecture. Matter of fact, we have already encountered them when we discussed Lebesgue integrals in Chapter 4 (Calculus Extensions). See Definition 4.2 (Borel sets) on p.85 and Theorem 4.1, which preceeds it. There Borel sets were introduced as some subset of $2^{\mathbb{R}^d}$ which is big enough to include all Riemann integrable sets and satisfies (4.4)–(4.6), what we now recognize as the formulas that define a σ–algebra.

Definition 5.6 (Borel σ–algebra). ⋆ For $d = 1, 2, \ldots$, we define

- $\mathfrak{B}^d := \sigma\{$ d–dimensional rectangles $\}$,
- $\mathfrak{B} := \mathfrak{B}^1 = \sigma\{$ all intervals of real numbers $\}$.

$\mathfrak{B}$ and $\mathfrak{B}^d$ are the **Borel σ–algebras** and their members are the **Borel sets** of $\mathbb{R}$ and $\mathbb{R}^d$. □

Remark 5.8. **(A)** Consider the following sets of intervals of real numbers.

$$\mathfrak{I}_1 := \{]a,b] : a < b\}, \quad \mathfrak{I}_2 := \{[a,b] : a < b\},$$
$$\mathfrak{I}_3 := \{]a,b[: a < b\}, \quad \mathfrak{I}_4 := \{[a,b[: a < b\}, \quad \boldsymbol{\mathcal{E}} := \{\,]-\infty, c] : c \in \mathbb{R}\,\}.$$

One can show that each one of those sets of intervals is big enough to generate the Borel sets of $\mathbb{R}$: $\mathfrak{B} = \sigma(\mathfrak{I}_1) = \sigma(\mathfrak{I}_2) = \sigma(\mathfrak{I}_3) = \sigma(\mathfrak{I}_4)$.

(B) The above generalizes to d–dimensional space: Let

$$\begin{aligned}
\mathfrak{I}_5 &:= \{]a_1,b_1]\times]a_2,b_2]\times\cdots\times]a_d,b_d] : a_1<b_1, a_2<b_2,\ldots,a_d<b_d\}\,,\\
\mathfrak{I}_6 &:= \{[a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_d,b_d] : a_1<b_1, a_2<b_2,\ldots,a_d<b_d\}\,,\\
\mathfrak{I}_7 &:= \{]a_1,b_1[\times]a_2,b_2[\times\cdots\times]a_d,b_d[: a_1<b_1, a_2<b_2,\ldots,a_d<b_d\}\,,\\
\mathfrak{I}_8 &:= \{[a_1,b_1[\times[a_2,b_2[\times\cdots\times[a_d,b_d[: a_1<b_1, a_2<b_2,\ldots,a_d<b_d\}\,,
\end{aligned}$$

one can show that $\mathfrak{B}^d = \sigma(\mathfrak{I}_5) = \sigma(\mathfrak{I}_6) = \sigma(\mathfrak{I}_7) = \sigma(\mathfrak{I}_8)$.

(C) In the parlance of Definition 5.5 (σ–algebra generated by a collection of sets) on p.114, each one of $\mathfrak{I}_1$–$\mathfrak{I}_4$ is a generator of the onedimensional Borel sets, and each one of $\mathfrak{I}_5$–$\mathfrak{I}_8$ is a generator of the d–dimensional Borel sets. □

Fact 5.1. ⋆ *For the following, note that the sets $\mathfrak{I}_1,\ldots,\mathfrak{I}_8$ were defined in Example 5.8 on p.114.*

(a) *Let $\mathfrak{I}$ denote one of the collections of half-open intervals, $\mathfrak{I}_1,\mathfrak{I}_4$. Let $\mathscr{E} := \mathfrak{I}\uplus\mathbb{R}$. Then any function $P_0:\mathscr{E}\to[0,1]$ which satisfies $P_0(\emptyset)=0,\ P_0(\mathbb{R})=1$ and σ–additivity on $\mathscr{E}$: $E_n\in\mathscr{E}$ disjoint such that $E := \biguplus_{n\in\mathbb{N}} \in\mathscr{E} \Rightarrow P_0(E)=\sum_{n\in\mathbb{N}} P_0(E_n)$, can be uniquely extended to a probability measure on $\mathfrak{B}$, the Borel sets of $\mathbb{R}$.*

(b) *Let $\mathfrak{I}$ denote one of the collections of d–dimensional rectangles $\mathfrak{I}_5,\mathfrak{I}_8$. Let $\mathscr{E} := \mathfrak{I}\cup\{\mathbb{R}^d\}$. Then any function $P_0:\mathscr{E}\to[0,1]$ which satisfies $P_0(\emptyset)=0,\ P_0(\mathbb{R}^d)=1$ and σ–additivity on $\mathscr{E}$: $E_n\in\mathscr{E}$ disjoint such that $E := \biguplus_{n\in\mathbb{N}} E_n\in\mathscr{E} \Rightarrow P_0(E)=\sum_{n\in\mathbb{N}} P_0(E_n)$, can be uniquely extended to a probability measure on $\mathfrak{B}^d$, the Borel sets of $\mathbb{R}^d$.* □

Remark 5.9. Consider this a continuation of Remark 5.7. We can summarize it as follows.
There are essentially only two kinds of probability spaces $(\Omega,\mathfrak{F},P)$ we are interested in.

(a) There is a countable subset A^* of Ω such that $\sum_{\omega\in A^*} P(\{\omega\}) = 1$ (discrete probability spaces).
Then $\boxed{\mathfrak{F}=2^\Omega}$, since the above allows us to define $P(A)$ for arbitrary $A\subseteq\Omega$ as

$$P(A) = \sum_{\omega\in A^*\cap A} P(\{\omega\})\,.$$

(b) $\Omega=\mathbb{R}$ or $\Omega=\mathbb{R}^n$. Then $\boxed{\mathfrak{F}=\text{the Borel sets.}}$

We note once more that all subsets of $\mathbb{R}^d$ that crop up in applications are Borel. See, e.g., Remark 4.9 on p.92. That allows us to behave as if we are in situation **(a)**, where $P(A)$ is defined for all $A\in 2^\Omega$, i.e., as if $\mathfrak{F}=2^\Omega$. But then there is no more need to worry about $\mathfrak{F}$ and we can and will henceforth, with very few exceptions, do the following.

We will ignore that probability measures cannot always be given on the entire power set 2^Ω. Accordingly, we will drop the σ–algebra $\mathfrak{F}$ from $(\Omega,\mathfrak{F},P)$.

- We often will refer to probability spaces (or sample spaces) (Ω,P),

rather than to probability spaces $(\Omega,\mathfrak{F},P)$. □

Notational conveniences for probabilities:

If we have a set that is written as $\{\dots\}$, i.e., with curly braces as delimiters, then we may write its probability as $P\{\dots\}$ instead of $P\big(\{\dots\}\big)$. Specifically for singletons $\{\omega\}$, it is OK to write $P\{\omega\}$.

The next theorem lists two important rules to determine probabilities.

Theorem 5.5 (WMS Ch.02.8, Theorem 2.6). *If A and B are two events in a probability space (Ω, P), then*

(5.17) ***Additive Law of Probability:*** $P(A \cup B) \;=\; P(A) \,+\, P(B) \,-\, P(A \cap B)\,.$

(5.18) ***Rule of the Complement:*** $P\big[A^{\complement}\big] \;=\; 1 \,-\, P[A]\,.$

PROOF of (5.17): We apply the σ–additivity of P as follows:

(1) $A \;=\; (A \setminus B) \uplus (A \cap B)$ and $B \;=\; (B \setminus A) \uplus (A \cap B)$
$\Rightarrow\; P(A) + P(B) \;=\; P(A \setminus B) + P(A \cap B) \;+\; P(B \setminus A) + P(A \cap B)$

(2) $A \cup B \;=\; (A \setminus B) \uplus (A \cap B) \uplus P(B \setminus A)$
$\Rightarrow\; P(A \cup B) \;=\; P(A \setminus B) + P(A \cap B) \;+\; P(B \setminus A)$

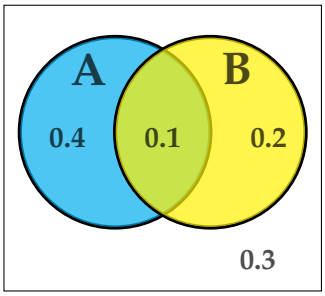
Thus, from **(1)** and **(2)**, $P(A) + P(B) \;=\; P(A \cup B) + P(A \cap B)$.
It follows that $P(A \cup B) \;=\; P(A) + P(B) - P(A \cap B)$.
PROOF of (5.18): Immediate from the σ–additivity of P and $\Omega \;=\; A \uplus A^{\complement}$. ■

Remark 5.10. If the events A and B are mutually exclusive, i.e., $A \cap B = \emptyset$, then $P[A \cap B] = 0$ and the additive law of probability simply is σ–additivity

(5.19) $$P(A \uplus B) \;=\; P(A) \;+\; P(B)\,.$$ □

Remark 5.11. The additive law of probability is very easy to apply, since all you need is $P(A)$, $P(B)$ and $P(A \cap B)$.

Nevertheless it might be fastest to draw a Venn diagram. Assume you know that $P(A) = 0.5$, $P(B) = 0.3$, $P(A \cap B) = 0.1$.
Clearly, $P(A \setminus B) = P(A) - P(A \cap B) = 0.4$
and $P(B \setminus A) = P(B) - P(A \cap B) = 0.2$.
It is now immediate that $P(A \cup B) = 0.7$ and we get for free that $P(A \cup B^{\complement}) = 0.3$.



The additive law of probability has generalizations for the probability of the union of three or more events.

Theorem 5.6 (Exclusion–Inclusion formula for 3 events). ⋆

If A_1, A_2, A_3 are events in a probability space (Ω, P), then

$$P(A_1 \cup A_2 \cup A_3) = \big[P(A_1) + P(A_2) + P(A_3)\big] - \big[P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3)\big] + P(A_1 \cap A_2 \cap A_n)\,. \tag{5.20}$$

PROOF: We apply the additive law of probability to the sets A_1 and $A_2 \cup A_3$ and obtain

$$P[A_1 \cup A_2 \cup A_3] = P[A_1] + P[A_2 \cup A_3] - P\big[A_1 \cap (A_2 \cap A_3)\big]\,. \tag{A}$$

Next, we apply the additive law of probability to A_2 and A_3:

$$P[A_2 \cup A_3] = P[A_2] + P[A_3] - P[A_2 \cap A_3]\,.$$

We substitute that in **(A)** which then reads

$$P[A_1 \cup A_2 \cup A_3] = P[A_1] + P[A_2] + P[A_3] - P[A_2 \cap A_3] - P\big[A_1 \cap (A_2 \cup A_3)\big]\,. \tag{B}$$

Since $A_1 \cap (A_2 \cup A_3) = (A_1 \cap A_2) \cup (A_1 \cap A_3)$, (see (2.37) on p.41: distributivity of unions and intersections), it follows from **(B)** that

$$P[A_1 \cup A_2 \cup A_3] = P[A_1] + P[A_2] + P[A_3] - P[A_2 \cap A_3] - P\big[(A_1 \cap (A_2) \cup (A_1 \cap A_3)\big]\,. \tag{C}$$

Finally, we apply the additive law of probability to the sets $A_1 \cap A_2$ and $A_1 \cap A_3$:

$$\begin{aligned} P[A_1 \cup A_2 \cup A_3] &= P[A_1] + P[A_2] + P[A_3] - P[A_2 \cap A_3] \\ &\quad - \big(P[A_1 \cap A_2] + P[A_1 \cap A_3] - P[A_1 \cap A_2 \cap A_1 \cap A_3]\big) \\ &= P[A_1] + P[A_2] + P[A_3] \\ &\qquad - P[A_2 \cap A_3] - P[A_1 \cap A_2] - P[A_1 \cap A_3] + P[A_1 \cap A_2 \cap A_3]\,. \quad \blacksquare \end{aligned}$$

Here is the general formula for any number of events.

Theorem 5.7 (Exclusion–Inclusion formula). ⋆

If $A_1, A_2, \cdots, A_n$ are events in a probability space (Ω, P), then

$$P(A_1 \cup A_2 \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \cdot P(A_1 \cap A_2 \cdots \cap A_n)\,. \tag{5.21}$$

PROOF: Will not be given here. ■

Remark 5.12. **This remark is preliminary.**

Randomness specifically:

(1) Random number generator of a statistics package: Generate a random a number $0 \leq x < 1$ with a precision of k decimals. We can have a big k, e.g., $k = 25$. For such a high precision we can model the set of potential outcomes Ω as the continuum $[0, 1[$. For any interval $[a, b[\subseteq [0, 1[$, we would obtain that $P([a, b[)$, the probability that a number that falls into $[a, b[$ is generated, equals $b - a$. In other words, this "uniform" probability on $[0, 1[$ is λ^1, the Lebesgue measure on $[0, 1[$. Note that $([0, 1[, \lambda^1)$ is not a discrete probability space.

(2) Roll a die: $|\Omega| = 6$

(3) Roll a die 3 times: $|\Omega| = 6^3$

(4) 20 coin tosses: $|\Omega| = 2^{20} \approx 10^6$ since $2^{10} = 1,024 \approx 10^3$.

(5) 10^9 coin tosses: $|\Omega| = 2^{10^9} = 2^{10 \cdot 10^8} = \left(2^{10}\right)^{10^8} \approx \left(10^3\right)^{10^8} = 10^{3 \cdot 10^8}$

(6) A selection of n items from a population is a sample of size n. The "most random" action of obtaining such a sample would be one where each subset of size n from that population is selected with the same likelihood as any other. □

5.2 Conditional Probability and Independent Events

Definition 5.7 (Conditional probability)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$ and two events $A, B \in \mathscr{F}$. We call

$$P(A \mid B) := \begin{cases} \dfrac{P(A \cap B)}{P(B)}, & \text{if } P(B) > 0, \\ \textbf{undefined}, & \text{if } P(B) = 0, \end{cases} \tag{5.22}$$

(read: "probability of A given B" or "probability of A conditioned on B") the **conditional probability** of the event A, given that the event B has occurred. □

Theorem 5.8.

Given are a probability space $(\Omega, \mathscr{F}, P)$ and an event $B \in \mathscr{F}$ such that $P(B) > 0$. Then

$$P(\cdot \mid B) : \mathfrak{F} \longrightarrow [0, 1]; \qquad A \mapsto P(A \mid B) \tag{5.23}$$

is another probability measure on $(\Omega, \mathscr{F})$.

In other words, $P(\cdot \mid B)$ satisfies (5.9) – (5.11) of Definition 5.2 (Probability measures and probability spaces) on p.106.

PROOF: First, it follows from $\emptyset \subseteq A \cap B \subseteq B$ that $P(A \cap B)/P(B) \geq 0$ and $P(A \cap B)/P(B) \leq 1$. This shows that $P(\cdot \mid B)$ indeed takes values between 0 and 1.

PROOF of (5.9): Since $P(\emptyset \cap B) = 0$, $\quad P(\emptyset \mid B) = 0/P(B) = 0$.

PROOF of (5.10): Since $\Omega \cap B = B$, $\quad P(\Omega \mid B) = P(\Omega \cap B)/P(B) = P(B)/P(B) = 1$.

PROOF of (5.11): Assume that $(A_n)_{n \in \mathbb{N}} \in \mathfrak{F}$ is a sequence of disjoint events. Then, for $i \neq j$,

$$(A_i \cap B) \cap (A_j \cap B) \subseteq A_i \cap A_j = \emptyset.$$

Thus, the sequence $(A_n \cap B)_{n\in\mathbb{N}}$ also is mutually disjoint. Further, by (2.37) on p.41,

$$\biguplus_{n\in\mathbb{N}} (B \cap A_n) \;=\; B \cap \biguplus_{n\in\mathbb{N}} A_n \,.$$

It follows from this and the σ–additivity of P that

$$\begin{aligned} P\left(\biguplus_{n\in\mathbb{N}} A_n \mid B\right) &= \frac{P\left(B \cap \biguplus_{n\in\mathbb{N}} A_n\right)}{P(B)} \;=\; \frac{P\left(\biguplus_{n\in\mathbb{N}} (B \cap A_n)\right)}{P(B)} \\ &= \frac{\sum_{n\in\mathbb{N}} P(B \cap A_n)}{P(B)} \;=\; \sum_{n\in\mathbb{N}} \frac{P(B \cap A_n)}{P(B)} \;=\; \sum_{n\in\mathbb{N}} P(A_n \mid B)\,. \end{aligned}$$

We have shown that $P(\cdot \mid B)$ is σ–additive and this proves (5.11). ■

It is immediate from the definition of $P(A \mid B)$ that

$$P(A \cap B) \;=\; P(A \mid B) \cdot P(B)\,.$$

This formula is referred to by WMS as the **multiplicative law of probability**. It can be extended to three events as follows.

Proposition 5.3. *If $(\Omega, \mathfrak{F}, P)$ is a probability space and $A, B, C \in \mathfrak{F}$, then*

$$P(A \cap B \cap C) \;=\; P(A \mid B \cap C) \cdot P(B \mid C) \cdot P(C)\,. \tag{5.24}$$

PROOF:

$$P(A \cap B \cap C) \;=\; P(A \mid B \cap C) \cdot P(B \cap C)\,. \;=\; P(A \mid B \cap C) \cdot P(B \mid C) \cdot P(C)\,. \;\blacksquare$$

The multiplicative law of probability generalizes to arbitrarily many sets as follows.

Proposition 5.4 (Multiplicative Law of Probability for n events)**.**

If $(\Omega, \mathfrak{F}, P)$ is a probability space, $n \in \mathbb{N}$ and $A_1, \ldots, A_n \in \mathfrak{F}$, then

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_n) \;&= P(A_1 \mid A_2 \cap \cdots \cap A_n) \cdot P(A_2 \mid A_3 \cdots \cap A_n) \cdots \\ &\quad \cdots P(A_{n-2} \mid A_{n-1} \cap A_n)\, P(A_{n-1} \mid A_n)\, P(A_n). \end{aligned} \tag{5.25}$$

PROOF:

It is easier to work with the reverse sequence $A_n \cap A_{n-1} \cap \cdots \cap A_1$ instead of $A_1 \cap A_2 \cap \cdots \cap A_n$. Repeated use of $P(U \cap V) = P(U \mid V)P(V)$ with $U = A_j$ and $V = A_{j-1} \cap \cdots \cap A_1$ yields

$$\begin{aligned} &P(A_n \cap A_{n-1} \cap \cdots \cap A_1) \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} \cap \cdots \cap A_1) \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} \mid A_{n-2} \cdots \cap A_1)\, P(A_{n-2} \cdots \cap A_1) \\ &= \ldots\ldots\ldots\ldots\ldots\ldots \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} \mid A_{n-2} \cdots \cap A_1) \cdots P(A_3 \mid A_2 \cap A_1)\, P(A_2 \mid A_1)\, P(A_1). \;\blacksquare \end{aligned}$$

Definition 5.8 (Two independent events)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$ and two events $A, B \in \mathscr{F}$. We say that A and B are **independent** if

$$P(A \cap B) \;=\; P(A) \cdot P(B)\,. \quad \square \tag{5.26}$$

Independence of three events is not defined as you may have guessed from that last definition.

Definition 5.9 (Three independent events)**.** Given are a probability space $(\Omega, \mathscr{F}, P)$ and three events $A, B, C \in \mathscr{F}$. We say that A, B and C are **independent** if

$$\begin{aligned} P(A \cap B \cap C) \;&= P(A) \cdot P(B) \cdot P(C)\,, \\ P(A \cap B) \;&= P(A) \cdot P(B)\,, \\ P(A \cap C) \;&= P(A) \cdot P(C)\,, \\ P(B \cap C) \;&= P(B) \cdot P(C)\,. \quad \square \end{aligned} \tag{5.27}$$

We can state (5.27) as follows. It must be true for <u>any</u> subsequence of events that the probability of the intersection equals the product of the probabilities of the individual events.

Remark 5.13. It is possible to construct a probability measure P and events A, B, C such that $P(A \cap B \cap C) \;=\; P(A) \cdot P(B) \cdot P(C)$ and $P(A \cap B) \;\neq\; P(A) \cdot P(B)$ $\square$

Definition 5.9 shows us how to generalize independence to any number of events.

Definition 5.10 (Finitely many independent events)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$, $n \in \mathbb{N}$ and events $A_1, A_2, \ldots, A_n \in \mathscr{F}$. We say that $A_1, A_2, \ldots, A_n$ are **independent** if, for ANY subselection of indices

$$1 \leq j_1 \;<\; j_2 \;<\; \cdots \;<\; j_k \;\leq\; n\,,$$

it is true that

$$P(A_{j_1} \cap A_{j_1} \cap A_{j_k}) \;=\; P(A_{j_1}) \cdot P(A_{j_2}) \cdot P(A_{j_k})\,. \quad \square \tag{5.28}$$

Finally, we define independence for infinitely many events.

Definition 5.11 (Sequences of independent events)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a sequence of events $A_1, A_2, \cdots \in \mathscr{F}$ We say that this sequence is **independent** if, for ANY FINITE subselection of distinct indices $j_1, j_2, \ldots, j_k \in \mathbb{N}$, it is true that

$$P(A_{j_1} \cap A_{j_2} \cap A_{j_k}) \;=\; P(A_{j_1}) \cdot P(A_{j_2}) \cdot P(A_{j_k})\,. \quad \square \tag{5.29}$$

Remark 5.14. Note that the number k in Definition 5.10 and Definition 5.11 is not fixed. □

We did not really define independence for any collection of infinitely many events, only for a sequence, i.e., a countable collection of events. The truly general case deals with families (see Definition 2.25 on p.38) of events

Definition 5.12 (Independence of arbitrarily many events). ⋆

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\big(A_i\big)_{i\in I}$ of events $A_i \in \mathscr{F}$. Here I denotes an arbitrary set of indices. We say that this family is **independent** if, for ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$, it is true that

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_k}) \;=\; P(A_{i_1}) \cdot P(A_{i_2}) \cdot P(A_{i_k})\,. \tag{5.30}$$ □

The next theorem is marked optional, but it is just as easy to remember as the corollary that follows it.

Theorem 5.9. ⋆

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\big(A_i\big)_{i\in I}$ of independent events $A_i \in \mathscr{F}$. Here I denotes an arbitrary set of indices. Then we have the following:

If some or all of the A_i are replaced by their complement $A_i^{\complement}$, then the resulting family of events also is independent.

In other words, for each $i \in I$, let B_i be either A_i or $A_i^{\complement}$. Then independence of $\big(A_i\big)_{i\in I}$ implies that of $\big(B_i\big)_{i\in I}$.

PROOF: Utilizes advanced probabilistic methods that are outside the scope of this course ■

Note that the following corollary is NOT marked as optional!

Corollary 5.2.

Given are a $(\Omega, \mathfrak{F}, P)$ is a probability space, $n \in \mathbb{N}$ and independent events $A_1, \ldots, A_n \in \mathfrak{F}$.
If some or all of the A_i are replaced by their complement $A_i^{\complement}$, then the resulting list of events also is independent.

In other words, for each $i = 1, 2, \ldots, n$, let B_i be either A_i or $A_i^{\complement}$. Then independence of $A_1, \ldots, A_n$ implies that of $B_1, \ldots, B_n$.

PROOF: ⋆

(A): The case $n = 2$ shows the essence of the proof: For convenience, let $B := A_2^{\complement}$. First, we show that A_1 and B are independent.

$$\begin{aligned} A_1 \;&= (A_1 \cap A_2) \uplus (A_1 \cap B) \;\Rightarrow\; P(A_1) \;=\; P(A_1 \cap A_2) \;+\; P(A_1 \cap B) \\ &= P(A_1) \cdot P(A_2) \;+\; P(A_1 \cap B) \\ \Rightarrow\; P(A_1 \cap B) \;&= P(A_1) \cdot \big(1 - P(A_2)\big) \;=\; P(A_1) \cdot P(B)\,. \end{aligned}$$

Thus, A_1 and $A_2^{\complement}$ are independent. Since intersection is commutative ($E \cap E' = E' \cap E$), it follows that $A_1^{\complement}$ and A_2 also are independent.

Knowing that $A_1^{\complement}$ and A_2 are independent, we can apply the proof above to those two independent events and obtain that $A_1^{\complement}$ and $A_2^{\complement}$ are independent. This finishes the proof for $n = 2$

(B): For general n, let $A_1, \ldots, A_n$ be independent. For convenience, let $B := A_1 \cap \cdots \cap A_{n-1}$.

Since $P(B \cap A_n) = P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n) = P(B) \cdot P(A_n)$, B and A_n are independent. We have shown in **(A)** that B and $A_n^{\complement}$ are independent, too.

We argue as in **(A)** and conclude from the commutativity of "$\cap$" that replacing any A_j with its complement, i.e., fixing an index j_1 and defining $B_j := A_j$ for $j \neq j_1$ and $B_{j_1} := A_{j_1}^{\complement}$, that $B_1, \ldots, B_n$ are independent In other words, replacing just one event with it complement maintains independence.

We apply this to the events $C_j := B_j$ for $j \neq j_2$ and $C_{j_2} := B_{j_2}^{\complement}$, where we assume that $j_2 \neq j_1$. The result is that $C_1, \ldots, C_n$ also are independent

At this point we know that replacing $k = 1$ or $k = 2$ events with their complements maintains independence. We apply this to the events $D_j := C_j$ for $j \neq j_3$ and $D_{j_3} := B_{j_3}^{\complement}$, where we assume that $j_2 \notin \{j_1, j_2$. The result is that $D_1, \ldots, D_n$ also are independent.

At this point we know that replacing $k \leq 3$ events with their complements maintains independence. We repeat the above with $k = 4$, then with $k = 5$, ..., then with $k = n$. This completes the proof. ■

Next, we examine connections between conditional probabilities and independence.

Theorem 5.10.

Given are a probability space $(\Omega, \mathscr{F}, P)$ and two events $A, B \in \mathscr{F}$ such that $P(B) > 0$. Then

(5.31) $$A \text{ and } B \text{ are independent} \quad \Leftrightarrow \quad P(A \mid B) = P(A)\,.$$

PROOF of "$\Rightarrow$":

Since A and B are independent and $P(B) > 0$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)\,.$$

PROOF of "$\Leftarrow$":

Since $P(A \mid B) = P(A)$ and $P(B) > 0$,

$$P(A) \cdot P(B) = P(A \mid B) \cdot P(B) = \frac{P(A \cap B)}{P(B)} \cdot P(B) = P(A \cap B)\,. \quad ■$$

Corollary 5.3.

If $(\Omega, \mathscr{F}, P)$ is a probability space and $A, B \in \mathscr{F}$ such that $P(A) > 0$ and $P(B) > 0$. Then

(5.32) $$A \text{ and } B \text{ are independent} \quad \Leftrightarrow \quad P(A \mid B) = P(A) \quad \Leftrightarrow \quad P(B \mid A) = P(B)\,.$$

PROOF: Obious ■

5.3 Random Elements and their Probability Distributions

Introduction 5.2. We continue with an observation we made in the introduction 2.3 to Section 2.5 (Preimages, p.41). There,

- $\Omega = \{1,2,\ldots,6\}^2$ and $\vec{\omega} = (\omega_1,\omega_2)$ represents a potential (two–number) outcome of two rolls of a fair die, i.e., $P(\{\vec{\omega}\}) = 1/|\Omega| = 1/36$.
- We defined the function $Y : \Omega \to \Omega' := \{2,3,4,\ldots,11,12\};\ \vec{\omega} \mapsto Y(\vec{\omega}) := \omega_1 + \omega_2$, which associates with $\vec{\omega} = (\omega_1,\omega_2)$ the sum of the two rolls.
- This function lead to a probability measure P' on Ω' by means of formula (2.39):

$$B \subseteq \Omega' \ \Rightarrow\ P'(B) \ = \ P\{\vec{\omega} \in \Omega : Y(\vec{\omega}) \in B\}\,.$$

  Observe that the set Ω' has been transformed into a probability space, (Ω', P')).
- With preimage notation and the notational shortcuts of Remark 2.4 on p.45, this can also be written as

$$P'(B) \ = \ P\big(Y^{-1}(B)\big) \ = \ P\{Y \in B\}\,.$$

These formulas can be written for an arbitrary probability space (Ω, P), an arbitrary nonempty set Ω', and an arbitrary function $Y : \Omega \to \Omega'$. Actually, that so only because we disregard the role of σ–algebras and measurability. [61] $\square$

The next theorem and the subsequent definitions are very important.

Theorem 5.11.

Let (Ω, P) be a probability space, Ω' a nonempty set, and $Y : \Omega \to \Omega'$ a function. Then the formula

(5.33) $$P_Y(B) \ := \ P\{Y \in B\} \ \ (B \subseteq \Omega')$$

defines a probability measure on Ω'.

PROOF: $\boxed{\star}$ It follows from $\{Y \in \emptyset\} = \emptyset$ and $\{Y \in \Omega'\} = \Omega$, that

$$P_Y(\emptyset) \ = \ P(\emptyset) \ = \ 0 \quad \text{and} \quad P_Y(\Omega') \ = \ P(\Omega) \ = \ 1\,.$$

Let $B \subseteq \Omega'$. From (2.46) on p.45, we obtain

$$P_Y(B^{\complement}) \ = \ P\{Y \in B^{\complement}\} \ = \ P\big(Y^{-1}(B^{\complement})\big) \ = \ P\big([Y^{-1}(B)]^{\complement}\big) \ = \ 1 \ - \ P\big(Y^{-1}(B)\big) \ = \ 1 \ - \ P_Y(B)\,.$$

To prove σ–additivity of P_Y, we apply (2.45) to the index set $\mathbb{N}$ of a sequence of disjoint subsets $B_1, B_2, \ldots$ of Ω'. Let $B := B_1 \uplus B_2 \uplus B_3 \uplus \cdot$. Then

$$P_Y(B) \ = \ P\big(Y^{-1}\left(\biguplus_{j\in\mathbb{N}} B_j\right)\big) \ = \ P\left(\bigcup_{j\in\mathbb{N}} Y^{-1}(B_j)\right)$$

[61] For measurability, see the optional section **??** (Advanced Topics – Measurable Functions)

By (2.47), the sets $Y^{-1}(B_j)$ are disjoint. Thus,

$$P_Y(B) \;=\; P\left(\biguplus_{j\in\mathbb{N}} Y^{-1}(B_j)\right) \;=\; \sum_{j\in\mathbb{N}} P\big(Y^{-1}(B_j)\big) \;=\; \sum_{j\in\mathbb{N}} P_Y(B_j)\,.$$

This proves σ–additivity. ■

Definition 5.13 (Probability Distribution)**.**

Let (Ω, P) be a probability space, Ω' a nonempty set, and $Y : \Omega \to \Omega'$ a function. Then the probability measure P_Y on Ω' of Theorem 5.11, given by

$$P_Y(A') \;:=\; P\{Y \in A'\} \;=\; P\left(Y^{-1}(A')\right) \quad (A' \subseteq \Omega')\,, \tag{5.34}$$

is called the **probability distribution** or just the **distribution** of Y with respect to P. Very often the probability space (Ω, P) is fixed for a long stretch. We then simply talk about the probability distribution of Y, without referring to P. □

Definition 5.14 (Random Variables and Random Vectors)**.** Let (Ω, P) be a probability space and let $n \in \mathbb{N}$.

Let $B \subseteq \mathbb{R}$. A function

$$Y \;:\; \Omega \longrightarrow B\,; \quad \omega \mapsto Y(\omega)$$

is called a **random variable** (in short, **r.v.** or **rv**,) on $(\Omega, \mathfrak{F}, P)$. Let $B' \subseteq \mathbb{R}^n$. A function

$$\vec{X} = \big(X_1, X_2, \ldots, X_n\big) \;:\; \Omega \longrightarrow B'\,; \quad \omega \mapsto \vec{X}(\omega) \;=\; \big(X_1\omega), \ldots, X_n(\omega)\big)$$

is called a **random vector** on $(\Omega, \mathfrak{F}, P)$.
If there is a countable subset $B^* = \{y_1, y_2, \ldots\}$ of B such that $\sum_j P_Y\{y_j\} = 1$ (i.e., $P\{Y \notin B^*\} = 0$), we call Y a **discrete random variable**. Likewise, if there is a countable subset B'^* of B' such that $P\{\vec{X} \notin B'^*\} = 0$, we call $\vec{X}$ a **discrete random vector**.

Note that random variables and vectors which have a countable range are discrete. □

Remark 5.15. In many instances the exact nature of the codomain B of a random variable Y is unimportant. Of course it must be a set of numbers, i.e., $B \subseteq \mathbb{R}$, and it must be big enough to accommodate all function values $Y(\omega)$, i.e., $Y(\omega) \subseteq B$. [62] Thus, here is some **good news**.

We often will just say something like "Let Y be a random variable on Ω" or, "Let Y be a discrete random vector on Ω" and not even mention the codomain of Y. □

Not all interesting functions on a probability space take values in $\mathbb{R}$ or $\mathbb{R}^n$. Here is an example.

[62] It only matters when we need the inverse function $\omega = Y^{-1}(y)$ of $y = Y(\omega)$. (Do not confuse inverse function and preimage, just because they use the same symbol Y^{-1}!) Then $Y^{-1}(y)$ must make sense for all $y \in B$ and that requires that B is minimal: $B = Y(\Omega)$. The same thought also applies to random vectors.

Example 5.9. The following describes a (unnecessarily complicated) way to simulate n tosses of a fair coin. Le Let $\Omega := [0,1[$, where we represent the real number $\omega \in \Omega$ as a decimal $0.d_1d_2d_3$ with inifinitely many decimal digits. If necessary, we append infinitely many zeroes to the right. For example, we write $0.25000\ldots$ for the number $1/4$. We write H for Heads and T for Tails and define the following function on (Ω, P).

$$\vec{X} : \Omega \to \{H,T\}^n$$

- $X_1(\omega) = H$ if d_1 is even, T else.
- $X_2(\omega) = H$ if d_2 is even, T else.
- .....
- $X_n(\omega) = H$ if d_n is even, T else.

Since $P_{\vec{X}}(\vec{x}) = 1/2^n$ for each $\vec{x} \in \{H,T\}^n$, each combination of a total of n Heads and Tails has the same chance to occur. That is our understanding of a fair coin. □

Considering that last example, it seems awkward not to call a function $\Omega \to \Omega'$ from a probability space (Ω, P) to a set Ω' a random variable only because its function values are not numbers. We give a name to such functions of randomness.

Definition 5.15 (Random element)**.**

Let $(\Omega, \mathfrak{F}, P)$ be a probability space and Ω' a nonempty set.
We call a function $X : \Omega \to \Omega'$ a **random element**, also: a **random item**, on Ω. □

Remark 5.16. We can phrase Theorem 5.11 on p.123 and the subsequent Definition 5.13 as follows. All random elements X on a probability space $(\Omega, \mathfrak{F}, P)$ have a distribution

$$P_X(B) \;=\; P\{X \in B\} \;=\; P\big(X^{-1}(B)\big) \;\; (B \subseteq \Omega'). \;\square$$

For a collection $\boldsymbol{\mathcal{A}}$ of subsets of Ω, $\sigma\{\boldsymbol{\mathcal{A}}\}$ denotes the minimal σ–algebra that contains $\boldsymbol{\mathcal{A}}$. [63] In particular, given random elements $X, X_i : (\Omega, P) \to \Omega', \; (i \in I)$, we can consider the sets of preimages

$$\boldsymbol{\mathcal{A}}_1 := \{X^{-1}(A') : A' \subseteq \Omega'\} \qquad \boldsymbol{\mathcal{A}}_2 := \{X_i^{-1}(A') : A' \subseteq \Omega', i \in I\} \tag{5.35}$$

and the σ–algebras that are generated by those collections of preimages, $\boldsymbol{\mathcal{A}}_1$ and $\boldsymbol{\mathcal{A}}_2$. The σ–algebras are so important that they have their own notation.

Definition 5.16 (σ–algebra generated by random elements [64])**.** ⋆ Let (Ω, P) be a probability space.

(a) Let $X : (\Omega, P) \to \Omega'$ be a random element on (Ω, P). We call

$$\sigma\{X\} \;:=\; \sigma\{X^{-1}(A') : A' \subseteq \Omega'\} \tag{5.36}$$

the **σ–algebra generated by the random element** X.

[63] See Definition 5.5 (σ–algebra generated by a collection of sets). on p.114

[64] A more mathy version of this is Definition 6.4 (Advanced Definition of σ–algebra generated by random elements) in the optional Chapter 6 (Advanced Topics – Measure and Probability). See p.143.

(b) Let $X_i : (\Omega, P) \to \Omega',\ i \in I$ be a family of random elements on (Ω, P). We call

(5.37) $$\sigma\big\{(X_i)_{i\in I}\big\} \ := \ \sigma\{X_i : i \in I\} \ := \ \sigma\{X_i^{-1}(A') : A' \subseteq \Omega',\ i \in I\}$$

the **σ–algebra generated by the family of random elements** $(X_i)_{i\in I}$.

Concerning notation:

- As usual, it is OK to omit the "$i \in I$" part if the meaning of I is unambiguous.
- The square braces of $\sigma\{\dots\}$ can be replaced with square braces or parentheses. For example, $\sigma\{Y_n : n = 1, \dots, 10\} = \sigma(Y_k : k = 1, \dots, 10)$, and $\sigma\{U\} = \sigma[U]$.
- But **BEWARE**: When we work on the applications side and X is a random variable, i.e., an $\mathbb{R}$–valued random element, it is very common practice do write $\sigma(X)$ or $\sigma[X]$ for the so called standard deviation of X. [65] There are alternate notations such as σ_X for this standard deviation, but WMS uses $\sigma(X)$ frequently. I try to stick with curly braces for σ–algebras generated by random elements and/or sets. □

Example 5.10. Let $H :=$ Heads, $T :=$ Tails, $\Omega' := \{H, T\}$, and $\mathfrak{F}' := 2^{\Omega'} = \{\emptyset, \{H\}, \{T\}, \Omega'\}$. For $j \in \{1, 2\}$, let

$$X_j : (\Omega, \mathfrak{F}, P) \to \Omega'; \qquad \omega \mapsto X_j(\omega)$$

denote two flips of a coin. The reader should verify that

$$\begin{aligned}\sigma\{X_1, X_2\} \ = \ \big\{&\emptyset,\ \Omega,\ \{X_1 = H\},\ \{X_1 = T\},\ \{X_2 = H\},\ \{X_2 = T\},\\ &\{X_1 = H, X_2 = H\},\ \{X_1 = H, X_2 = T\},\ \{X_1 = T, X_2 = H\},\ \{X_1 = T, X_2 = T\}\big\}.\end{aligned}$$

By the way, this his is the set of all finite unions that can be obtained from the partition

$$\Omega \ = \ \{X_1 = H, X_2 = H\} \uplus \{X_1 = H, X_2 = T\}, \uplus \{X_1 = T, X_2 = H\}, \uplus \{X_1 = T, X_2 = T\}$$

of Ω. □

Example 5.11. ⋆ The simplest examples for Definition 5.16 (σ–algebra generated by random elements) are given by random elements that only take one or two function values.

(a): Let $X(\omega) = c'$ for all $\omega \in \Omega$, i.e., X is constant on Ω. There are only two types of of sets $B' \subseteq \Omega'$. Either $c' \in B'$ or $c' \notin B'$.

(a) Let $X(\omega) = c'$ for all $\omega \in \Omega$, i.e., X is constant on Ω. There are only two types of of sets $B' \subseteq \Omega'$. Either $c' \in B'$ or $c' \notin B'$.
• $c' \in B' \ \Rightarrow \ X^{-1}(B) = \Omega$, • $c' \notin B' \ \Rightarrow \ X^{-1}(B) = \emptyset$.
Thus, $\sigma\{X\} = \sigma\{\emptyset, \Omega\}$. Since $\{\emptyset, \Omega\}$ itself is a σ–algebra, we obtain that $\sigma\{X\} = \{\emptyset, \Omega\}$.

(b) Let $A \subseteq \Omega$. Consider the random variable $\mathbf{1}_A$, the indicator function of A. There are four types of of sets $B' \subseteq \Omega'$.
• $0 \in B', 1 \in B' \ \Rightarrow \ X^{-1}(B) = \Omega$, • $0 \in B', 1 \notin B' \ \Rightarrow \ X^{-1}(B) = A^{\complement}$,
• $0 \notin B', 1 \in B' \ \Rightarrow \ X^{-1}(B) = A$, • $0 \notin B', 1 \notin B' \ \Rightarrow \ X^{-1}(B) = \emptyset$.
Thus, $\sigma\{X\} = \sigma\{\emptyset, A, A^{\complement}\Omega\} = \{\emptyset, A, A^{\complement}\Omega\}$ (since $\{\emptyset, A, A^{\complement}\Omega\}$ already is a σ–algebra). □

[65] See Definition 9.3 (Variance and standard deviation of a random variable) on p.195

Remark 5.17. ⋆ We compare Example 5.11**(b)** with Example 5.8 on p.114 and see the following:

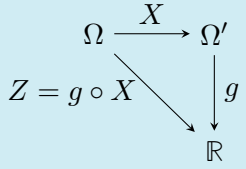
Let $A \subseteq \Omega$. Then A and $\mathbf{1}_A$, the indicator function of A, generate the same σ–algebra

$$\sigma\{A\} \;=\; \sigma\{\mathbf{1}_A\} \;=\; \{\emptyset, A, A^{\complement}\Omega\}\,. \;\square$$

Since an element x of the domain of a function f (an argument) is assigned to only one function value $y = f(x)$, one should expect that a function of a discrete random element should again be discrete. This is the assertion of the next proposition and the corollary that follows it.

Proposition 5.5. ⋆

- *Let $X : (\Omega, P) \to \Omega'$ be a random element and $g : \Omega' \to \mathbb{R}$.*
- *Let Z be the random variable $\omega \mapsto Z(\omega) := g\big(X(\omega)\big)$.*
- *Let $B^* \in \Omega'$ such that $P_X(B^*) = 1$ and let $C^* := \{g(x) : x \in B^*\}$ be the direct image $g(B^*)$ of B^* under g. (See Definition 2.28 on p.46.)*



Then $P_Z(C^*) = 1$.

PROOF: Let

(A) $$A_1 \;:=\; \{\omega \in \Omega : Z(\omega) \notin C^*\} \;= \{\omega \in \Omega : g\big(X(\omega)\big) \notin C^*\}\,.$$

(B) $$\text{Then } \tilde{\omega} \in X^{-1}(B^*) \;\Leftrightarrow\; X(\tilde{\omega}) \in B^* \;\Rightarrow\; Z(\tilde{\omega}) \;=\; g\big(X(\tilde{\omega})\big) \in g(B^*) \;=\; C^*$$

Here, "⇔" follows from the definition of X^{-1}. From **(A)** + **(B)** we see that $A_1 \cap X^{-1}(B^*) = \emptyset$.

(C) Thus, $A_1 \subseteq \big[X^{-1}(B^*)\big]^{\complement}$.

(D) $$\text{Since } P\big[X^{-1}(B^*)\big] \;=\; P_X(B^*) \;=\; 1 \quad (\text{by definition of } B^*)\,,$$

we obtain from **(C)** that $P(A_1) = 0$ and then, from **(A)**, that

(E) $$P_Z(C^*) \;= P\{\omega \in \Omega : Z(\omega) \in C^*\} \;=\; P(A_1^{\complement}) \;=\; 1\,. \;\blacksquare$$

Corollary 5.4. *Let $X : (\Omega, P) \to \Omega'$ be a random element and $g : \Omega' \to \mathbb{R}$. Further, let Z be the random variable $g \circ X : \omega \mapsto Z(\omega) = g\big(X(\omega)\big)$. In other words, Z is the composition of g with X. Then*

- ***(a)*** *If $\omega \mapsto X(\omega)$ only assumes finitely many (distinct) values $x_1, \ldots, x_n$, then $\omega \mapsto Z(\omega)$ only assumes finitely many values $z_1, \ldots, z_m$ (and $m \le n$).*
- ***(b)*** *If $\omega \mapsto X(\omega)$ only assumes an infinite sequence of (distinct) values (x_j), then $\omega \mapsto Z(\omega)$ assumes a countable set of function values. (This set forms a finite or infinite sequence. (See Definition 2.24 (Countable and uncountable sets) on p.37).*
- ***(c)*** *If X is a discrete random element, then $Z = g(X)$ is a discrete random variable.*

PROOF of **(a):** ⋆ The potential function values of Z are

$$z'_1 := g(x_1), z'_2 := g(x_2), \ldots, z'_n := g(x_n)$$

If g is not injective, there may be duplicate z'_j which must be removed. Thus, Z assumes at m distinct values for some suitable $m \leq n$. We rename them $z_1, \ldots, z_m$.

PROOF of **(b):** ⋆ The potential function values of Z the members of the sequence $z'_j = g(x_j)$, where $j \in \mathbb{N}$. Removing the duplicates leaves us with a finite or infinite subsequence of distinct items z_j and those form the countable set of all function values of Z.

PROOF of **(c):** ⋆ Since X is discrete, there is a countable set $B^* \subseteq \Omega'$ such that $P_X(B^*) = 1$. We have seen in the proof of **(b)** that a function g transports countably many arguments b^* into countably many function values $c^* = g(b^*)$. Thus, the set $C^* := \{g(b^*) : b^* \in B^*\}$ is countable. It follows from Proposition5.5 on p.127 that $P_Z(C^*) = 1$. Since C^* is countable, Z is discrete. ■

Remark 5.18. If $Q(E) = 1$ for some probability measure Q and some event E on some probability space, then $Q(F) = 0$ for all events $F \subseteq E^{\complement}$. That does not necessarily make it impossible for F to happen, but it would be so improbable, we do not take this possibility into account. A good way to think about the probability measure Q in relation to E is that Q "lives" on E.

This situation happens twice in the context of Proposition 5.5 on p.127.

(1) On the probability space (Ω', P_X): $P_X(B^*) = 1$. (Here, $Q = P_X$, and $E = B^*$).
(2) On the probability space $(\mathbb{R}, P_Z)$: $P_Z(C^*) = 1$. (Here, $Q = P_Z = P_{g\circ X}$, and $E = C^*$).

Considering that $C^* = g(B^*)$, Proposition 5.5 states the following:

(3) If P_X lives on B^* then $P_Z = P_{g\circ X}$ lives on $g(B^*)$.

We improve on **(3)** by showing that the distribution of $g \circ X$ under the probability measure P equals the distribution of g under the probability measure P_X, In short, we will show that

(4) $P_{g\circ X} = \big(P_X\big)_g$, i.e., $P_{g\circ X}(C) = \big(P_X\big)_g(C)$, for all arguments $C \subseteq \mathbb{R}$.

To prove **(4)**, note that for any probability space $(\widetilde{\Omega}, \widetilde{P})$, random element $\widetilde{X} : \widetilde{\Omega} \to \widetilde{\Omega}'$, and $B \subseteq \widetilde{\Omega}'$, the expression $\{\widetilde{X} \in B\}$ merely is a notational convenience for the preimage of B under $\widetilde{X}$:

$$\{\widetilde{X} \in B\} = \widetilde{X}^{-1}(B) = \{\tilde{\omega} \in \widetilde{\Omega} : \widetilde{X}(\tilde{\omega}) \in B\}.$$

Thus, by (5.34) on p.124, $\widetilde{P}_{\widetilde{X}}(B) = \widetilde{P}\big[\widetilde{X}^{-1}(B)\big]$. It follows that

$$P_{g\circ X}(C) = P\big[(g\circ X)^{-1}(C)\big] \quad \text{(with } P = P,\ \widetilde{X} = g\circ X, B = C\text{)}, \tag{5.38}$$

$$P_X\big(g^{-1}(C)\big) = P\big[X^{-1}\big(g^{-1}(C)\big)\big] \quad \text{(with } \widetilde{P} = P,\ \widetilde{X} = X,\ B = g^{-1}(C)\text{)}, \tag{5.39}$$

$$(P_X)_g(C) = P_X\big[g^{-1}(C)\big] \quad \text{(with } \widetilde{P} = P_X,\ \widetilde{X} = g,\ B = C\text{)}. \tag{5.40}$$

Also recall (2.50) on p.46 for the preimage of the composition of functions:

$$(g\circ X)^{-1}(B) = X^{-1}\big(g^{-1}(B)\big). \tag{5.41}$$

We have everything in place to show that **(4)** is true. Let $C \subseteq \mathbb{R}$. Then

$$P_{g\circ X}(C) \overset{(5.38)}{=} P\big[(g\circ X)^{-1}(C)\big] \overset{(5.41)}{=} P\big[X^{-1}\big(g^{-1}(C)\big)\big] \overset{(5.39)}{=} P_X\big[g^{-1}(C)\big] \overset{(5.40)}{=} (P_X)_g(C).$$

We have shown **(4)**. □

Remark 5.19. Consider the following of a philosophical rather than mathematical nature. Not all mathematicians agree with it.

I like to think of a probability space (Ω, P) as a seat of randomness in the following sense. Some all–powerful supreme being or supreme force of nature, let's call it $\boxed{\mathcal{SB}}$, decides to pick "this" particular $\omega_0 \in \Omega$. As a result, all random elements $X, Y, \vec{Z}, \ldots$ that have Ω as domain are invoked with ω_0 as argument, resulting in the outcomes $X(\omega_0), Y(\omega_0), \vec{Z}(\omega_0), \ldots$. With this interpretation it makes a lot of sense to talk about functions on (Ω, P) as **random** elements since, when we interpret $\omega \in \Omega$ as "randomness",

$$x \;=\; X(\omega) \;\text{ simply means that }\; x \text{ is a function of randomness.}$$

Only $\boxed{\mathcal{SB}}$ knows what ω_0 will be picked. But if we know, say, the distribution P_Y of a certain random variable Y, then we can at least quantify the likelihood that $\boxed{\mathcal{SB}}$ is going to chose an ω such that $17.8 \le Y(\omega) \le 21.3$. It is $P_Y([17.8, 21.3]) = P\{17.8 \le Y \le 21.3\}$. □

Example 5.12. Often it only is the distribution P_X of a random element

$$X : (\Omega, P) \longrightarrow (\Omega', P_X)$$

with values in a set Ω' that matters. Accordingly, only the set Ω' and the probability measure P_X on that set are specified by the problem. On the other hand, there are infinitely many different choices of the probability space (Ω, P) plus the random element X which result in thatsame probability measure on Ω'. We illustrate that with two more settings for the modeling of the distribution of n tosses of a fair coin on the space $\{H, T\}^n$. See Example 5.9 on p.125. We fix $n = 3$ since the resulting specific example illustrates all essential points. **(a)**, **(b)** and **(c)** give three different choices of a probability space (Ω, P) and a random element X on that probability space, such that for each one of **(a)**, **(b)** and **(c)**,

(1) $\Omega' = \{H, T\}^3$,

(2) (Ω, P) and $X : (\Omega, P) \longrightarrow \Omega'$ are constructed such that P_X, the distribution of X on Ω', is that of a fair coin: $P_X\{\omega'\} = 1/8$, for each one of the 8 outcomes $\omega' \in \Omega'$.

(a) Let $\Omega_1 := \{0, 1\}^3$ with the probability measure $P\{(a, b, c)\} = 1/|\Omega_1| = 1/8$.

Let $\vec{X}_1 : \Omega_1 \to \{H, T\}^3$ be the random element that changes each 1 into an H and each 0 into a T. For example, $\vec{X}_1(1, 0, 1) = (H, T, H)$ and $\vec{X}_1(0, 0, 1) = (T, T, H)$.

Then $P_{\vec{X}_1}$ is the same probability measure as $P_{\vec{X}}$ of **(2)**, since both assign the number $1/8$ to each element of $\{H, T\}^3$.

(b) Let $\Omega_2 := \{H, T\}^3$ with the probability measure $P\{(a, b, c)\} = 1/|\Omega_2| = 1/8$. (Same as in **(a)**, except that now a, b, c represent either of H or T rather than 0 or 1.)

Let the random element $\vec{X}_2 : \Omega_2 \to \{H, T\}^3$ be the **identity** (also, **identity function**) on Ω_2. That is the "do nothing" function which assigns each element of a set to itself, i.e., $\vec{X}_2(\omega) = \omega$ for all $\omega \in \Omega_2$. For example, $H, T, H) = (H, T, H)$ and $\vec{X}_1(T, T, H) = (T, T, H)$.

Clearly, $P_{\vec{X}_2}$ also assigns probability $P_{\vec{X}_2}(\{\omega\}) = 1/8$ to each element of $\{H, T\}^3$.

(c) Let $\Omega_3 := \{H, T\}^3 \times \{1, 2, 3, 4\}$ with the probability measure $P\{(a, b, c, d)\} = 1/|\Omega_3| = 1/32$.

Let $\vec{X}_3 : \Omega_3 \to \{H,T\}^3$ be the function defined as $\vec{X}_3(a,b,c,d) := (a,b,c)$. We compute the distribution $P_{\vec{X}_3}$ for the outcomes (a,b,c) of the probability space $(\{H,T\}^3, P_{\vec{X}_3})$ as follows.

$$\begin{aligned}(a,b,c) \in \vec{X}_3 \;\Rightarrow P_{\vec{X}_3}\{(a,b,c,d)\} \;&=\; P\{\vec{X}_3 = (a,b,c,d)\} \\ &= P\{(a,b,c,1),(a,b,c,2),(a,b,c,3),(a,b,c,4)\} \;=\; 4(1/32) \;=\; 1/8\,.\end{aligned}$$

We have have obtained in this example and Example 5.9 on p.125 the probability P' which models three tosses of a fair coin, i.e., $P'\{(a,b,c)\} = 1/8$ for each $(a,b,c) \in \{H,T\}^3$, as the distribution of four different random elements, $\vec{X}$ (see Example 5.9), $\vec{X}_1, \vec{X}_2, \vec{X}_3$, which were defined on four different probability spaces. This clearly demonstrates that we have multiple choices of probability spaces and random items to model a distribution. You will hopefully agree that $\vec{X}_1$ and $\vec{X}_2$ are much better choices than $\vec{X}$ and $\vec{X}_3$. □

The next remark lists the different types of probability.

Remark 5.20 (Types of probability). ⋆ We have encountered the following types of probability:

- The **empirical probability** of an event A is the relative frequency of its occurrence in the long run: if an experiment is performed n times and the event A is observed n_k times, $P(A) = \lim_{k\to\infty} n_k/k$. See Example 1.1 on p.6.
- **Equiprobability**: The probability space consists of a finite number N of outcomes and each outcome $\{\omega\}$ is assigned the same probability, $\{\omega\} = 1/N$. Other names for this probability are **theoretical probability**, **Laplace probability**. See Definition 5.3 on p.109.
- The **subjective probability** of an event reflects an individual's personal judgment or own experience about whether it is likely to occur. Subjective probability contains no formal calculations and only reflects the subject's opinions and past experience. [66] An example would be a student's assessment that her/his probability of getting an A or A- is between 0.75 and 0.9.
- **Axiomatic probability**: This is an abstract mathematical construct: the function values $P(A)$ of a probability measure $P : \Omega \to \mathbb{R}$ which obeys certain rules such as σ–addititivity. See Definition 5.2 on p.106. The axiomatic definition of probability is by far the most general and includes all of the other definitions presented here. □

5.4 Independence of Random Elements

Introduction 5.3. According to Definition 5.8 (Two independent events) on p.120, two events A and B are independent if

$$P(A \cap B) \;=\; P(A)\cdot P(B)\,. \tag{5.42}$$

The justification for doing so comes from (5.32) on p.122: If $P(A) > 0$ and $P(B) > 0$, then

$$A \text{ and } B \text{ are independent} \quad\Leftrightarrow\quad P(A \mid B) \;=\; P(A) \quad\Leftrightarrow\quad P(B \mid A) \;=\; P(B)\,.$$

[66] Source: Investopedia: Subjective Probability: How it Works, and Examples.

This formula is a good characterization of independence, since it states that there is no dependency between A and B in the sense that conditioning of one event on the other has no effect. We extended (5.42) to define independence of an arbitrary collection $(A_i)_{i\in I}$ of events in Definiton 5.12 (Independence of arbitrarily many events) 121 as follows:

ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ must satisfy

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_k}) \;=\; P(A_{i_1}) \cdot P(A_{i_2}) \cdot P(A_{i_k})\,. \tag{5.43}$$

It has been proven extremely fruitful from an application oriented perspective to model the independence of random elements on the above framework, by demanding the independence of the events that are associated with those random elements.

What are those events? It turns out that they are the elements of the σ–algebras

$$\sigma\{X_i\} \;=\; \sigma\{X_i^{-1}(A_i') : A_i' \subseteq \Omega'\}\,, \tag{5.44}$$

where $\sigma\{X_i\}$ is the σ–algebra generated by the random element X_i. [67] If we accept this, the appropriate way to define the independence of two random elements X_1, X_2, defined on a common probability space $(\Omega, \mathfrak{F}, P)$, should be

$$P(A_1 \cap A_2) \;=\; P(A_1) \cdot P(A_2)\,, \quad \text{for all } A_1 \in \sigma\{X_1\} \text{ and } A_2 \in \sigma\{X_2\}.$$

Further, the independence of an arbitrary family, $(X_i)_{i\in I}$, of random elements on $(\Omega, \mathfrak{F}, P)$, should be defined as follows: ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ must satisfy

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \;=\; P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k})\,, \text{ if } A_{i_j} \in \sigma\{X_{i_j}\}, \text{ and } j = 1, 2, \ldots, k. \tag{5.45}$$

One can show the following. [68] If $X : (\Omega, P) \to \Omega'$ is a random element, then

$$\sigma\{X\} \;=\; \{X^{-1}(A') : A' \subseteq \Omega'\}\,. \tag{5.46}$$

We recall that alternate notation for $\{X^{-1}(A')$ is $\{X \in A'$ and obtain from (5.46) that

$$A \in \sigma\{X\} \;\Leftrightarrow\; A \;=\; \{X \in A'\}, \text{ for some suitable } A' \subseteq \Omega'\,. \tag{5.47}$$

We rewrite (5.45) with the relevant events for the random elements X_i expressed as in (5.47) and arrive at the formal definition of the independence of those random elements. □

Definition 5.17 (Independence of arbitrarily many random elements)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\big(X_i\big)_{i\in I}$ of random elements on Ω. Here, I denotes an arbitrary set of indices. We say that this family is **independent** if, for ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ and $j = 1, 2, \ldots, k$,

$$\begin{aligned} &P\{\, X_{i_1} \in A_{i_1}',\, X_{i_2} \in A_{i_2}',\, \ldots,\, X_{i_k} \in A_{i_k}' \,\} \\ &\qquad = P\{X_{i_1} \in A_{i_1}'\} \cdot P\{X_{i_2} \in A_{i_2}'\} \cdots P\{X_{i_k} \in A_{i_k}'\}\,, \quad \text{for all } A_{i_j}' \subseteq \Omega'. \;\square \end{aligned} \tag{5.48}$$

[67] See Definition 5.16**(a)** (σ–algebra generated by random elements) on p.125.

[68] See Chapter 6 (Advanced Topics – Measure and Probability), Theorem 6.5**(a)** on p.144.

Example 5.13. Let $H := \text{Heads}$, $T := \text{Tails}$, $\Omega' := \{H, T\}$, and $\mathfrak{F}' := 2^{\Omega'} = \{\emptyset, \{H\}, \{T\}, \Omega'\}$. For $j \in \{1, 2\}$, let

$$X_j : (\Omega, \mathfrak{F}, P) \to \Omega'; \qquad \omega \mapsto X_j(\omega)$$

denote two flips of a coin. The reader should verify that

$$\begin{aligned} \sigma\{X_1\} &= \big\{\emptyset,\, \Omega,\, \{X_1 = H\},\, \{X_1 = T\}\,\big\}\,, \\ \sigma\{X_2\} &= \big\{\emptyset,\, \Omega,\, \{X_2 = H\},\, \{X_2 = T\}\,\big\}\,. \end{aligned}$$

To show the independence of X_1 and X_2, we must verify, for example, that

$$P\big(\,\{X_1 = T\} \cap \{X_2 = H\}\,\big) \;=\; P\{X_1 = T\} \cdot P\{X_2 = H\}\,.$$

This is true, since $P\{X_1 = T, X_2 = H\} = 1/4$ and $P\{X_1 = T\} = P\{X_2 = H\} = 1/2$. The other cases are dealt with just as easily. □

We modify the last example so it will be about an uncountable collection of independent random elements.

Example 5.14. Let $H := \text{Heads}$, $T := \text{Tails}$, $\Omega' := \{H, T\}$, and $\mathfrak{F}' := 2^{\Omega'} = \{\emptyset, \{H\}, \{T\}, \Omega'\}$. For $0 \le t \le 1$, let

$$X_t : (\Omega, \mathfrak{F}, P) \to \Omega'; \qquad \omega \mapsto X_t(\omega)$$

denote the flip of a fair coin at time t. Let $t \in [0, 1]$. By the same reasoning as in Example 5.13, we obtain

$$\sigma\{X_t\} \;= \big\{\emptyset,\, \Omega,\, \{X_t = H\},\, \{X_t = T\}\,\big\}\,.$$

To show the independence of $(X_t)_{0 \le t \le 1}$, we must verify that, for any selection of times,

$$0 \;\le\; t_1 \;<\; t_2 \;<\; \cdots \;<\; t_k \;\le\; 1\,,$$

it is true that, for any $j = 1, \ldots, k$ and for any choice of either $\omega_{t_j} = H$ or $\omega_{t_j} = T$,

$$P\big(\,\{X_{t_1} = \omega_{t_1}\} \cap \cdots \cap \{X_{t_k S} = \omega_{t_k}\} \;=\; P\{X_{t_1} = \omega_{t_1}\} \cdots P\{X_{t_k} = \omega_{t_k}\}\,.$$

It is intuitive clear (and can be proven with combinatorial methods), that both sides are equal to

$$\frac{1}{2} \cdot \frac{1}{2} \cdots \frac{1}{2} \;=\; \frac{1}{2^n}\,.$$

Thus, the throws are independent. □

Definition 5.17 (Independence of arbitrarily many random elements) applies to the most general kind of random elements. It turns out that the equations (5.48) need not be verified for all $A'_{i_j} \subseteq \Omega'$, if all random elements are discrete.

Fact 5.2. ⋆ *[Independence of discrete random elements]*

Assume that the random elements X_i of Definition 5.17 are discrete and that $\Omega'_* \subseteq \Omega'$ is countable and satisfies $P\{X_i \in \Omega'_*\} = 1$. Then it suffices to show that (5.48) is satisfied for events of the form $\{X_{i_j} = \omega'\}$, where $\omega' \in \Omega'_*$. In other words, it suffices to verify the following.

- *For ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ and $j = 1, 2, \ldots, k$,*

$$P\{X_{i_1} = \omega'_{i_1}, \ldots, X_{i_k} = \omega'_{i_k}\} = P\{X_{i_1} = \omega'_{i_1}\} \cdots P\{X_{i_k} = \omega'_{i_k}\} \tag{5.49}$$

is satisfied for all $\omega'_{i_j} \in \Omega'_*$.

From this general case, we obtain the case $I = 1, 2$ as follows.

Independence of two random elements, X_1, X_2: For all $\omega', \tilde{\omega}' \in \Omega'_*$,

$$P\{X_1 = \omega', X_2 = \tilde{\omega}'\} = P\{X_1 = \omega'\} \cdot P\{X_2 = \tilde{\omega}'\}\,. \tag{5.50}$$

For $I = 1, 2, 3$, we obtain

Independence of three random elements, X_1, X_2, X_3:

(1) *For all subselections $i_1 < i_2$ of $k = 2$ elements of $\{1, 2, 3\}$ (there are 3 such subselections) and for all $\omega'_{i_1}, \omega'_{i_2} \in \Omega'_*$,*

$$P\{X_{i_1} = \omega'_{i_1}, X_{i_2} = \omega'_{i_2}\} = P\{X_{i_1} = \omega'_{i_1}\} \cdot P\{X_{i_2} = \omega'_{i_2}\}\,, \tag{5.51}$$

(2) *For $k = 3$ (i.e., $i_1 = 1, i_2 = 2, i_1 = 3$) and for all $\omega'_1, \omega'_2, \omega'_3 \in \Omega'_*$,*

$$\begin{aligned} P\{X_1 = \omega'_1, X_2 =& \omega'_2, X_3 = \omega'_3\} \\ &= P\{X_1 = \omega'_1\} \cdot P\{X_2 = \omega'_2\} \cdot P\{X_3 = \omega'_3\}\,. \end{aligned} \tag{5.52}$$

For $I = 1, 2, \ldots, n$, we obtain

Independence of n random elements, $X_1, X_2, \ldots, X_n$:
For <u>EACH</u> $k = 2, 3, \ldots, n-1, n$, the following must be true: For all subselections $i_1 < \cdots < i_k$ of k elements of $\{1, \ldots, n\}$ and for all $\omega'_{i_j} \in \Omega'_*$, $(1 \le j \le k)$,

$$P\{X_{i_1} = \omega'_{i_1}, \ldots, X_{i_k} = \omega'_{i_k}\} = P\{X_{i_1} = \omega'_{i_1}\} \cdots P\{X_{i_k} = \omega'_{i_k}\}\,. \tag{5.53}$$

For $I = \mathbb{N}$, we obtain

Independence of an infinite sequence $X_1, X_2, \ldots,$ of random elements:
For <u>EACH</u> $k = 2, 3, 4, \ldots,$ the following must be true: For all subselections $i_1 < \cdots < i_k$ of k elements of $\mathbb{N}$ and for all $\omega'_{i_j} \in \Omega'_*$, $(1 \le j \le k)$,

$$P\{X_{i_1} = \omega'_{i_1}, \ldots, X_{i_k} = \omega'_{i_k}\} = P\{X_{i_1} = \omega'_{i_1}\} \cdots P\{X_{i_k} = \omega'_{i_k}\}\,. \tag{5.54}$$

Remark 5.21. ⋆

Note that the X_{i_j} in Fact 5.2 are distinct: If $j \neq m$ then $X_{i_j} \neq X_{i_m}$. In contrast, each ω'_{i_j} can take on

any value $\omega' \in \Omega'_*$

- For example, let $I = \mathbb{N}$ and X_i represents the ith roll of a die, and $\Omega' = \Omega'_* = \{1, 2, \ldots, 6\}$. Let $i_1 = 5, i_2 = 8, i_3 = 9, i_4 = 14$. Then one of the 6^4 equations (5.54) to be checked is for $\omega_{i_1} = 4, \omega_{i_2} = 1, \omega_{i_3} = 4, \omega_{i_4} = 4$.
- If we choose another selection of 4 indices, e.g., $i_1 = 8, i_2 = 13, i_3 = 89, i_4 = 1477$, Then another 6^4 equations (5.49) must be checked.

If that looks like bad news, it gets worse:

- For any $k \in \mathbb{N}$, there are, of course, infinitely many ways to pick integers $0 < i_1 < \cdots < i_k$. Thus, infinitely many equations (5.54) must be checked.

So, what is good for? The answer is as follows. It is often easier to prove, for general k and $0 < i_1 < \cdots < i_k$, that (5.54) holds true. An example follows this remark. □

Example 5.15. Let the random variables $Y_1, Y_2, \ldots$ denote an infinite sequence of rolls of a fair die. Is that an independent sequence of random variables?

Solution:

Before we start, let us agree that the answer to that question better be **yes**, since the outcome of the kth roll is in no way influenced by those of the other rolls.

What are the domain $(\Omega, \mathfrak{F}, P)$ and codomain $(\Omega', \mathfrak{F}')$ for the random variables Y_j? The obvious choice for the codomain is $\Omega' = \{1, 2, \ldots, 6\}$ and $\mathfrak{F}' = 2^{\Omega'}$. As usual, we leave $(\Omega, \mathfrak{F}, P)$ unspecified.

Note however, that we know the following about P: Let $y \in \Omega'$, i.e., $y \in \{1, \ldots, 6\}$. Since $P\{Y_j = y\}$ denotes the probability of $Y_j(\omega)$ resulting in the outcome y, it follows that

$$P\{Y_j = y\} \;=\; \frac{1}{6}\,. \tag{5.55}$$

These probabilities are the only ones that occur in (5.50), and we are able to work with that formula.

So let $k \in \mathbb{N}$ and $i_1 < i_2 < \cdots < i_k$ be an arbitrary selection of k indices.

- Since there are 6 possible outcomes y_{i_1} for Y_{i_1},
- and each of those can be combined with 6 possible outcomes y_{i_2} for Y_{i_2},
- and each of those combined outcomes can be combined with 6 possible outcomes y_{i_3} for Y_{i_3},
- .....
- and each of those combined outcomes can be combined with 6 possible outcomes y_{i_k} for Y_{i_k},

there are 6^k outcomes $\{Y_{i_1} = y_{i_1}, Y_{i_2} = y_{i_2}, \ldots, Y_{i_k} = y_{i_k}\}$, and each one of those is as likely to happen as any other. Thus,

$$P\{Y_{i_1} = y_{i_1}\, Y_{i_2} = y_{i_2}, \ \ldots, \ Y_{i_k} = y_{i_k}\} \;=\; \frac{1}{6^k}\,. \tag{5.56}$$

By (5.55), $P\{Y_{i_j} = y_{i_j}\} = 1/6$, for $j = 1, 2, \ldots, k$. Thus,

$$P\{Y_{i_1} = y_{i_1}\} \cdot P\{Y_{i_2} = y_{i_2}\} \cdots P\{Y_{i_k} = y_{i_k}\} \;=\; \frac{1}{6^k}\,. \tag{5.57}$$

Since (5.56) and (5.57) have matching right sides, (5.50) of Fact 5.2 is satisfied. This shows that $Y_1, Y_2, \ldots$ form an independent sequence of random variables. □

We noted in Fact 5.2 (Independence of discrete random elements) the following.

If the random elements are discrete, the condition $A'_{i_j} \subseteq \Omega'$ in (5.48) of Definition 5.17 (Independence of arbitrarily many random elements) only needs to be satisfied for specific A'_{i_j}:

$$A'_{i_j} \;=\; \{X_{i_j} = \omega'\}\,, \quad \text{where } \omega' \in \Omega' \text{ satisfies } \; P\{X_{i_j} = \omega'\} > 0.$$

An analogous situation exists if all random elements are random variables, except that the singletons $\{\omega'\} \subseteq \mathbb{R}$ will be replaced with intervals. (Recall that $\Omega' = \mathbb{R}$ for random variables!)

Fact 5.3. ⋆ *[Independence of random variables]*

Assume that the random elements X_i of Definition 5.17 are random variables. Then it suffices to show that (5.48) is satisfied for events of the form $\{ X_{i_j} \in]-\infty, \beta_{i_j}] \}$, for all $\beta_{i_j} \in \mathbb{R}$. In other words, it suffices to verify the following.

- *For ANY FINITE subselection of distinct indices $i_1, i_2, \dots, i_k \in I$ and $j = 1, 2, \dots, k$,*

$$P\{X_{i_1} \le \beta_{i_1}, \dots, X_{i_k} \le \beta_{i_k}\} \;=\; P\{X_{i_1} \le \beta_{i_1}\} \cdots P\{X_{i_k} \le \beta_{i_k}\}\,, \tag{5.58}$$

is satisfied for all $\beta_{i_j} \in \mathbb{R}$.

From this general case, we obtain the case $I = 1, 2$ as follows.

Independence of two random variables, Y_1, Y_2: For all $\beta_1, \beta_2 \in \mathbb{R}$,

$$P\{Y_1 \le \beta_1, Y_2 \le \beta_2\} \;=\; P\{Y_1 \le \beta_1\} \cdot P\{Y_2 \le \beta_2\}\,. \tag{5.59}$$

For $I = 1, 2, 3$, we obtain

Independence of three random variables, Y_1, Y_2, Y_3:

(1) *For all subselections $i_1 < i_2$ of $k = 2$ elements of $\{1, 2, 3\}$ (there are 3 such subselections) and for all $\beta_{i_1}, \beta_{i_2} \in \mathbb{R}$,*

$$P\{Y_{i_1} \le \beta_{i_1}, Y_{i_2} \le \beta_{i_2}\} \;=\; P\{Y_{i_1} \le \beta_{i_1}\} \cdot P\{Y_{i_2} \le \beta_{i_2}\}\,, \tag{5.60}$$

(2) *For $k = 3$ (i.e., $i_1 = 1, i_2 = 2, i_1 = 3$) and for all $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$,*

$$\begin{aligned} &P\{Y_1 \le \beta_1,\, Y_2 \le \beta_2,\, Y_3 \le \beta_3\} \\ &\qquad = P\{Y_1 \le \beta_1\} \cdot P\{Y_2 \le \beta_2\} \cdot P\{Y_3 \le \beta_3\}\,. \end{aligned} \tag{5.61}$$

For $I = 1, 2, \dots, n$, we obtain

Independence of n random variables, $Y_1, Y_2, \dots, Y_n$:
For <u>EACH</u> $k = 2, 3, \dots, n-1, n$, the following must be true: For all subselections $i_1 < \cdots < i_k$ of k elements of $\{1, \dots, n\}$ and for all $\beta_{i_j} \in \mathbb{R}$, $(1 \le j \le k)$,

$$P\{Y_{i_1} \le \beta_{i_1}, \dots, Y_{i_k} \le \beta_{i_k}\} \;=\; P\{Y_{i_1} \le \beta_{i_1}\} \cdots P\{Y_{i_k} \le \beta_{i_k}\}\,. \tag{5.62}$$

For $I = \mathbb{N}$, we obtain

Independence of an infinite sequence $Y_1, Y_2, \dots,$ of random variables:
For <u>EACH</u> $k = 2, 3, 4, \dots,$ the following must be true: For all subselections $i_1 < \cdots < i_k$ of k elements of $\mathbb{N}$ and for all $\beta_{i_j} \in \mathbb{R}$, $(1 \le j \le k)$,

$$P\{Y_{i_1} \le \beta_{i_1}, \dots, Y_{i_k} \le \beta_{i_k}\} \;=\; P\{Y_{i_1} \le \beta_{i_1}\} \cdots P\{Y_{i_k} \le \beta_{i_k}\}\,. \tag{5.63}$$

The next definition is extremely important, since the notion of a "random sample" uses it.

We give a special name to collections of random elements that are independent and also share the same probability distribution.

Definition 5.18 (iid families)**.**

Let $(X_i)_{i\in I}$ be a family of random elements $X_i : (\Omega, P) \to \Omega'$. We speak of an **independent and identically distributed family**, aka **iid family** of random elements, if

(1) the X_i are independent,

(2) they all have the same distribution:

$$P_{X_i}(B) \;=\; P_{X_j}(B)\,, \qquad \text{for all } \; i,j \in I \;\text{ and all }\; B \subseteq \Omega'.$$

Note that this can also be written

$$P\{X_i \in B\} \;=\; P\{X_j \in B\}\,, \qquad \text{for all } \; i,j \in I \;\text{ and all }\; B \subseteq \Omega'.$$

In the special case of a sequence $X_1, X_2, \ldots$ of iid random elements we speak of an **iid sequence** of random elements. □

6 Advanced Topics – Measure and Probability ⋆

6.1 Random Variables as Measurable Functions

Introduction 6.1. The definition of the distribution P_X of a random element $X : (\Omega, P) \to \Omega'$ was based on Theorem 5.11. (See p.123). It asserts that the formula

$$A' \mapsto P_X(A') \;=\; P\{X \in A'\} \;=\; P\big[\,X^{-1}(A')\,\big]\,, \quad A' \subseteq \Omega', \tag{6.1}$$

defines a probability measure on all subsets of Ω'.

A. This is fine for the applications, since we decided to ignore σ–algebras whenever possible. We can do so when practical applications are involved, since we deal with one of the following two situations when doing computations (see Remark 5.7 on p.112):

- **(1)** Either P_X discrete and there will be no issues with defining $P_X(A')$ for all $A' \in \Omega'$; in this case we work with $\mathfrak{F}' = 2^{\Omega'}$.
- **(2)** Or X is a random variable or random vector, i.e., $\Omega' = \mathbb{R}$ or $\Omega' = \mathbb{R}^d$. Then practical computations always involve sets $B' \in \mathbb{R}^d$ for which the Riemann integral $\int\!\!\int \cdots \int_{B'} d\vec{x}$ exists. Since this necessitates that B' is Borel, we can work with $\mathfrak{F}' = \mathfrak{B}^d$. As we previously mentioned, only very strange and nonsensical subsets of $\mathbb{R}^d$ are not Borel and we may as well act as if $P_X(B')$ exists for all $B' \in \mathbb{R}^d$.

In summary, it is OK to assume for the applications that the domain of $A' \mapsto P(A')$ is $2^{\Omega'}$, the entire power set of Ω'. There is no need to restrict ourselves to a potentially smaller σ–algebra $\mathfrak{F}' \subseteq 2^{\Omega'}$.

B. All that having been said, let us now consider the mathematical aspects of probability theory. What if we cannot make the assumption that all sets are events, i.e., can be assigned a probability? We have to consider this issue for both the domain Ω and the codomain Ω' of the random element X. Accordingly, we need a σ–algebra $\mathfrak{F}$ as the domain of $A \mapsto P(A)$, and another σ–algebra $\mathfrak{F}'$ as the domain of $A' \mapsto P_X(A')$. In other words, we have to consider X as a function

$$X : (\Omega, \mathfrak{F}, P) \longrightarrow (\Omega', \mathfrak{F}', P_X)\,.$$

Of course, some conditions to ensure that $\mathfrak{F}, P, \mathfrak{F}', X$ compatible may have to be imposed. (The distribution P_X does not play a part here, since it is completely determined by the other four items.)

To understand what such conditions should be, consider the following.

- $\mathfrak{F}$ is given, as part of the original probability space $(\Omega, \mathfrak{F}, P)$.
- As we saw in **(1)** and **(2)**, there also is not much leeway as far as $\mathfrak{F}'$ is concerned.

So what conditions must X satisfy that it has a distribution P_X, defined at least on $\mathfrak{F}'$?

- The distribution of X is given by $P_X(A') \;=\; P\{X \in A'\} \;=\; P\big(f^{-1}(A')\big)$. (See (6.1).) Accordigly, the answer is as follows: $P\{X \in A'\}$ must exist at least for all $A' \in \mathfrak{F}'$
- That can only happen if X satisfies $\boxed{\{X \in A'\} \;=\; X^{-1}(A') \in \mathfrak{F}, \text{ whenever } A' \in \mathfrak{F}'}$ □

Based on those introductory remarks, we introduce the concept of measurability. Since it has per se nothing to do with probabilities, we switch the function symbol to f.

Definition 6.1 (Measurable functions). ⋆

(a) Let Ω be a nonempty set and $\mathfrak{F}$ a σ–algebra on Ω. We call the pair $(\Omega, \mathfrak{F})$ a **measurable space**. (This is not worthwhile remembering, but the remainder of this definition is.)

(b) Let $f : (\Omega, \mathfrak{F}) \longrightarrow (\Omega', \mathfrak{F}')$ be a function which has measurable spaces both as domain and codomain. We call this function **measurable with respect to $\mathfrak{F}$ and $\mathfrak{F}'$**, a.k.a. $(\mathfrak{F}, \mathfrak{F}')$**–measurable**, if

$$A' \in \mathfrak{F}' \;\Rightarrow\; f^{-1}(A') \in \mathfrak{F}\,. \tag{6.2}$$

(c) If f is $\mathbb{R}^d$–valued, in particular if f is real–valued, and if we refer to f as being **$\mathfrak{F}$–measurable** or **Borel measurable**, then it is implied that $\mathfrak{F}' = \mathfrak{B}^d$, the Borel σ–algebra of $\mathbb{R}^d$. □

Next comes a straight translation of Definition 4.3 (Simple Function on $\mathbb{R}^d$) on p.87.

Definition 6.2 (Simple Function on Ω). ⋆

Let $(\Omega, \mathfrak{F})$ be a measurable space, $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathfrak{F}$. Further, let $c_1, c_2, \ldots, c_n$ be a corresponding set of real numbers. Let

$$f : \Omega \longrightarrow \mathbb{R}\,; \qquad \omega \mapsto f(\omega) \;:=\; \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\omega) \tag{6.3}$$

Then we call f a **simple function**. We say that f is in **standard form**, if the numhers c_j are distinct, i.e., $c_i \neq c_j$, for $i \neq j$. □

Remark 6.1. ⋆ Assume that f is a simple function $f = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}$ on $(\Omega, \mathfrak{F})$.

(a) It is not assumed that the numbers c_j are distinct. In that case, $\{f = c_j\} = f^{-1}\{c_j\}$ does not equal A_j. Assume for example, that $c_3 = c_5 = c_{16}$ and that all other c_j are different from c_3. Then $\{f = c_3\} = A_3 \uplus A_3 \uplus A_5 \uplus A_{16}$. Note that, since $A_k \in \mathfrak{F}$ for all k, $\{f = c_j\} \in \mathfrak{F}$.

(b) Let $B \in \mathfrak{B}^1$. Let $J := \{j : c_j \in B\}$ be the set of all indices j such that $c_j \in B$. Then

$$f^{-1}(B) \;=\; \biguplus_{j \in J} A_j\,. \qquad \text{Thus,}\;\; f^{-1}(B) \in \mathfrak{F},\;\; \text{for all } B \in \mathfrak{B}^1.$$

In other words, all simple functions are $(\mathfrak{F}, \mathfrak{B}^1)$–measurable. □

Proposition 6.1. ⋆

Let $f = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}$ be a simple function. Then f has a representation in standard form. This standard representation is

$$f(\omega) \;=\; \sum_{i=1}^{k} d_i \mathbf{1}_{\{f=d_i\}}(\omega)\,, \quad \textit{with distinct numbers}\;\; d_1, \ldots, d_k. \tag{6.4}$$

PROOF: Since there are only finitely many terms $c_j \mathbf{1}_{A_j}$, the range of f [69] is a finite list,

$$f(\Omega) \;=\; \{d_1, \ldots, d_k\}\,. \quad \text{of distinct numbers } \; d_1, \ldots, d_k.$$

Regardless of the nature of the c_j and A_j, $f(\omega) = \alpha \;\Leftrightarrow\; \omega \in f^{-1}\{\alpha\}$, for any $\alpha \in \mathbb{R}$. Thus,

$$f(\omega) \;=\; d_i \;\Leftrightarrow\; \omega \in f^{-1}\{d_i\} \;\Leftrightarrow\; f(\omega) \;=\; d_i \cdot \mathbf{1}_{f^{-1}\{d_i\}}\,.$$

Since the d_i are distinct, the sets $f^{-1}\{d_i\}$ are disjoint. Thus,

$$f(\omega) \;=\; d_i \;\Rightarrow\; d_i \cdot \mathbf{1}_{f^{-1}\{d_i\}}(\omega) \;=\; \sum_{m=1}^{k} d_m \cdot \mathbf{1}_{f^{-1}\{d_m\}}(\omega)\,.$$

Since the right–hand side does not depend on i and each $\omega \in \Omega$ satisfies $f(\omega) = d_i$ for some $i = 1, \ldots, k$, we conclude that $f(\omega) \;=\; \sum_{m=1}^{k} d_m \cdot \mathbf{1}_{f^{-1}\{d_m\}}(\omega)$, for all $\omega \in \Omega$. ■

The next theorem is a straight translation of Theorem 6.1 on p.139.

It asserts that about anything that can be done with a countable collection of real–valued, Borel measurable functions results again in a Borel measurable function. Note that we have suppressed the arguments in the functions listed there. For example, $\max(f_1, f_2)$ is the function $\omega \mapsto \max(f_1(\omega),\, f_2(\omega))$, and $\sum_{j=1}^{\infty} f_j$ is the function $\omega \mapsto \sum_{j=1}^{\infty} f_j(\omega)$.

Theorem 6.1. ⋆ *Assume that $f_1, f_2, \ldots$ are Borel measurable functions, $c_1, c_2, \cdots \in \mathbb{R}$, $B \in \mathfrak{F}$.*

Each of the following also is a Borel measurable function:

- c_1 *(constant function)* • $c_1 f_1$ • $f_1 \pm f_2$ • $f_1 f_2$ • $\mathbf{1}_B f_1$ • f_1/f_2 *(if* $f_2 \neq 0$*)* • $\sum_{j=1}^{n} c_j f_j$
- $\min(f_1, f_2)$ • $\max(f_1, f_2)$ • $\min_{j=1,\ldots,n} f_j$ • $\max_{j=1,\ldots,n} f_j$ • $\inf_{j\in\mathbb{N}} f_j$ • $\sup_{j\in\mathbb{N}} f_j$ □

If they exist (see the subsequent remark), the following also are measurable functions:

- $\lim_{j\to\infty} f_j$ • $\sum_{j=1}^{\infty} f_j$ • $\min_{j\in\mathbb{N}} f_j$ • $\max_{j\in\mathbb{N}} f_j$

PROOF: ■

The next theorem will be key in later extending integration with respect to Lebesgue measure to a class of measures so general that it includes all probability measures.

Theorem 6.2. ⋆ *Let $(\Omega, \mathfrak{F})$ be a measurable space.*

Let $f : (\Omega, \mathfrak{F}) \longrightarrow [0, \infty[$ be a nonnegative, $(\mathfrak{F}, \mathfrak{B}^1)$–measurable function. Then there exists a sequence $0 \leq f_1 \leq f_2 \leq \cdots$ of simple functions such that $f_n \uparrow f$ as $n \to \infty$. In other words,

$$\lim_{n\to\infty} f_n(\omega) \;=\; f(\omega), \quad \text{for all } \omega \in \Omega.$$

[69] See Definition 2.17 (Function) on p.33.

PROOF: This proof can be found in greater detail in Remark 4.5 on p.89.
Fix $n \in \mathbb{N}$, and define, for $k \in \mathbb{N}$,

$$I_{k,n} := \left] \frac{k-1}{2^n}, \frac{k}{2^n} \right] .$$

Note that $[0,\infty[= \{0\} \uplus \left(\biguplus [I_{k,n} : k \in \mathbb{Z}] \right)$ partitions the codomain into small intervals. Let

$$A_{k,n} := \left\{ \omega \in \Omega : \frac{k-1}{2^n} < f(\omega) \leq \frac{k}{2^n} \right\} \quad (k = 1, \ldots, 4^n),$$

Note that $\omega \in A_{k,n} \Leftrightarrow (k-1)/2^n < f(\omega) \leq k/2^n$. Next, we define

$$f_n(\omega) := \sum_{k=1}^{4^n} \frac{k-1}{2^n} \cdot \mathbf{1}_{A_{k,n}}(\omega). \tag{6.5}$$

Remark 4.5 contains a picture which demonstrates how the simple functions $f_n \uparrow f$ are constructed. Observe that

$$f_n(\omega) = \frac{k-1}{2^n} \text{ on } A_{k,n} = \left\{ \omega \in \Omega : \frac{k-1}{2^n} < f(\omega) \leq \frac{k}{2^n} \right\} .$$

Further,

$$0 \leq f(\omega) - f_n(\omega) \leq \frac{1}{2^n}, \quad \text{for } \omega \in A_{k,n}.$$

Let $A_0 := \{\omega \in \Omega : f(\omega) = 0\}$. Since $f \geq 0$, (6.5) implies that $f_n(\omega) = f(\omega) = 0$ on A_0, we see that

$$0 \leq f(\omega) - f_n(\omega) \leq \frac{1}{2^n}, \quad \text{for} \quad \omega \in A_0 \cup A_{1,n} \cup A_{2,n} \cup \cdots \cup A_{4^n,n} .$$

Since $1 \leq k \leq 4^n$ is equivalent to $0 \leq (k-1)/2^n < k/2^n \leq 4^n/2^n = 2^n$, we obtain

$$0 \leq f(\omega) - f_n(\omega) \leq \frac{1}{2^n}, \quad \text{for} \quad f(\omega) \leq 2^n .$$

Finally, since $f(\omega) < \infty$ for all $\omega \in \Omega$ and $2^{-n} \to 0$ and $2^n \to \infty$ as $n \to \infty$, we conclude that

$$f_n(\omega) \uparrow f(\omega), \quad \text{for} \quad \omega \in \Omega . \blacksquare$$

Measurability will only be useful if it helps to construct distributions on $(\Omega', \mathfrak{F}'$. The next theorem, a modified version of Theorem 5.11 on p.123, shows that such is the case.

Theorem 6.3. ⋆

Let $(\Omega, \mathfrak{F}, P)$ be a probability space, $(\Omega', \mathfrak{F}')$ a measurable space, and

$$X : (\Omega, \mathfrak{F}, P) \longrightarrow (\Omega', \mathfrak{F}', P_X) .$$

an $(\mathfrak{F}, \mathfrak{F}')$–measurable function. Then the formula

$$P_X(A') := P\{X \in A'\} \quad (A' \in \mathfrak{F}') \tag{6.6}$$

defines a probability measure on $\mathfrak{F}'$.

PROOF: Very similar to that of Theorem 5.11. The fact that $\mathfrak{F}$ and $\mathfrak{F}'$ are σ–algebras guarantees that all sets A in the proof that need $P(A)$ defined are indeed elements of $\mathfrak{F}$, and all sets A' in the proof that need $P_X(A')$ defined are indeed elements of $\mathfrak{F}'$.

First, we establish that all probabilities required exist, i.e., that $\mathfrak{F}'$ contains all relevant subsets of Ω'. Assume fo the following that $'A, B', A'_1, A'_2, \cdots \in \mathfrak{F}'$, with A_j pairwise disjoint and $B' = \biguplus_j A'_j$.

- Since $\mathfrak{F}'$ is a σ–algebra, $\emptyset \in \mathfrak{F}'$ and $\Omega' \in \mathfrak{F}'$. Thus, both $P_X(\emptyset)$ and $P_X(\Omega')$ exist.
- Since $\mathfrak{F}'$ is a σ–algebra, $A'^{\complement} \in \mathfrak{F}'$. Thus, $P_X\left(A'^{\complement}\right)$ exists.
- Since $\mathfrak{F}'$ is a σ–algebra, $B' \in \mathfrak{F}'$. Thus, $P_X(B')$ exists.

The remainder of the proof is word for word the same as that of Theorem 5.11.
It follows from $\{X \in \emptyset\} = \emptyset$ and $\{X \in \Omega'\} = \Omega$, that

$$P_X(\emptyset) \;=\; P(\emptyset) \;=\; 0 \qquad \text{and} \qquad P_X(\Omega') \;=\; P(\Omega) \;=\; 1\,.$$

From (2.46) on p.45, we obtain

$$P_X\big(A'^{\complement}\big) \;=\; P\{X \in A'^{\complement}\} \;=\; P\big(X^{-1}(A'^{\complement})\big) \;=\; P\big([X^{-1}(A')]^{\complement}\big) \;=\; 1 \,-\, P\big(X^{-1}(A')\big) \;=\; 1 \,-\, P_X(A')\,.$$

We apply (2.45) to the sequence of disjoint subsets $A'_1, A'_2, \ldots$ of Ω' and obtain

$$P_X(B') \;=\; P\big(X^{-1}\left(\biguplus_{j\in\mathbb{N}} A'_j\right)\big) \;=\; P\left(\bigcup_{j\in\mathbb{N}} X^{-1}(A'_j)\right)$$

By (2.47), the sets $X^{-1}(A'_j)$ are disjoint. σ–additivity now follows from

$$P_X(B') \;=\; P\left(\biguplus_{j\in\mathbb{N}} X^{-1}(A'_j)\right) \;=\; \sum_{j\in\mathbb{N}} P\big(X^{-1}(A'_j)\big) \;=\; \sum_{j\in\mathbb{N}} P_X(A'_j)\,. \;\blacksquare$$

Remark 6.2. ⋆

(a) Note that every probability space $(\Omega, \mathfrak{F}, P)$ is a measurable space.

(b) A key property of random elements X, in particular random variables and random vectors, is that they induce a distribution $P_X(A') = P\{X \in A'\}$ on the codomain. This is so essential, that measurability becomes part of the definition of a random element in basically all graduate level texts on probability, since there one does not gloss over the role of σ–algebras and measurability like we do in this course. □

We amend the definition of random elements accordingly.

Definition 6.3 (Advanced level definition of random variables and random elements)**.** ⋆

Only for the remainder of this chapter 6.1 (Advanced Topics – Measurable Functions), we modify Definitions 5.14 on p.124 and 5.15 on p.125 as follows.

Given are a probability space $(\Omega, \mathfrak{F}, P)$, a measurable space $(\Omega', \mathfrak{F}')$, $d \in \mathbb{N}$, and $(\mathfrak{F}, \mathfrak{F}'$–measurable

$$X : (\Omega, \mathfrak{F}, P) \longrightarrow (\Omega', \mathfrak{F}') .$$

(a) We call X a **random element**.
(b) If $(\Omega', \mathfrak{F}') = (\mathbb{R}, \mathfrak{B}))$, we also call X a **random variable**.
(c) If $(\Omega', \mathfrak{F}') = (\mathbb{R}^d, \mathfrak{B}^d))$, we also call X a **random vector**. □

The material covered so far in this section should leave no doubt that understanding how preimages relate to σ–algebras is very impportant. We now turn our attention to σ–algebras.

Remark 6.3. ⋆

We repeat here some findings and definitions from earlier chapters.

(a) Given are a function $f : \Omega \to \Omega'$ and an arbitrary family of sets $A'_i \subset \Omega', i \in I$. By Proposition 2.7 (p.44) and Theorem 2.2 (45), the preimages $\big(f^{-1}(B_i)\big)_{i\in I}$ satisfy the following.
⊡ $f^{-1}(\emptyset) = \emptyset$ ⊡ $f^{-1}(\Omega') = \Omega$ ⊡ $B_i \subseteq B_j \;\Rightarrow\; f^{-1}(B_i) \subseteq f^{-1}(B_j)$
⊡ $f^{-1}(\bigcap\limits_{i\in I} B_j) = \bigcap\limits_{j\in J} f^{-1}(B_j)$ ⊡ $f^{-1}(\bigcup\limits_{i\in I} B_j) = \bigcup\limits_{j\in J} f^{-1}(B_j)$ ⊡ $f^{-1}(B_i^{\complement}) = \big(f^{-1}(B_i)\big)^{\complement}$
⊡ $B_i \cap B_j = \emptyset \;\Rightarrow\; f^{-1}(B_i) \cap f^{-1}(B_j) = \emptyset$; thus, $f^{-1}(\text{ partition of } \Omega') = \text{ partition of } \Omega$

(b) The intersection of an arbitrary collection of σ–algebras is a σ–algebra. (Theorem 5.3 on p.113)

(c) Let $\boldsymbol{\mathcal{A}} \subseteq 2^{\Omega}$. By **(b)**, $\sigma\{\boldsymbol{\mathcal{A}}\} = \bigcap \big[\, \widetilde{\mathfrak{F}} : \widetilde{\mathfrak{F}} \supseteq \boldsymbol{\mathcal{A}}$ and $\widetilde{\mathfrak{F}}$ is a σ–algebra $\big]$ is a σ–algebra: the smallest one that contains $\boldsymbol{\mathcal{A}}$. It is the σ–algebra generated by $\boldsymbol{\mathcal{A}}$. See Definition 5.5 on p.114

(d) $\sigma\{\mathfrak{B}^d\} = \sigma\{$ all d–dimensional rectangles $\}$ constitutes the Borel sets of $\mathbb{R}^d$. □

Theorem 6.4. ⋆

Let Ω be a nonempty set and let $\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ be three collections of subsets of Ω. Then

$$\boldsymbol{\mathcal{E}}_1 \subseteq \boldsymbol{\mathcal{E}}_2 \quad\Rightarrow\quad \sigma(\boldsymbol{\mathcal{E}}_1) \subseteq \sigma(\boldsymbol{\mathcal{E}}_2) , \tag{6.7}$$

$$\sigma\big(\sigma(\boldsymbol{\mathcal{E}})\big) = \sigma(\boldsymbol{\mathcal{E}}) , \tag{6.8}$$

$$\sigma(\boldsymbol{\mathcal{E}}_1) \supseteq \boldsymbol{\mathcal{E}}_2 \;\text{ and }\; \sigma(\boldsymbol{\mathcal{E}}_2) \supseteq \boldsymbol{\mathcal{E}}_1 \quad\Rightarrow\quad \sigma(\boldsymbol{\mathcal{E}}_1) = \sigma(\boldsymbol{\mathcal{E}}_2) . \tag{6.9}$$

PROOF of (6.7):

Any σ–algebra $\mathfrak{G}$ that contains $\boldsymbol{\mathcal{E}}_2$ also contains $\boldsymbol{\mathcal{E}}_1$. Thus more sets are intersected in

$$\sigma(\boldsymbol{\mathcal{E}}_1) = \bigcap\{\mathfrak{G} : \mathfrak{G} \supseteq \boldsymbol{\mathcal{E}}_1 \text{ and } \mathfrak{G} \text{ is a } \sigma\text{–algebra for } \Omega\}.$$

than in

$$\sigma(\boldsymbol{\mathcal{E}}_2) = \bigcap\{\mathfrak{G} : \mathfrak{G} \supseteq \boldsymbol{\mathcal{E}}_2 \text{ and } \mathfrak{G} \text{ is a } \sigma\text{–algebra for } \Omega\}.$$

It follows that $\sigma(\boldsymbol{\mathcal{E}}_1) \subseteq \sigma(\boldsymbol{\mathcal{E}}_2)$.

PROOF of (6.8):

Since $\boldsymbol{\mathcal{E}} \subseteq \sigma(\boldsymbol{\mathcal{E}})$, we obtain from (6.7) (already proven) that $\sigma(\boldsymbol{\mathcal{E}}) \subseteq \sigma\big(\sigma(\boldsymbol{\mathcal{E}})\big)$. Now we show "$\supseteq$":

(1) Let $\mathscr{D} := \{\widetilde{\mathfrak{F}} : \widetilde{\mathfrak{F}} \supseteq \sigma(\boldsymbol{\mathcal{E}})$ and $\widetilde{\mathfrak{F}}$ is a σ–algebra $\}$. Note that $\sigma(\boldsymbol{\mathcal{E}}) \in \mathscr{D}$. Thus, $\{\sigma(\boldsymbol{\mathcal{E}})\} \subseteq \mathscr{D}$.
(2) It is true in general that $\mathscr{U}_1 \subseteq \mathscr{U}_2 \;\Rightarrow\; \bigcap[U : U \in \mathscr{U}_1] \;\supseteq\; \bigcap[U : U \in \mathscr{U}_2]$.
(3) By **(1)**, $\mathscr{U}_1 := \{\sigma(\boldsymbol{\mathcal{E}})\} \subseteq \mathscr{U}_2 =: \mathscr{D}$.
Thus, by **(2)**, $\sigma(\boldsymbol{\mathcal{E}}) = \bigcap[U : U \in \mathscr{U}_1] \supseteq \bigcap[U : U \in \mathscr{U}_2] = \sigma\big(\sigma(\boldsymbol{\mathcal{E}})\big)$.

Thus, $\sigma(\boldsymbol{\mathcal{E}}) \supseteq \sigma\big(\sigma(\boldsymbol{\mathcal{E}})\big)$. That concludes the proof of (6.8).

PROOF of (6.9):

(1) $\sigma(\boldsymbol{\mathcal{E}}_1) \supseteq \boldsymbol{\mathcal{E}}_2 \overset{(6.7)}{\Rightarrow} \sigma\big(\sigma(\boldsymbol{\mathcal{E}}_1)\big) \supseteq \sigma(\boldsymbol{\mathcal{E}}_2) \overset{(6.8)}{\Rightarrow} \sigma(\boldsymbol{\mathcal{E}}_1) \supseteq \sigma(\boldsymbol{\mathcal{E}}_2)$.
(2) $\sigma(\boldsymbol{\mathcal{E}}_2) \supseteq \boldsymbol{\mathcal{E}}_1 \overset{(6.7)}{\Rightarrow} \sigma\big(\sigma(\boldsymbol{\mathcal{E}}_2)\big) \supseteq \sigma(\boldsymbol{\mathcal{E}}_1) \overset{(6.8)}{\Rightarrow} \sigma(\boldsymbol{\mathcal{E}}_2) \supseteq \sigma(\boldsymbol{\mathcal{E}}_1)$.

It follows from **(1)** and **(2)** that $\sigma(\boldsymbol{\mathcal{E}}_1) = \sigma(\boldsymbol{\mathcal{E}}_2)$. ■

We now use that last theorem to prove some of the assertions made in Remark 5.8 on p.114.

Example 6.1. Consider the following subsets of the real numbers.

$$\begin{aligned}
&\mathfrak{B} = \{\text{ the Borel sets of } \mathbb{R}\} = \sigma\{\text{ all intervals of } \mathbb{R}\},\\
&\mathfrak{I}_1 := \{\,]a,b] : a<b\,\}, \quad \mathfrak{I}_2 := \{\,[a,b] : a<b\,\},\\
&\mathfrak{I}_3 := \{\,]a,b[: a<b\,\}, \quad \mathfrak{I}_4 := \{\,[a,b[: a<b\,\}, \quad \boldsymbol{\mathcal{E}} := \{\,]-\infty,c] : c\in\mathbb{R}\,\}.
\end{aligned}$$

Then $\mathfrak{B} = \sigma(\mathfrak{I}_1) = \sigma(\mathfrak{I}_2) = \sigma(\mathfrak{I}_3) = \sigma(\mathfrak{I}_4) = \sigma(\boldsymbol{\mathcal{E}})$.
For example, to prove that $\mathfrak{I}_2 = \mathfrak{I}_3$, it suffices according to Theorem 6.4 to show that

$$\text{any closed interval } [a,b] \text{ belongs to } \mathfrak{I}_3, \quad \text{any open interval }]a,b[\text{ belongs to } \mathfrak{I}_2.$$

Since $\sigma(\mathfrak{I}_3)$ contains all countable intersections of sequences in $\mathfrak{I}_3$ and $\sigma(\mathfrak{I}_2)$ contains all countable unions of sequences in $\mathfrak{I}_2$, this follows from

$$[a,b] = \bigcap_n \Big]a-\frac{1}{n}, b+\frac{1}{n}\Big[\quad \text{and} \quad]a,b[= \bigcup_n \Big[a+\frac{1}{n}, b-\frac{1}{n}\Big].$$

As another example, we show that $\sigma(\mathfrak{B}) = \sigma(\boldsymbol{\mathcal{E}})$. Note that

$$]a,b] =]-\infty,b\,] \cap]-\infty,a\,], \quad (a,b\in\mathbb{R}).$$

Thus, $\mathfrak{I}_1 \subseteq \sigma(\boldsymbol{\mathcal{E}})$. Since also $\boldsymbol{\mathcal{E}} \subseteq \mathfrak{I}_1$, it follows that $\mathfrak{I}_1 = \sigma(\boldsymbol{\mathcal{E}})$. □

In Definition 5.16 (σ–algebra generated by random elements) on p.125, we discussed the σ–algebras

$$\begin{aligned}
\sigma\{X\} &= \sigma\{X^{-1}(A') : A' \subseteq \Omega'\},\\
\sigma\big\{(X_i)_{i\in I}\big\} &= \sigma\{X_i : i\in I\} = \sigma\{X_i^{-1}(A') : A'\subseteq\Omega',\ i\in I\},
\end{aligned}$$

for random elements X and families of random elements $(X_i)_{i\in I}$ with domain (Ω, P) and codomain Ω'. See (5.36) and (5.37). This was done without taking into account the role of σ–algebras on Ω and Ω'. Note that P does not appear in the formulas that define those σ–algebras. Accordingly, probability measures will not appear in the replacement definition that follows.

Definition 6.4 (Advanced Definition of σ–algebra generated by random elements)**.** ⋆

We define for a function f and a family of functions $(f_i)i \in I$,

$$f, f_i : \Omega \longrightarrow (\Omega', \mathfrak{F}'),\ i \in I\ :$$

(6.10) $$\sigma\{f\} \ := \ \sigma\{f^{-1}(A') : A' \in \mathfrak{F}'\}$$

(6.11) $$\sigma\big\{(f_i)_{i\in I}\big\} \ := \ \sigma\{f_i : i \in I\} \ := \ \sigma\{f_i^{-1}(A') : A' \in \mathfrak{F}',\ i \in I\}$$

(a) We call $\sigma\{f\}$ the **σ–algebra generated by the function f.**
(b) We call $\sigma\big\{(f_i)_{i\in I}\big\}$ the **σ–algebra generated by the family of functions $(f_i)_{i\in I}$.** □

Remark 6.4. ⋆

(a) No assumption was made about $(\mathfrak{F}, \mathfrak{F}')$–measurability for obvious reasons. After all, there may not even be a σ–algebra $\mathfrak{F}$.
(b) We cannot call those functions random elements, because there is no probability measure.
(c) Note that the only difference between (5.36) and (5.37) on the one hand, and (6.10) and (6.11) on the other hand, is as follows: We have replaced $A' \subseteq \Omega'$ with $A' \in \mathfrak{F}'$. Hence, both definitions are identical if $\mathfrak{F}'$ denotes $2^{\Omega'}$, the biggest σ–algebra that exists on Ω'.
(d) If $\mathfrak{F}'$ is very small, then $\sigma\{f\}$ and $\sigma\big\{(f_i)_{i\in I}\big\}$ also might be very small.
In the extreme case, consider $\mathfrak{F}' := \{\emptyset, \Omega'\}$, the smallest σ–algebra for Ω'. Since $g^{-1}(\emptyset) = \emptyset$ and $g^{-1}(\Omega') = \Omega$ for ALL functions $g : \Omega \to \Omega'$, we obtain

$$\sigma\{f\} \ = \ \sigma\big\{(f_i)_{i\in I}\big\} \ = \ \{\emptyset, \Omega\}\,. \ \square$$

The next theorem shows that formula (6.10) (the definition of the σ–algebra generated by a single function) simplifies to

$$\sigma\{f\} \ = \ \{f^{-1}(A') : A' \in \mathfrak{F}'\}\,.$$

Theorem 6.5. ⋆

Given is a function f with measurable spaces $(\Omega, \mathfrak{F})$ as domain and $(\Omega', \mathfrak{F}')$ as codomain:

$$f : (\Omega, \mathfrak{F}) \longrightarrow (\Omega', \mathfrak{F}')\,.$$

No assumption is made about $(\mathfrak{F}, \mathfrak{F}')$–measurability. Then
(a) *$\sigma\{f\} \ = \ \{f^{-1}(A') : A' \in \mathfrak{F}'\}$. In particular, $\{f^{-1}(A') : A' \in \mathfrak{F}'\}$is a σ–algebra for Ω.*
(b) *f is $(\mathfrak{F}, \mathfrak{F}')$–measurable $\Leftrightarrow$ $\sigma\{f\} \subseteq \mathfrak{F}$.*
(c) *We can strengthen assertion **(b)** as follows: Let $\boldsymbol{\mathcal{E}}'$ be a generator of $\mathfrak{F}'$. Then*
f is $(\mathfrak{F}, \mathfrak{F}')$–measurable $\Leftrightarrow$ $\{f^{-1}(E') : E' \in \boldsymbol{\mathcal{E}}'\} \subseteq \mathfrak{F}$.

PROOF: We write $\mathscr{A}$ for the set $\{f^{-1}(A') : A' \in \mathfrak{F}'\}$.
PROOF of **(a)**: The assertions of Remark 6.3**(a)** on p.142 show that $\mathscr{A}$ is a σ–algebra.
It follows that $\mathscr{A} = \sigma\{\mathscr{A}\}$. Moreover, $\sigma\{f\} = \sigma\{\mathscr{A}\}$, by definition of $\sigma\{\dots\}$. Thus, $\sigma\{f\} = \mathscr{A}$.
PROOF of **(b)**: This follows from the definition of measurable functions.
PROOF of **(c)**, "⇒": Assume that f is measurable. Then $f^{-1}(A') \in \mathfrak{F}$, for all $A' \in \mathfrak{F}'$.
Since $\boldsymbol{\mathcal{E}}' \subseteq \mathfrak{F}'$, it follows that $f^{-1}(E') \in \mathfrak{F}$, for all $E' \in \boldsymbol{\mathcal{E}}'$.

PROOF of **(c)**, "⇐": This is not as easy as the "⇐" direction. First, we show that

$$\mathcal{G}' := \{G' \subseteq \Omega' : f^{-1}(G') \in \mathfrak{F}\} \tag{6.12}$$

is a σ–algebra for Ω'. We only show $G' \in \mathcal{G}' \Rightarrow G'^{\complement} \in \mathcal{G}'$ and leave the remainder as an exercise.
So let $G' \in \mathcal{G}'$. By (6.12) (the definition of $\mathcal{G}'$), $f^{-1}(G') \in \mathfrak{F}$. Since $\mathfrak{F}$ is a σ–algebra, $\big(f^{-1}(G')\big)^{\complement} \in \mathfrak{F}$.
By Remark 6.3**(a)**, $\big(f^{-1}(G')\big)^{\complement} = f^{-1}\big(G'^{\complement}\big)$. Thus, $f^{-1}\big(G'^{\complement}\big) \in \mathfrak{F}$, i.e., $G'^{\complement} \in \mathcal{G}'$.
The proof that $\mathcal{G}'$ satisfies the other properties of a σ–algebra is just as straightforward.
Now let us prove that $\{f^{-1}(E') : E' \in \mathcal{E}'\} \subseteq \mathfrak{F} \Rightarrow f$ is $(\mathfrak{F}, \mathfrak{F}')$–measurable.
So assume that $\{f^{-1}(E') : E' \in \mathcal{E}'\} \subseteq \mathfrak{F}$. By (6.12), each $E' \in \mathcal{E}'$ belongs to $\mathcal{G}'$, i.e., $\mathcal{E}' \subseteq \mathcal{G}'$.
It follows that $\sigma\{\mathcal{E}'\} \subseteq \sigma\{\mathcal{G}'\}$. Since $\mathcal{E}'$ generates $\mathfrak{F}'$ and $\mathcal{G}'$ is a σ–algebra, we obtain that $\mathfrak{F}' \subseteq \mathcal{G}'$.
Thus, $A' \in \mathfrak{F}' \Rightarrow A' \in \mathfrak{G}' \overset{(6.12)}{\Rightarrow} f^{-1}(A') \in \mathfrak{F}$. This proves that f is $(\mathfrak{F}, \mathfrak{F}')$–measurable. ■

Definition 5.17 on p.131 stated the independence of an arbitrary family $(X_i)_{i\in I}$ of random elements. It is not mathematically precise, since the role of σ–algebras is not considered there. Now, that the concept of measurability has been made available, we can rewrite that definition in its precise form. For the following, recall Definition 6.3 (Advanced level definition of random variables and random elements) on p.141.

Definition 6.5 (Independence of arbitrarily many random elements – advanced definition). ⋆

Given are a probability space $(\Omega, \mathfrak{F}, P)$, a measurable space $(\Omega', \mathfrak{F}')$, and a family of random elements,

$$X_i : (\Omega, \mathfrak{F}, P) \longrightarrow (\Omega', \mathfrak{F}') \qquad (i \in I).$$

We say that this family is **independent** if, for ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ and $j = 1, 2, \ldots, k$,

$$\begin{aligned} &P\{\, X_{i_1} \in A'_{i_1},\, X_{i_2} \in A'_{i_2},\, \ldots,\, X_{i_k} \in A'_{i_k}\,\} \\ &\qquad = P\{X_{i_1} \in A'_{i_1}\} \cdot P\{X_{i_2} \in A'_{i_2}\} \cdots P\{X_{i_k} \in A'_{i_k}\}\,, \quad \text{for all } A'_{i_j} \in \mathfrak{F}'. \ \square \end{aligned} \tag{6.13}$$

Remark 6.5. ⋆ Convince yourself that the only difference between (6.15) and (5.48) is this: "$A'_{i_j} \subseteq \mathfrak{F}'$" has been replaced with the (easier to satisfy) condition "$A'_{i_j} \subseteq \Omega'$". $\square$

For completeness' sake, we give the definition of independence of a family of sets. Note that there will be no more codomain $(\Omega', \mathfrak{F}')$, because there will be no more functions X_i on $(\Omega, \mathfrak{F}, P)$. The sets $A_{i_j} \in \mathcal{E}_{i_j}$ given there will be subsets of Ω!

Definition 6.6 (Independence of a family of sets of measurable sets). ⋆

Given are a probability space $(\Omega, \mathfrak{F}, P)$, and a family

$$\boldsymbol{\mathcal{E}}_i \subseteq \mathfrak{F} \qquad (i \in I).$$

We say that this family is **independent** if, for ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ and $j = 1, 2, \ldots, k$, and for any choices $A_{i_j} \in \boldsymbol{\mathcal{E}}_{i_j}$,

(6.14) $$P\{ A_{i_1} \cap A_{i_2} \cap \cdots, \cap A_{i_k} \} = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k}), \quad \text{for all } A_{i_j} \in \boldsymbol{\mathcal{E}}_{i_j}. \ \square$$

Definition 6.7 can be expressed in terms of Definition 6.6 as follows.

Proposition 6.2. ⋆

Given are a probability space $(\Omega, \mathfrak{F}, P)$, a measurable space $(\Omega', \mathfrak{F}')$, and a family of random elements,

$$X_i : (\Omega, \mathfrak{F}, P) \longrightarrow (\Omega', \mathfrak{F}') \qquad (i \in I).$$

Then,

The family $(X_i)_i \in I$ is independent $\Leftrightarrow$ the family $\sigma\{X_i\}_{i \in I}$ is independent.

PROOF: In Definition 6.6, set $\boldsymbol{\mathcal{E}}_i := \sigma\{X_i\}$. ■

Remark 6.6. ⋆ What we discuss now about functions is a very general phenomenon in mathematics. So let us switch briefly the notation to that of general mathematics. Let

$$X \xrightarrow{f} Y; \qquad x \mapsto y = f(x),$$

be a function f with domain X and codomain Y. One can consider Y as being in forward direction and X in backward direction of the function arrow $\xrightarrow{f}$.

(a) Let $\mathscr{S}_X$ be some mathematical structure on X, the domain of f. Assume that f can be used to generate a corresponding mathematical structure, $\mathscr{S}_Y$, on Y, the codomain of f. Since $\mathscr{S}_Y$ was created from $\mathscr{S}_X$ by "pushing forward" that structure from X to Y by means of f, mathematicians will speak of $\mathscr{S}_Y$ as the **push–forward** of $\mathscr{S}_X$ by f.

(b) Let $\widetilde{\mathscr{S}}_Y$ be some mathematical structure on Y, the codomain of f. Assume that f can be used to generate a corresponding mathematical structure, $\widetilde{\mathscr{S}}_X$, on X, the domain of f. Since $\widetilde{\mathscr{S}}_X$ was created from $\widetilde{\mathscr{S}}_Y$ by "pulling back" that structure from Y to X by means of f, mathematicians will speak of $\widetilde{\mathscr{S}}_X$ as the **pull–back** of $\widetilde{\mathscr{S}}_Y$ by f.

Here is a very good example of a push–forward. Let $X : (\Omega, \mathfrak{F}, P) \to (\Omega', \mathfrak{F}')$ denote a random element. The probability measure P certainly is a mathematical structure on Ω, the domain of X.

- The function X pushes P forward to the distribution P_X of X, a probability measure on $(\Omega', \mathfrak{F}')$, by means of the formula $P_X(A') = P\{X \in A'\}, \ A' \in \mathfrak{F}'$.

P_X certainly is a mathematical object of Ω', the codomain of X.

For the following example of a pull–back, let $f : \Omega \to (\Omega', \mathfrak{F}')$. Here, Ω is some nonempty set, and $(\Omega', \mathfrak{F}')$ denotes a measurable space. Note that he σ–algebra $\mathfrak{F}'$ will be very important here!

- The function f pulls $\mathfrak{F}'$ back to $\sigma\{f\}$, a σ–algebra on Ω, by means of the formula [70]

$$\sigma\{f\} \;=\; \sigma\{f^{-1}(A') : A' \in \mathfrak{F}'\}\,.$$

How does the above relate to a pull–back, as discussed in **(b)**? The σ–algebra $\mathfrak{F}'$ is a mathematical structure on Ω', the codomain of f. This function pulls $\mathfrak{F}'$ back to a σ–algebra $\sigma\{f\}$ on Ω. This σ–algebra certainly is a mathematical object of Ω, the domain of f.
As an aside, one would also refer to $\sigma\big\{(f_i)_{i\in I}\big\}$ as the pull-back of $\mathfrak{F}'$ by means of a family $(f_i)_{i\in I}$ of functions $f_i : \Omega \to (\Omega', \mathfrak{F}')$. □

It is definitely OK to skip this next remark.

Remark 6.7. ⋆
Like σ–algebras, measurability is a theoretical concept that aids in the development of probability theory as a mathematical theory. Such concepts are tools that help understand why the practical things taught here about solving applications oriented problems yield the intended results. This in turn helps to see various items as connected rather than unrelated, and that in turn makes it easier to remember the applications oriented material and use it in situations that are not an obvious fit for any of the cookbook recipes.
That having been said, measurability will not be an issue in this course! □

Definition 5.17 on p.131 concerning the independence of an arbitrary family $(X_i)_{i\in I}$ of random elements) is not mathematically precise, since the role of σ–algebras is not considered there. Now, that the concept of measurability has been made available, we can state that definition in its precise form.

Definition 6.7 (Independence of arbitrarily many random elements – precise definition)**.** ⋆

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\big(X_i\big)_{i\in I}$ of random elements on Ω. Here, I denotes an arbitrary set of indices. We say that this family is **independent** if, for ANY FINITE subselection of distinct indices $i_1, i_2, \ldots, i_k \in I$ and $j = 1, 2, \ldots, k$,

$$\begin{aligned} &P\{\, X_{i_1} \in A'_{i_1},\, X_{i_2} \in A'_{i_2},\, \ldots,\, X_{i_k} \in A'_{i_k}\,\} \\ &\qquad = P\{X_{i_1} \in A'_{i_1}\} \cdot P\{X_{i_2} \in A'_{i_2}\} \cdots P\{X_{i_k} \in A'_{i_k}\}\,, \quad \text{for all } A'_{i_j} \subseteq \Omega'. \end{aligned} \tag{6.15}$$

□

6.2 Measures

This chapter is very selective and incomplete at this point in time. Additions will be made as time allows.

Introduction 6.2. There is so much commonality between Lebesgue measure and probability measures that it justifies an overarching term, that of a measure. Many results that hold true for both Lebesgue measure and a probability measure also are true for measures. □

[70] See Definition 6.4 (Advanced Definition of σ–algebra generated by random elements) on p.143.

Based on those introductory remarks, we introduce the concept of an abstract (i.e., general) measure.

Definition 6.8 (Abstract measures). ⋆

Let $(\Omega, \mathfrak{F})$ be a measurable space. A **measure** on $\mathfrak{F}$ is an extended real–valued function

$$\mu : \mathfrak{F} \to [0, \infty]; \quad A \mapsto \mu(A) \qquad \text{such that}$$

(6.16) $\mu(\emptyset) = 0$,

(6.17) $A, B \in \mathfrak{F}$ and $A \subseteq B \quad \Rightarrow \quad \mu(A) \leq \mu(B)$, (monotony)

(6.18) $(A_n)_{n\in\mathbb{N}} \in \mathfrak{F}$ disjoint $\quad \Rightarrow \quad \mu\Big(\biguplus_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n)$. ($\sigma$–additivity)

- The triplet $(\Omega, \mathfrak{F}, \mu)$ is called a **measure space**
- We call any set $N \subseteq \Omega$ with measure zero a μ **Null set**.
- We call μ a **discrete measure** if there is a countable $A^* \in \mathfrak{F}$ such that $\mu\left(A^{*\complement}\right) = 0$. We then call $(\Omega, \mathfrak{F}, \mu)$ a **discrete measure space**
- We call μ a **finite measure** on $\mathfrak{F}$ if $\mu(\Omega) < \infty$.
- We call μ a **σ–finite measure** on $\mathfrak{F}$ if one can find a sequence $A_n \in \mathfrak{F}$ such that $\mu(A_n) < \infty$ and $\Omega = \bigcup_n A_n$.

Footnotes for measurable spaces, [71] extended real numbers, [72] and μ–null sets. [73] □

Do not confuse measurable spaces $(\Omega, \mathfrak{F})$ and measure spaces $(\Omega, \mathfrak{F}, \mu)$!

Remark 6.8. ⋆

(a) Lebesgue measure λ^d is a measure on $\mathfrak{B}^d$, and $(\mathbb{R}^d, \mathfrak{B}^d, \lambda^d)$ and $(\mathbb{R}, \mathfrak{B}, \lambda^1)$ are measure spaces. Note that λ^d is infinite, since $\lambda^d(\mathbb{R}^d) = \infty$.
On the other hand, λ^d is σ–finite, since $K_n := [-n, n]^d \uparrow \mathbb{R}^d$ and $\lambda^d(K_n) = (2n)^d < \infty$.

(b) A probability measure is a finite measure. Probability spaces are measure spaces.

(c) A measure μ is a probability measure $\Leftrightarrow \quad \mu(\Omega) = 1$. □

Example 6.2. ⋆ Let $A \in \mathfrak{B}^d$ and $\mathfrak{B}(A) := \{B \in \mathfrak{B}^d : B \subseteq A\}$. Consider

$$\lambda^d\Big|_A : \mathfrak{B}(A) \to [0, \infty]; \qquad B \mapsto \lambda^d\Big|_A(B) := \lambda^d(B).$$

Then $(A, \mathfrak{B}(A), \lambda^d\Big|_A)$ is a measure space. Note that $\lambda^d\Big|_A$ is the restriction of the function $\lambda^d : \mathfrak{B}^d \to [0, \infty]$ to the subset $\mathfrak{B}(A)$ of $\mathfrak{B}^d$.

[71] See Definition 6.1 (Measurable functions) on p.137.
[72] See Definition 2.14 (Extended real numbers) on p.28.
[73] Strictly speaking any set N such that $N \subseteq A$ and $\mu(A) = 0$ is said to be μ Null. We ignore such fine points.

For example, let $A := \{\vec{x} = (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 25\}$. Let $B := [-1, 2] \times]1, 3[$.
Then $\lambda^2\Big|_A(B) = \lambda^2(B) = 6$.

It is customary to call $\lambda^d\Big|_A$ the **Lebesgue measure on** A, to write λ^d for $\lambda^d\Big|_A$, and to call $\mathfrak{B}(A)$ the **Borel sets of** A. □

Example 6.3. ⋆ Let Ω be a nonempty, countable set. Let $g : \Omega \to [0, \infty]$ be an arbitrary function (satisfying $0 \leq g(\omega) \leq \infty$ for all ω). We associate with g the function

$$\mu : 2^\Omega \longrightarrow [0, \infty]\,; \qquad A \mapsto \mu(A) := \sum_{\omega \in A} g(\omega)\,. \tag{6.19}$$

Then μ defines a measure on 2^Ω. We can see this as follows.
Clearly, $\mu(\emptyset) = 0$. Let $A \subseteq B \subseteq \Omega$ and $C := B \setminus A$.. Since $B = A \uplus C$ and $g \geq 0$, we obtain

$$\mu(B) = \sum_{\omega \in B} g(\omega) = \sum_{\omega \in A \uplus C} g(\omega) = = \sum_{\omega \in A} g(\omega) + \sum_{\omega \in C} g(\omega) = \mu(A) + \mu(C) \geq \mu(A)\,. \tag{6.20}$$

Key in (6.22) is that the disjointness of A and C allowed us to write $\sum_{\omega \in A \uplus C} = \sum_{\omega \in A} + \sum_{\omega \in C}$.
To prove σ–additivity, we use that same trick for a disjoint sequence $A_j \subseteq \Omega$ and $A := \biguplus_j A_j$:

$$\mu(A) = \sum_{\omega \,\in\, \biguplus_j A_j} g(\omega) = \sum_{j \in \mathbb{N}} \sum_{\omega \in A_j} g(\omega) = \sum_{j \in \mathbb{N}} \mu(A_j)\,. \quad \square$$

Example 6.4. ⋆ Let Ω be a nonempty, countable set. Let

$$\boldsymbol{\mathcal{E}} := \big\{\, \{\omega\} : \omega \in \Omega \,\big\} = \{\text{ all singleton sets of } \Omega \,\}\,.$$

Let $\mu_0 : \boldsymbol{\mathcal{E}} \to [0, \infty]$ be an arbitrary function on $\boldsymbol{\mathcal{E}}$ (satisfying $0 \leq \mu_0\{\omega\} \leq \infty$ for all $\omega \in \Omega$). We associate with μ_0 the function

$$h : \Omega \longrightarrow [0, \infty]\,; \qquad \omega \mapsto h(\omega) := \mu_0\{\omega\}\,. \tag{6.21}$$

We have seen in Example 6.4, with $g(\omega) := h(\omega)$, that $\mu(A) := \sum_{\omega \in A} h(\omega)$ defines a measure on 2^Ω.
Of course there will be some relationship between μ_0 and μ. It is as follows.

$$\mu\{\omega\} = \sum_{\omega' \in \{\omega\}} h(\omega') = h(\omega) \overset{(6.21)}{=} \mu_0\{\omega\}\,. \tag{6.22}$$

In other words, μ is the unique extension of the function μ_0 to a measure on 2^Ω. □

Example 6.5. ⋆ In examples 6.3 and 6.4 It was not all that important that Ω itself be countable. All that is needed that there is a countable set $A^* \subseteq \Omega$ as follows.

- **(a)** in Example 6.3: $g(\omega) = 0$ for $\omega \notin A^*$;
- **(b)** in Example 6.4: $\mu_0\{\omega\} = 0$ for $\omega \notin A^*$.

The reason is that adding zeroes has no effect: If $A \subseteq \Omega$, then $\sum_{\omega \in A} = \sum_{\omega \in A \cap A^*}$.

- **(c)** The existence of such countable A^* is equivalent to the measure μ being discrete. That applies to both **(a)** (i.e., Example 6.3) and **(b)**)i.e., Example 6.4). □

Remark 6.9. ⋆

The measure μ of Example 6.5 is a discrete probability measure
$\Leftrightarrow$ g of (6.19) satisfies $\sum_{\omega\in A^*} g(\omega) = 1$
$\Leftrightarrow$ μ_0 of (6.22) satisfies $\sum_{\omega\in A^*} \mu_0\{\omega\} = 1$

Thus, it is easy to derive Theorem 5.2 on p.111 and Corollary 5.1 from Examples 6.3–6.5. □

Next, we consider what happens in Example 6.5 if $g(\omega) = \mathbf{1}_{A^*}(\omega) = 1$, for $\omega \in A^*$, and 0, else. Then

$$\mu(A) = \sum_{\omega\in A\cap A^*} 1 = |A \cap A^*|$$

In other words, $\mu(A)$ counts how many elements of A^* fall into A.

Definition 6.9 (Counting measure). ⋆

Let $(\Omega, \mathfrak{F})$ be a measurable space, $A^* \neq \emptyset$ a countable subset of Ω
(a) We call the measure Σ_* on $\mathfrak{F}$, defined by

$$\Sigma_*(A) := |A \cap A^*|$$

the **counting measure** on $\mathfrak{F}$ with respect to A^*.
(b) In particular, if $\Omega \subseteq \mathbb{R}$ and $A^* = \mathbb{N}$, we call Σ_* the **standard counting measure** on $\mathfrak{F}$.
(c) If no reference to a σ–algebra is made, we set $\mathfrak{F} := 2^\Omega$. □

Example 6.6. ⋆ Here are some examples for the counting measure.

(a) $(\Omega, \mathfrak{F}) = (\mathbb{R}, 2^{\mathbb{R}})$, $A^* = \mathbb{N}$: ⊡ $\Sigma_*([3,5]) = 3$ ⊡ $\Sigma_*(]3,5[) = 1$ ⊡ $\Sigma_*([-\pi, e]) = 2$ ⊡ $\Sigma_*([-\pi, e] \cup \{-8, 2, 3, 3.5, 4, \}) = 4$.

(b) $(\Omega, \mathfrak{F}) = (\mathbb{R}, 2^{\mathbb{R}})$, $A^* = \mathbb{Z}$: ⊡ $\Sigma_*([3,5]) = 3$ ⊡ $\Sigma_*(]3,5[) = 1$ ⊡ $\Sigma_*([-\pi, e]) = 6$ ⊡ $\Sigma_*([-\pi, e] \cup \{-8, 2, 3, 3.5, 4, \}) = 9$.

(c) Given the measurable space $(\mathbb{R}^3, \mathfrak{B}^3)$ and $\vec{a} = (1, 0, 5)$, let $\delta_{\vec{a}}$ be the counting measure on $\{\vec{a}\}$. The symbols need not be $\Omega, \mathfrak{F}, A^*, \Sigma_*$!
⊡ $\delta_{\vec{a}}(\mathbb{R}^3) = \delta_{\vec{a}}([0.8, 1]\times] - 1, 1[\times\{5, 9\}) = \delta_{\vec{a}}(\vec{a}) = 1$ ⊡ $\delta_{\vec{a}}\Big(\big([0,5]^3\big)^{\complement}\Big) = 0$
⊡ $B \in \mathfrak{B}^3 \Rightarrow \delta_{\vec{a}}(B) = \mathbf{1}_B(\vec{a}) = 1$ if $\vec{a} \in B$ and 0 otherwise.

(d) We generalize **(c)** as follows. Given an arbitrary set Ω and $\omega_0 \in \Omega$, let δ_{ω_0} be the counting measure on $\{\omega_0\}$. No σ–algebra $\mathfrak{F}$ was mentioned, so $\mathfrak{F} = 2^\Omega$. Then
⊡ $A \subseteq \Omega \Rightarrow \delta_{\omega_0}(A) = \mathbf{1}_A(\omega_0)$, i.e., $\delta_{\omega_0}(A) = \begin{cases} 1, & \text{if } \omega_0 \in A\,, \\ 0, & \text{if } \omega_0 \notin A\,. \end{cases}$
Since $\delta_{\omega_0}(\Omega) = 1$, δ_{ω_0} is a probability measure on 2^Ω. It often is referred to as the **unit mass** at ω_0. □

Here are two not very useful measures which are easy to understand.

Example 6.7. ⋆

One can easily verify that the following set functions μ_1 and μ_2 define measures on an arbitrary nonempty set Ω with an arbitrary σ–algebra $\mathfrak{F}$.

(6.23) $\mu_1(A) := 0$ for all $A \in \mathfrak{F}$, **zero measure or Null measure**

(6.24) $\mu_2(\emptyset) := 0; \qquad \mu(A) := \infty$ if $A \neq \emptyset$.

Keep the second example in mind when you work with non–finite measures. □

Remark 6.10. ⋆

(1) We emphasize that the only difference between (general) measures and probability measures is that the latter must assign a measure of one to the entire space Ω.

(2) Many things that apply to probabilities can be extended to general measures, and this will matter to us even if we are only interested in probability spaces, since will see in the context of the expectation $E[Y]$ of a random variable Y, that assignments of the form

$$A \mapsto E[X \cdot \mathbf{1}_A] \text{ where } A \in \mathfrak{F} \text{ and } \mathbf{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A\,, \\ 0 & \text{if } \omega \notin A\,, \end{cases}$$

i.e., $\mathbf{1}_A$ is the indicator function of A, define a measure on $(\Omega, \mathfrak{F})$.

(3) A measure space can support many different measures: If μ is a measure on $\mathfrak{F}$ and $\alpha \geq 0$ then $\alpha\mu : A \mapsto \alpha\mu(A)$ also is a measure on $\mathfrak{F}$. □

The following generalizes Fact 5.1 on p.115 from probability measures to σ–finite measures (and thus to all reasonable measures).

Fact 6.1. ⋆ *For the following, note that the sets $\mathfrak{I}_1, \ldots, \mathfrak{I}_8$ were defined in Example 5.8 on p.114.*

- *Let $\mathfrak{I} = \mathfrak{I}_5$ or $\mathfrak{I} = \mathfrak{I}_8$. Let the function $\mu_0 : \mathfrak{I} \to [0, \infty[$ (so $E \in \mathfrak{I} \Rightarrow \mu_0(E) < \infty$) satisfy $\mu_0(\emptyset) = 0$, $\mu_0(\mathbb{R}^d) = 1$ and σ–additivity on $\mathfrak{I}$: $E_n \in \mathfrak{I}$ disjoint such that $E := \biguplus_{n\in\mathbb{N}} E_n \in \mathfrak{I} \Rightarrow \mu_0(E) = \sum_{n\in\mathbb{N}} \mu_0(E_n)$.*
  Then μ_0 can be uniquely extended to a measure on $\mathfrak{B}^d$, the Borel sets of $\mathbb{R}^d$.

  - *One can drop the requirement that $\mu_0(A) < \infty$ for all $A \in \mathfrak{I}$, but then the extension μ is no more guaranteed to be unique.*
  - *Note that for $d = 1$, the following sets are equal: $\mathfrak{I}_5 = \mathfrak{I}_1$, $\mathfrak{I}_8 = \mathfrak{I}_4$, $\mathfrak{B}^1 = \mathfrak{B}$.* □

Here is a simple consequence of the monotone convergence theorem for Lebesgue integrals.

Theorem 6.6. ⋆

Let $f : \mathbb{R}^n \to \mathbb{R}$ be nonnegative and Borel–measurable. Then the set function

(6.25) $$\mu : \mathfrak{B}^d \longrightarrow [0, \infty], \qquad \mu(A) := \int_A f \, d\lambda^d$$

defines a measure on $\mathfrak{B}^d$.

PROOF: We must show that

(1) $\mu(A) \geq 0$ for $A \in \mathfrak{B}^d$,

(2) $\mu(\emptyset = 0)$,

(3) If $A, B \in \mathfrak{B}^d$ and $A \subseteq B$, then $\mu(A) \leq \mu(B)$,

(4) If $A_j \in \mathfrak{B}^d$ are mutually disjoint ($j \in \mathbb{N}$), then $\mu\left(\biguplus_j A_j\right) = \sum\limits_{j\in\mathbb{N}} \mu(A_j)$.

Since $\mathbf{1}_\emptyset = 0$, $\mathbf{1}_\emptyset \cdot f = 0$. Thus

$$\mu(\emptyset) \;=\; \int \mathbf{1}_\emptyset \cdot f \, d\lambda^d \;=\; \int 0 \, d\lambda^d \;=\; 0.$$

This proves **(2)**. To show **(1)** and **(3)**, we will use the monotonicity of the Lebesgue integral: [74]

$\mathbf{1}_A \cdot f \geq 0$ on $\mathbb{R}^d \;\Rightarrow\; 0 \;=\; \int 0 \, d\lambda^d \;\leq\; \int \mathbf{1}_A \cdot f \, d\lambda^d \;=\; \mu(A)$. This proves **(1)**.

$A \subseteq B \;\Rightarrow\; \mathbf{1}_A \cdot f \leq \mathbf{1}_B \cdot f$ on $\mathbb{R}^d \;\Rightarrow\; \mu(A) \;=\; \int \mathbf{1}_A \cdot f \, d\lambda^d \;\leq\; \int \mathbf{1}_B \cdot f \, d\lambda^d \;=\; \mu(B)$.

This proves **(3)**. It remains to prove σ–additivity. Let

$$B_n \;:=\; \biguplus_{j\leq n} A_j, \quad B \;:=\; \biguplus_{j\in\mathbb{N}} A_j \;=\; \bigcup_{j\in\mathbb{N}} B_j, \quad g_n \;:=\; \mathbf{1}_{B_n} f, \quad g \;:=\; \mathbf{1}_B f\,.$$

We claim that

$$0 \;\leq\; g_n \;\uparrow\; g \text{ as } n \to \infty, \quad \text{i.e.,} \quad 0 \;\leq\; g_n(\vec{x}) \;\uparrow\; g(\vec{x}), \quad \text{for each } \vec{x} \in \mathbb{R}^d. \tag{6.26}$$

(A) First, let $\vec{x} \notin B$.

Then $\mathbf{1}_{B_n}(\vec{x}) = \mathbf{1}_B(\vec{x}) = 0 \;\Rightarrow\; g_n(\vec{x}) = g(\vec{x}) = 0$, for all n, it follows that $g_n(\vec{x}) = g(\vec{x})$.

(B) Now we assume that $\vec{x} \in B$.

Since $B = \bigcup_n B_n$, $\vec{x} \in B_{n_0}$ for some index n_0. Since $B_n \uparrow$, $\vec{x} \in B_j$ for all $j \geq n_0$.

Thus $g_j(\vec{x}) = g(\vec{x})$ for all $j \geq n_0$. Of course, n_0 will vary with $\vec{x}$, but that is OK.

We have shown for both cases **(A)** and **(B)** that $g_n(\vec{x}) = g(\vec{x})$ for large enough n.

Moreover, $B_n \uparrow$ implies $g_n = \mathbf{1}_{B_n} \uparrow$. We have shown that $g_n(\vec{x}) \uparrow g(\vec{x})$ for all $\vec{x}$.

Finally, note that $\mathbf{1}_{B_n} \geq 0$ and $f \geq 0 \;\Rightarrow\; g_n = \mathbf{1}_{B_n} f \geq 0$. We have shown (6.26).

We apply the definitions of g_n and g to (6.26) and obtain

$$0 \;\leq\; \mathbf{1}_{\biguplus[A_j : j\leq n]} f \;=\; \mathbf{1}_{B_n} f \;=\; g_n \;\uparrow\; g \;=\; \mathbf{1}_B f \;=\; \mathbf{1}_{\biguplus[A_j : j\in\mathbb{N}]} f\,.$$

We apply the monotone convergence property of Lebesgue integrals, [75] and obtain

$$\int \mathbf{1}_{\biguplus[A_j : j\leq n]} f \, d\lambda^d \;\uparrow\; \int \mathbf{1}_{\biguplus[A_j : j\in\mathbb{N}]} f \, d\lambda^d \;=\; \int_{\biguplus[A_j : j\in\mathbb{N}]} f \, d\lambda^d \;=\; \mu\Big(\biguplus_{j\in\mathbb{N}} A_j\Big)\,. \tag{6.27}$$

Since the A_j are disjoint, $\mathbf{1}_{\biguplus[A_j : j\in\mathbb{N}]} \;=\; \sum\limits_{j\leq n} \mathbf{1}_{A_j}$. By Linearity II of Lebesgue integrals [76] plus (6.27),

[74] see Theorem 4.5**(b)** on p.94

[75] see Theorem 4.5**(d)** on p.94

[76] see Theorem 4.5**(c)** on p.94

$$\sum_{j=1}^{n}\mu(A_j) \;=\; \sum_{j=1}^{n}\int f\cdot\mathbf{1}_{A_j}\,d\lambda^d \;=\; \int\Big(\sum_{j=1}^{n} f\cdot\mathbf{1}_{A_j}\Big)\,d\lambda^d \;=\; \int \mathbf{1}_{\biguplus[A_j:j\le n]}f\,d\lambda^d \;\uparrow\; \mu\Big(\biguplus_{j\in\mathbb{N}} A_j\Big)\,.$$

We have shown that μ also is σ–additive. It follows that μ is a measure. ■

The next theorem corresponds to Theorem 5.1 on p.109. But note the additional requirement $\mu(B_1)<\infty$ for the case of nonincreasing sequences of measurable sets. It makes (6.29) significantly different from the corresponding formula (5.16) for probability measures.

Theorem 6.7 (Continuity property of measures). ⋆

Let $(\Omega,\mathfrak{F},\mu)$ be a measure space. If $A_n,B_n\in\mathfrak{F}$, then the following is true:

$$A_n\uparrow \Rightarrow\; P(A_n)\uparrow \mu\left(\bigcup_{n\in\mathbb{N}} A_n\right)\,, \tag{6.28}$$

$$B_n\downarrow \text{ and } \mu(B_1)<\infty \;\Rightarrow\; P(B_n)\downarrow P\left(\bigcap_{n\in\mathbb{N}} B_n\right)\,. \tag{6.29}$$

PROOF: The proof of (6.28) is very similar to that of thm-x:prob-meas-continuity-prop:eqn01 (for probability measures) and left as an exercise.

Proof of (6.29) – Outline: Modify the proof of (5.16) as follows:

- Replace all complements $U^{\complement}$ with set differences$B_1\setminus U$.
- Use the relation $\mu(U)+\mu(B_1\setminus U)=\mu(B_1)<\infty$ instead of $P(U)+P(U^{\complement})=1$. ■

Proposition 6.3. *Let $(\Omega,\mathfrak{F},\mu)$ be a measure space and $(\Omega',\mathfrak{F}')$ a measurable space.*

Let $f:\Omega\to\Omega'$ be $(\mathfrak{F},\mathfrak{F}')$ measurable. Then the set function

$$\mu_f:\mathfrak{F}'\to[0,\infty];A'\mapsto \mu\{f\in A'\} \;=\; \mu\{\omega\in\Omega\,:\,f(\omega)\in A'\} \tag{6.30}$$

defines a measure on $(\Omega',\mathfrak{F}')$. Moreover, if μ is a probability measure on $\mathfrak{F}$, i.e., $\mu(\Omega=1)$, then μ_f is a probability measure on $\mathfrak{F}'$.

PROOF: ⋆ $\mu_f(\emptyset)=0$, since $f^{-1}(\emptyset)=\emptyset$, and μ is a measure.

We show here in detail that μ_f is monotone: $A'\subseteq B'\;\Rightarrow\;\mu_f(A')\le\mu_f(B')$, for all $A',B'\in\mathfrak{F}'$. According to Proposition 2.7 on p.44, $A'\subseteq B'$ implies $f^{-1}(A')\subseteq f^{-1}(B')$. Since μ is a measure, this implies $\mu\big(f^{-1}(A')\big)\le\mu\big(f^{-1}(B')$, i.e., by definition of μ_f, $\mu_f(A')\le\mu_f(B')$

The proof that $\mu_f\left(\biguplus_n B_n\right)=\sum_n\mu_f(B_n)$ for any disjoint sequence $B_n\in\mathfrak{F}'$, is just as simple, since the order of taking preimages and unions can be switched. See Theorem 2.2 (f^{-1} is compatible with all basic set ops) on p.45. ■

Definition 6.10 (Image measure). ⋆

(1) We call the measure μ_f of Proposition 6.3 the **image measure** of μ under f or the **measure induced** by μ and f.

(2) We now switch notation from f and μ to the more customary X and P for the sake of clarity. In the case of a random element X on a probability space $(\Omega, \mathfrak{F}, P)$ with codomain $(\Omega', \mathfrak{F}')$, we call the image measure P_X of P under X which is, according to (6.30), given by

$$P_X(B) := P\{X \in B\} = P\{\omega \in \Omega : X(\omega) \in B\}, \quad (B \in \mathfrak{B}^1) \tag{6.31}$$

the **probability distribution** or simply the **distribution** of X. □

Remark 6.11. ⋆ Except for the added measurability conditions (which you may ignore if you like), the definition above of a probability distribution matches that of Definition 5.13 (Probability Distribution) on p.124. □

6.3 Abstract Integrals

This chapter is very selective and incomplete at this point in time. Additions will be made as time allows.

Introduction 6.3. In Chapter 4 (Calculus Extensions) we introduced the Lebesgue integral, $\int f d\lambda^d$, as an extension of the Riemann integral, $\int f(\vec{x}) d\vec{x}$, to a larger class of integrands. In practice, all functions one deals with are Riemann integrable. What then is the purpose of the Lebesgue integral as an alternate definition?

The answer is that the Lebesgue integral allows the mathematician to prove certain assertions that are of huge practical importance. We mention the monotone convergence property. [77] It was used to show that $A \mapsto \int_A f d\lambda^d$ defines a measure, [78] but it also is useful in practical applications that require computing certain integrals.

Only a fairly limited amount of changes in that theory is needed to define integrals $\int f d\mu$ of real valued functions f with respect to an abstract measure μ.

Even though all definitions, theorems, and remarks make extensive use of σ–algebras, they can be ignored by the student who is not interested in understanding the proof. The reasons have been stated more than once already:

- All sets A that occur in practice can be assigned a measure $\mu(A)$, in particular, a probability $P(A)$ or Lebesgue measure $\lambda^d(A)$.
- No matter what kind of measurable space $(\Omega, \mathfrak{F})$ is considered one can act as if $\mathfrak{F}) = 2^\Omega$, the collection of all subsets of Ω. □

The next definition corresponds to Definition 4.4 on p.88.

[77] see Theorem 4.5**(d)** on p.94
[78] see Theorem 6.6 on p.151

Definition 6.11 (Abstract integral for simple functions). ⋆ Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space [79]

Let $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathfrak{F}$, $c_1, \ldots, c_n \in [0, \infty[$ Let

$$f : (\Omega, \mathfrak{F}, \mu) \longrightarrow \mathbb{R}; \qquad f(\omega) \;=\; \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\omega)\,.$$

The **abstract integral** aka **integral** of the simple function f with respect to μ is

$$\int f\,d\mu \;:=\; \int f(\omega)\,d\mu(\omega) \;:=\; \int f(\omega)\mu(d\omega) \;:=\; \sum_{j=1}^{n} c_j \mu(A_j)\,. \quad \square \tag{6.32}$$

Proposition 6.4. ⋆

Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space. Let $f, g_n, h_n : (\Omega, \mathfrak{F}, \mu) \longrightarrow \mathbb{R}$ be nonnegative, $(\mathfrak{F}, \mathfrak{B}^1)$measurable functions. Assume further that the functions g_n and h_n are simple. Then the following is true:

$$\textit{If } \; g_n \uparrow f \; \textit{ and } \; h_n \uparrow f, \; \textit{ then } \quad \lim_{n\to\infty} \int g_n d\mu \;=\; \lim_{n\to\infty} \int h_n d\mu\,. \tag{6.33}$$

PROOF: ■

By Theorem 6.2 on p.139, any nonnegative and $(\mathfrak{F}, \mathfrak{B}^1)$measurable function f can be approximated from below by a sequence of nonnegative, simple functions f_n. There potentially is a huge number of such function sequences, but the previous proposition shows that $\lim_{n\to\infty} \int f_n d\mu$ does not depend on the particular approximating sequence. This enables us to make the next definition, which is the counterpart of Definition 4.5 (Lebesgue integral) on p.90

Definition 6.12 (Abstract integral for measurable functions). ⋆

(a) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space, $f, f_n : (\Omega, \mathfrak{F}, \mu) \longrightarrow \mathbb{R}$ $(\mathfrak{F}, \mathfrak{B}^1)$measurable, and assume that the functions f_n are simple and • **either** $0 \le f_n \uparrow f$ • **or** $0 \ge f_n \downarrow f$. Then

$$\int f d\mu \;:=\; \lim_{n\to\infty} \int f_n d\mu \tag{6.34}$$

is called the **abstract integral** aka **integral of f with respect to μ**. $\square$

(b) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and $f : (\Omega, \mathfrak{F}, \mu) \longrightarrow \mathbb{R}$ $(\mathfrak{F}, \mathfrak{B}^1)$measurable, such that

- both f^+ and f^- are limits of nondecreasing sequences of simple functions ≥ 0;
- at least one of $\int f^+ d\mu$, $\int f^- d\mu$ is finite. (According to **(a)**, those integrals exist, but neither of them was guaranteed to be finite.)

Then we define the **abstract integral** aka **integral of f with respect to μ**, as the expression

$$\int f\,d\mu \;=\; \int (f^+ - f^-)\,d\mu \;:=\; \int f^+\,d\mu \;-\; \int f^-\,d\mu\,. \tag{6.35}$$

[79] see Definition 6.8 (Abstract measures) on p.148

(c) We call a real–valued function f **μ–integrable**, if $\int f\,d\mu$ exists and is finite. □

Remark 6.12. ⋆ Note that $\int f d\mu$ may be infinite, even for simple and bounded f.
As an example, let $\Omega := \{0\}, \mathfrak{F} = \{\emptyset, \{0\}\}$, μ the measure defined by $\mu(\emptyset) = 0$ and $\mu(\{0\}) = \infty$. [80]

Since $0 \cdot \infty = 0$, [81] $\displaystyle\int f d\mu \;=\; f(0) \cdot \mu\{0\} \;=\; \begin{cases} \infty, & \text{if } f(0) > 0\,, \\ -\infty, & \text{if } f(0) < 0\,, \\ 0, & \text{if } f(0) = 0\,. \end{cases}$

Accordingly, some care must be exercised when defining the integral for functions which can take both positive and negative values. □

Assumption 6.1.

Unless explicitly stated otherwise, we assume the following for the remainder of this chapter (Chapter 6 (Advanced Topics – Measure and Probability)).
- The underlying measurable space is $(\Omega, \mathfrak{F})$.
- The underlying measure is μ.
- "measurable" means "$(\mathfrak{F}, \mathfrak{B}^1)$measurable". □

Here are some simple examples for integrals $\int Y\,dP = \int Y(\omega)\,P(d\omega)$ of a random variable Y with respect to a probability measure P.

We state again that you may assume that the probabilities of all events exist and therefore can ignore the σ–algebras.

Example 6.8. Assume that $Y : (\Omega, \mathfrak{F}, P) \to \mathbb{R}$ is a random variable on a probability space $(\Omega, \mathfrak{F}, P)$ which only takes finitely many distinct values, $c_1, \ldots, c_n$, i.e., $Y(\omega) \in \{c_1, \ldots, c_n\}$, for all $\omega \in \Omega$. Note the following. If $c_j \geq 0$ for all j, then such Y is a simple function in standard form, in the sense of Definition 6.2 (Simple Function on Ω) on p.138, since

$$Y(\omega) = c_j \;\Leftrightarrow\; \omega \in \{Y = c_j\} \;=\; Y^{-1}\{c_j\}. \quad \text{Thus,}\quad Y \;=\; \sum_{j=1}^{n} c_j\,\mathbf{1}_{A_j},\quad \text{with } A_j = \{Y = c_j\}.$$

(a) If $c_j \geq 0$ for all j, then (6.32) directly applies and

$$\int Y d\,P \;=\; \sum_{j=1}^{n} c_j P(A_j) \;=\; \sum_{j=1}^{n} c_j P\{Y = c_j\}\,.$$

(b) Otherwise, $[1, \ldots, n]_{\mathbb{Z}} = J_1 \uplus J_2$, where $j \in J_1 \;\Rightarrow\; c_j \geq 0$, and $j \in J_2 \;\Rightarrow\; c_j < 0$.
Convince yourself that $Y^+ = \sum\limits_{j \in J_1} c_j\,\mathbf{1}_{\{Y = c_j\}}$, and $Y^- = \sum\limits_{j \in J_2} (-c_j)\,\mathbf{1}_{\{Y = c_j\}}$. Thus,

$$\int Y d\,P \;=\; \int Y^+ d\,P - \int Y^- d\,P \;=\; \sum_{j \in J_1} c_j P\{Y = c_j\} - \sum_{j \in J_2} (-c_j) P\{Y = c_j\} \;=\; \sum_{j=1}^{n} c_j P\{Y = c_j\}\,.$$

[80] see Example 6.7 on p.150
[81] see Remark 2.8 (Extended real numbers arithmetic) on p.28

(c) In particular, assume that Y represents the toss of a fair coin which is marked 0 on one side and 1 on the other. Then [82]

$$Y \;=\; 0\cdot\mathbf{1}_{\{Y=0\}} \,+\, 1\cdot\mathbf{1}_{\{Y=1\}} \;=\; \mathbf{1}_{\{Y=1\}}\,.$$

$$\text{Thus, } \int Y\,dP \;=\; \int 1\cdot\mathbf{1}_{\{Y=1\}}\,dP \;=\; 1\cdot P\{Y=1\} \;=\; 0.5\,. \;\square$$

Example 6.9. Assume that $Y:(\Omega,\mathfrak{F},P)\to\mathbb{R}$ is a random variable on a probability space $(\Omega,\mathfrak{F},P)$ which only takes countably many distinct values, $c_1, c_2, \ldots$. We also assume that $c_j \geq 0$, for all j.

$$\text{Then } Y \;=\; \sum_{j=1}^{\infty} c_j\,\mathbf{1}_{\{Y=c_j\}}\,. \quad \text{Thus, } Y_n \;:=\; \sum_{j=1}^{n} c_j\,\mathbf{1}_{\{Y=c_j\}} \;\uparrow\; Y\,.$$

By Example 6.8**(a)**, $\displaystyle\int Y_n\,dP \;=\; \sum_{j=1}^{n} c_j P\{Y=c_j\}\,.$

By Definition 6.12 (Abstract integral for measurable functions) on p.155, $\displaystyle\int Y\,dP \;=\; \lim_{n\to\infty}\int Y_n\,dP\,.$

Thus, as expected, $\displaystyle\int Y\,dP \;=\; \sum_{j=1}^{\infty} c_j P\{Y=c_j\}\,.$

For example, assume that $c_j = j$ and $P(A_j) = (1/2)^j$, for all $j\in\mathbb{N}$. [83] Then $\int Y\,dP = \displaystyle\sum_{j=1}^{\infty} j\left(\frac{1}{2}\right)^j$.

In Section 9.3 (Geometric + Negative Binomial + Hypergeometric Distributions) we will learn that Y has a geom$(1/2)$ distribution. Also, the proof of Theorem 9.12 in that section (see p.200) shows that $\displaystyle\sum_{j=1}^{\infty} j\left(\frac{1}{2}\right)^j = 2$. Thus, $\int Y dP = 2$. $\square$

Here are some simple examples for integrals with respect to discrete measures.

Example 6.10. Assume that $f:(\Omega,\mathfrak{F},\mu)\to\mathbb{R}$ is a measurable function on a measure space $(\Omega,\mathfrak{F},\mu)$ which only takes finitely many distinct values, $c_1,\ldots,c_n$, i.e., $f(\omega)\in\{c_1,\ldots,c_n\}$, for all $\omega\in\Omega$. Such f is a simple function in standard form.

(a) If $c_j\geq 0$ for all j, then (6.32) directly applies and

$$\int f\,d\mu \;=\; \sum_{j=1}^{n} c_j\mu(A_j) \;=\; \sum_{j=1}^{n} c_j\mu\{f=c_j\}\,.$$

(b) Otherwise, $1,\ldots,k]_{\mathbb{Z}} = J_1 \uplus J_2$, where $j\in J_1 \;\Rightarrow\; c_j\geq 0$, and $j\in J_2 \;\Rightarrow\; c_j<0$.
Since $f^+ = \displaystyle\sum_{j\in J_1} c_j\,\mathbf{1}_{\{f=c_j\}}$, and $f^- = \displaystyle\sum_{j\in J_2}(-c_j)\,\mathbf{1}_{\{f=c_j\}}$,

$$\int f\,d\mu \;=\; \int f^+\,d\mu \,-\, \int f^-\,d\mu \;=\; \sum_{j\in J_1} c_j\mu\{f=c_j\} \,-\, \sum_{j\in J_2}(-c_j)\mu\{f=c_j\} \;=\; \sum_{j=1}^{n} c_j\mu\{f=c_j\}\,. \;\square$$

[82] We will call Y in Definition 9.4 (Bernoulli trials and variables) on p.197 a 0–1 encoded Bernoulli trial

[83] Then $P(\Omega) = \displaystyle\sum_{j=1}^{\infty}\left(\frac{1}{2}\right)^j = 1$ (geometric series). We see that indeed the measure P is a probability measure.

Example 6.11. Assume that $f : (\Omega, \mathfrak{F}, \mu) \to \mathbb{R}$ only takes countably many distinct values, $c_1, c_2, \ldots$. We also assume that $c_j \geq 0$, for all j.

$$\text{Then } f = \sum_{j=1}^{\infty} c_j \mathbf{1}_{\{Y=c_j\}}. \quad \text{Thus, } f_n := \sum_{j=1}^{n} c_j \mathbf{1}_{\{f=c_j\}} \uparrow f.$$

By Example 6.10**(a)**, $\int f_n d\mu = \sum_{j=1}^{n} c_j \mu\{f = c_j\}$.

By Definition 6.12 (Abstract integral for measurable functions) on p.155, $\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$.

Thus, as expected, $\int f d\mu = \sum_{j=1}^{\infty} c_j \mu\{f = c_j\}$. □

Example 6.12. Assume that $A^* = \{\omega_1, \omega_2, \ldots\}$ is a countable subset of a set Ω and Σ_* is the counting measure on 2^Ω with respect to A^*: $\mu\{\omega_j\} = 1$ for all j, and $\mu\left(A^{*\complement}\right) = 0$.
Let $f : (\Omega, \mathfrak{F}, \mu) \to \mathbb{R}$ such that $f(\omega_j) \geq 0$ for all j, and $f(\omega) = 0$ for $\omega \notin A^*$. For all j, let $c_j := f(\omega_j)$.

(a) If $c_j \neq 0$ for only finitely many j, say, $c_j = 0$ for $j > n$, then $f = \sum_{j=1}^{n} c_j \mathbf{1}_{\{\omega_j\}}$ is simple. Thus,

$$\int f \, d\Sigma_* = \sum_{j=1}^{n} c_j \, \Sigma_*\{\omega_j\} = \sum_{j=1}^{n} c_j$$

(b) Otherwise ($c_j \neq 0$ for infinitely many j), $f_n := \sum_{j=1}^{n} c_j \mathbf{1}_{\{\omega_j\}} \uparrow f$, and

$$\int f \, d\Sigma_* = \sum_{j=1}^{\infty} c_j \, \Sigma_*\{\omega_j\} = \sum_{j=1}^{\infty} c_j$$

We see that abstract integrals with respect to counting measure simply becomes summation. □

The next theorem and subsequent definition correspond to Theorem 4.4 and Definition 4.7. See pages 93 and 93.

Theorem 6.8. ⋆ *abstract integrals satisfy the following.*

Let $A \in \mathfrak{F}$ and assume that f is $(\mathfrak{F}, \mathfrak{B}^1)$measurable. Then
(a) *If $\int f \, d\mu$ exists, then $\int \mathbf{1}_A f \, d\mu$ exists.*
(b) *If f is μ–integrable, then $\mathbf{1}_A f$ is μ–integrable.*

PROOF: ■

This last theorem allows us to make the following definition.

Definition 6.13. ⋆

Let $A \in \mathfrak{F}$ and assume that f is a measurable function on $(\Omega, \mathfrak{F}, \mu)$, for which the abstract integral $\int f d\mu$ exists. The **abstract integral of f on A** or **over A** is defined by the expression

$$\int_A f\, d\mu \;:=\; \int_A f(\omega) d\,\mu(\omega) \;:=\; \int_A f(\omega)\mu(d\omega) \;:=\; \int \mathbf{1}_A f\, d\mu\,. \tag{6.36}$$

We say that f is **μ–integrable on A**, if $\int_A f\, d\mu$ exists and is finite. □

The next proposition and subsequent theorem correspond to Proposition 4.2 (Integrability criterion) on p.94 and Theorem 4.5 on p.94.

Proposition 6.5 (Integrability criterion). ⋆ *Let f be a measurable function and $A \in \mathfrak{F}$. Then*

$$f \text{ is integrable on } A \;\Leftrightarrow\; \int_A |f|\, d\mu < \infty \;\Leftrightarrow\; \text{both } \int_A f^+\, d\mu < \infty \text{ and } \int_A f^-\, d\mu < \infty.$$

PROOF: ■

Theorem 6.9. ⋆ *Assume that $f, g, f_1, f_2, \ldots$ are measurable functions, $c, c_1, c_2, \cdots \in \mathbb{R}$, and $A \in \mathfrak{F}$. Then μ–integrals on A satisfy the following.*

(a) Positivity: $\displaystyle\int_A 0\, d\mu = 0; \qquad f \geq 0 \text{ on } A \;\Rightarrow\; \int_A f\, d\mu \geq 0,$

(b) Monotonicity: $\mu\{\omega \in A : f(\omega) \geq g(\omega)\} = 0 \;\Rightarrow\; \displaystyle\int_A f\, d\mu \leq \int_A g\, d\mu.$

In particular, $f \leq g \text{ on } A \;\Rightarrow\; \displaystyle\int_A f\, d\mu \leq \int_A g\, d\mu,$

and also, $\mu\{\omega \in A : f(\omega) \neq g(\omega)\} = 0 \;\Rightarrow\; \displaystyle\int_A f\, d\mu = \int_A g\, d\mu.$

(b) Linearity I: f, g *integrable on* $A \;\Rightarrow\; \displaystyle\int_A (f \pm g)\, d\mu = \int_A f\, d\mu \pm \int_A g\, d\mu$

and also, $\displaystyle\int_A (cf)\, d\mu = c \int_A f\, d\mu.$

Linearity II: $f_1 \ldots, f_n$ *integrable* $\;\Rightarrow\; \displaystyle\int_A \Big(\sum_{j=1}^n f_j\Big)\, d\mu = \sum_{j=1}^n c_j \int_A f_j\, d\mu.$

(d) ***Monotone Convergence:*** *Assume that* $0 \le f_1 \le f_2 \le \cdots$, $0 \ge g_1 \ge g_2 \ge \cdots$.

Then $\int_A f_n \, d\mu \uparrow \int_A \left(\sup_{n\in\mathbb{N}} f_n \right) d\mu$ *and* $\int_A g_n \, d\mu \downarrow \int_A \left(\inf_{n\in\mathbb{N}} g_n \right) d\mu$ *as* $n \to \infty$.

(e) ***Dominated Convergence:*** *Assume that*

- $\lim_{n\to\infty} f_n$ *exists,*
- $f_n \le g$ *for all* $n \in \mathbb{N}$,
- $\int_A g \, d\mu < \infty$.

Then $\lim_{n\to\infty} \int_A f_n \, d\mu = \int_A \left(\lim_{n\to\infty} f_n \right) d\mu$ *as* $n \to \infty$.

PROOF: ■

Remark 6.13. ⋆ We recall Fubini's Theorem for Lebesgue integrals (Theorem 4.6 on p.95):

Let $d_1, d_2 \in \mathbb{N}$ and $d := d_1 + d_2$. Let $A_1 \in \mathfrak{B}^{d_1}$ and $A_2 \in \mathfrak{B}^{d_2}$. Note that

(1) $\mathbb{R}^d = \mathbb{R}^{d_1+d_2} = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, **(2)** $A_1 \times A_2 \in \mathfrak{B}^{d_1} \times \mathfrak{B}^{d_2}$, **(3)** $\lambda^d(A_1 \times A_2) = \lambda^{d_1}(A_1) \cdot \lambda^{d_2}(A_2)$.

Let $f : \mathbb{R}^d \to \mathbb{R}$, i.e., $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, be Borel measurable and λ^d–integrable. Then

$$\begin{aligned} \int_{B_1\times B_2} f(\vec{x},\vec{y}) \, \lambda^d\big(d(\vec{x},\vec{y})\big) &= \int_{B_1} \left(\int_{B_2} f(\vec{x},\vec{y}) \, \lambda^{d_2}(d\vec{y}) \right) \lambda^{d_1}(d\vec{x}) \\ &= \int_{B_2} \left(\int_{B_1} f(\vec{x},\vec{y}) \, \lambda^{d_1}(d\vec{x}) \right) \lambda^{d_2}(d\vec{y}) \,. \end{aligned} \tag{6.37}$$

One can show the following relation for the Borel σ–algebras $\mathfrak{B}^{d_1}, \mathfrak{B}^{d_2}$, and $\mathfrak{B}^{d_1+d_2}$:

(6.38) Let $\mathfrak{B}^{d_1} \otimes \mathfrak{B}^{d_2} := \sigma\{A_1 \times A_2 : A_1 \in \mathfrak{B}^{d_1}, A_2 \in \mathfrak{B}^{d_2}\}$. Then $\mathfrak{B}^{d_1} \otimes \mathfrak{B}^{d_2} = \mathfrak{B}^{d_1+d_2}$.

Since "$\times$" occurs in **(3)**, it seems reasonable to replace the symbol λ^d with the symbol $\lambda^{d_1} \times \lambda^{d_2}$:

$$\textbf{(3a)}\ \lambda^d(A_1 \times A_2) = \lambda^{d_1} \times \lambda^{d_2}(A_1 \times A_2) = \lambda^{d_1}(A_1) \cdot \lambda^{d_2}(A_2) \,.$$

This, with **(1) (2)** and (6.38), means that

$$(\mathbb{R}^d, \mathfrak{B}^d, \lambda^d) = (\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \mathfrak{B}^{d_1} \otimes \mathfrak{B}^{d_2}, \lambda^{d_1} \times \lambda^{d_2}) \,. \tag{6.39}$$

The general setting for Fubini's Theorem is obtained as follows. Let

$$(\Omega_1, \mathfrak{F}_1, \mu_1), (\Omega_2, \mathfrak{F}_2, \mu_2)$$

be two measure spaces with σ–finite measures μ_1 and μ_2. [84] We replace

- **(a)** $(\mathbb{R}^{d_1}, \mathfrak{B}^{d_1}, \lambda^{d_1})$ with the measure space $(\Omega_1, \mathfrak{F}_1, \mu_1)$,
  $(\mathbb{R}^{d_2}, \mathfrak{B}^{d_2}, \lambda^{d_2})$ with the measure space $(\Omega_2, \mathfrak{F}_2, \mu_2)$,
- **(b)** the cartesian product $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with the cartesian product $\Omega_1 \times \Omega_2$,
- **(c)** the definition of $\mathfrak{B}^{d_1} \otimes \mathfrak{B}^{d_2}$ in (6.38) with the definition

$$\mathfrak{F}_1 \otimes \mathfrak{F}_2 := \sigma\{A_1 \times A_2 : A_1 \in \mathfrak{F}_1, A_2 \in \mathfrak{F}_2\} \,, \tag{6.40}$$

[84] σ–finiteness is a very technical condition. It means that one can find two sequences $A_n \in \mathfrak{F}_1$ and $B_n \in \mathfrak{F}_2$ such that $\mu(A_n) < \infty$ and $\mu(B_n) < \infty$ for all n, and $\Omega_1 = \bigcup_n A_n$, $\Omega_2 = \bigcup_n B_n$. See Definition 6.8 (Abstract measures) on p.148.

(d) the "product measure" $\lambda^{d_1} \times \lambda^{d_2}(A_1 \times A_2) = \lambda^{d_1}(A_1) \cdot \lambda^{d_2}(A_2)$ with the measure

$$\mu_1 \times \mu_2 : \mathfrak{F}_1 \otimes \mathfrak{F}_2 \longrightarrow [0,\infty]\,; \qquad A_1 \times A_2 \mapsto \mu_1 \times \mu_2(A_1 \times A_2) := \mu_1(A_1) \cdot \mu_2(A_2)\,. \tag{6.41}$$

Here, $A_1 \in \mathfrak{F}_1$ and $A_2 \in \mathfrak{F}_2$. Thus, (6.41) defines $\mu_1 \times \mu_2(A)$ only for measurable rectangles, $A_1 \times A_2$. However, one can show that $\mu_1 \times \mu_2$ can be extended to a measure on all of $\mathfrak{F}_1 \otimes \mathfrak{F}_2$, and that this extension is unique. □

Definition 6.14 (Product measure space). ⋆ The following is based on Remark 6.13.

Let $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_1, \mathfrak{F}_1, \mu_1)$ be measure spaces with σ–finite measures μ_1, μ_2. Let

$$\mathfrak{F}_1 \otimes \mathfrak{F}_2 := \sigma\{A_1 \times A_2 : A_1 \in \mathfrak{F}_1,\ A_2 \in \mathfrak{F}_2\}\,. \tag{6.42}$$

Let $\mu_1 \times \mu_2 : \mathfrak{F}_1 \otimes \mathfrak{F}_2 \longrightarrow [0,\infty]$ be the measure which is uniquely determined by

$$\mu_1 \times \mu_2(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2)\,, \quad \text{for } A_1 \in \mathfrak{F}_1 \text{ and } A_2 \in \mathfrak{F}_2. \tag{6.43}$$

We call the measure space $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mu_1 \times \mu_2)$ the **product measure space** aka **product space** of the factors $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_1, \mathfrak{F}_1, \mu_1)$, $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ the **product σ–algebra** of the factors $\mathfrak{F}_1$ and $\mathfrak{F}_2$, and $\mu_1 \times \mu_2$ the **product measure** of the factors μ_1 and μ_2.

There are alternate ways to denote integrals with respect to $\mu_1 \times \mu_2$.

$$\begin{aligned} \int f\, d\,\mu_1 \times \mu_2 &= \int f(\omega_1,\omega_2)\, d\,\mu_1 \times \mu_2(\omega_1,\omega_2) \\ &= \int f(\omega_1,\omega_2)\, \mu_1 \times \mu_2\big(d(\omega_1,\omega_2)\big) = \int f(\omega_1,\omega_2)\, \mu_1 \times \mu_2(d\omega_1, d\omega_2) \end{aligned} \tag{6.44}$$

See (6.32) and (6.36). □

Theorem 6.10 (Fubini's theorem for abstract integrals). ⋆

Let $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_1, \mathfrak{F}_1, \mu_1)$ be measure spaces with σ–finite measures μ_1, μ_2. Let

$$f : (\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2,\ \mu_1 \times \mu_2) \longrightarrow \mathbb{R}; \quad (\omega_1,\omega_2) \mapsto f(\omega_1,\omega_2), \quad \text{be } \mathfrak{F}_1 \otimes \mathfrak{F}_2\text{–measurable.}$$

Assume that f is nonnegative and/or $(\mu_1 \times \mu_2)$–integrable, and that $A_1 \in \mathfrak{F}_1$, $A_2 \in \mathfrak{F}_2$. Then

$$\begin{aligned} \int_{A_1 \times A_2} f\, d\mu_1 \times \mu_2 &= \int_{A_1} \left(\int_{A_2} f\, d\mu_2 \right) d\mu_1 \\ &= \int_{A_2} \left(\int_{A_1} f\, d\mu_1 \right) d\mu_2\,. \end{aligned} \tag{6.45}$$

When we supply the arguments, ω_1 and ω_2, (6.45) reads

$$\text{(6.46)}\qquad \int_{A_1\times A_2} f(\omega_1,\omega_2)\,\mu_1\times\mu_2\big(d(\omega_1,\omega_2)\big) = \int_{A_1}\left(\int_{A_2} f(\omega_1,\omega_2)\,\mu_{d_2}(d\omega_2)\right)\mu_{d_1}(d\omega_1) = \int_{A_2}\left(\int_{A_1} f(\omega_1,\omega_2)\,\mu_{d_1}(d\omega_1)\right)\mu_{d_2}(d\omega_2)\,.$$

PROOF: ■

Remark 6.14. ⋆ One can defined product measure spaces

$$(\Omega_1\times\cdots\times\Omega_n, \mathfrak{F}_1\otimes\cdots\otimes\mathfrak{F}_n,\ \mu_1\times\cdots\times\mu_2)$$

and Fubini's theorem for more than two factors. □

6.4 The ILMD Method

Introduction 6.4. ⋆

The abstract integral was defined or computed in the following stages:

(2) For simple functions $f(\omega) = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}(\omega)$, we defined $\int f\,d\mu = \sum_{j=1}^{n} c_j\mu(A_j)\,.$

(3) For any nonnegative (measurable) function f, choose simple functions $0\le f_n\uparrow f$. By monotone convergence, $\int f\,d\mu = \lim_{n\to\infty}\int f_n\,d\mu.$

(4) For arbitrary (measurable) $f=f^+-f^-$ such that $\int f^+\,d\mu<\infty$ or $\int f^-\,d\mu<\infty$, we defined $\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu\,.$

Note that replacing f and f_n with $f\mathbf{1}_A$ and $f_n\mathbf{1}_A$, $A\in\mathfrak{F}$, also covers $\int_A\cdots d\mu$.

Why is **(1)** missing? We reserve that case for particularly simple simple functions, the indicator functions. We could have preceded Definition 6.11 (Abstract integral for simple functions) on p.155, which handles **(2)**, by the following.

(1) For $A\in\mathfrak{F}$, define $\int \mathbf{1}_A\,d\mu = \mu(A)\,.$

This section describes a general method for proving statements that are about integrals. □

Remark 6.15 (The ILMD Mehod)**.** ⋆ If one wants to prove a theorem in which integration plays a central role, the following procedure, which we call the **ILMD method**, [85] often is successful.

[85] When googling the phrase "ILMD Mehod", the author found the following result:

- The Improved Local Mean Decomposition (ILMD) is employed to decompose remanufacturing cost time series data into several components with smooth, periodic fluctuation and use this as input.

So be sure to explain the term when you use it in discussions with others! Other authors use different terms. For example, [11] Shreve, Steve: Stochastic Calculus for Finance II: Continuous-Time Models refers to the ILMD method as the "Standard Machine".

I Prove the statement for integrands which are **I**ndicator functions $\mathbf{1}_A(\omega)$.
L **L**inearity of the integral (Theorem 6.9**(b)** on p.159 often extends the result to simple functions at little or no cost.
M **M**onotone convergence (Theorem 6.9**(c)** on p.159 often extends the result to non-negative integrands at little or no cost.
D Writing f as the **D**ifference of two nonnegative functions, e.g., $f = f^+ - f^-$, extends the result to general integrands. This might prove more difficult than the preceding two steps, since expressions of the form $\infty - \infty$ must be avoided. □

The proof of the next theorem demonstrates the usefulness of ILMD. We state again that you can ignore the σ–algebras and assume that every function is measurable and $\mu(A)$ is defined for every set A.

Theorem 6.11 (Integrals under Transforms). ⋆ *Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and let $(\Omega', \mathfrak{F}')$ be a measurable space. Assume that $f : \Omega \to \Omega'$ is $(\mathfrak{F}, \mathfrak{F}')$–measurable. and $g : \Omega' \to \mathbb{R}$ is $(\mathfrak{F}', \mathfrak{B}^1)$–measurable. μ_f denotes the image measure of μ under f on $\mathfrak{F}'$. It was defined in Definition 6.10 on p.153 as*

$$\mu_f(A') \;=\; \mu\{f \in A'\} \;=\; \mu\big(f^{-1}(A')\big)\,.$$

If $g \geq 0$ or $g \circ f$ is integrable then

(6.47) $$\int g \circ f \, d\mu \;=\; \int g \, d\mu_f, \quad \textit{i.e.,} \quad \int_\Omega g\big(f(\omega)\big)\, d\mu(\omega) \;=\; \int_{\Omega'} g(\omega')\, d\mu_f(\omega')\,.$$

PROOF:

Step 1. Assume that $g = \mathbf{1}_{A'}$ for some $A' \in \mathfrak{F}'$. Note that

$$\mathbf{1}_{A'}\big(f(\omega)\big) \;=\; 1 \;\Leftrightarrow\; f(\omega) \in A' \;\Leftrightarrow\; \omega \in f^{-1}(A')\,.$$

Thus,

$$\int_\Omega \mathbf{1}_{A'}\big(f(\omega)\big)\, d\mu(\omega) \;=\; \int_\Omega \mathbf{1}_{f^{-1}(A')}(\omega)\, d\mu(\omega) \;=\; \mu\big(f^{-1}(A')\big) \;=\; \mu_f(A') \;=\; \int_{\Omega'} \mathbf{1}_{A'}(\omega')\, d\mu_f(\omega').$$

This proves (6.47) for $g = \mathbf{1}_{A'}$.

Step 2. Let $g \geq 0$ be a simple function $g = \sum_{j=1}^n c_j \mathbf{1}_{A'_j}$ ($n \in \mathbb{N}, c_j \geq 0, A_j \in \mathfrak{F}$). It then follows from the linearity of the integral and what we obtained in step 1, that

$$\int_\Omega g \circ f \, d\mu \;=\; \sum_{j=1}^n c_j \int_\Omega \mathbf{1}_{A'_j} \circ f \, d\mu \;=\; \sum_{j=1}^n c_j \int_{\Omega'} \mathbf{1}_{A'_j}\, d\mu_f \;=\; \int_{\Omega'} g \, d\mu_f.$$

Step 3. Assume that g is a nonnegative, $(\mathfrak{F}', \mathfrak{B}^1)$–measurable function. Let $(g_n)_n$ be a sequence of simple functions such that $g_n \uparrow g$. By **Step 2** and the monotone convergence property,

$$\int_\Omega g \circ f \, d\mu \;=\; \lim_{n\to\infty} \int_\Omega g_n \circ f \, d\mu \;=\; \lim_{n\to\infty} \int_{\Omega'} g_n \, d\mu_f \;=\; \int_{\Omega'} g \, d\mu_f.$$

Step 4. Since the proof is finished for $g \geq 0$, we may From now on assume that $g \circ f$ is μ–integrable, i.e., both $\int (g \circ f)^+ d\mu < \infty$ and $\int (g \circ f)^- d\mu < \infty$. We have shown in step 3 that the nonnegative functions $g^+ \circ f$ and $g^- \circ f$ satisfy

$$\int_\Omega g^+ \circ f \, d\mu \;=\; \int_{\Omega'} g^+ \, d\mu_f, \qquad \int_\Omega g^- \circ f \, d\mu \;=\; \int_{\Omega'} g^- \, d\mu_f, \tag{6.48}$$

We also have

$$\begin{aligned} (g^+ \circ f)(\omega) \;&=\; g^+\big(f(\omega)\big) \;=\; \big[g\big(f(\omega)\big)\big]^+ \;=\; \big[(g \circ f)(\omega)\big]^+ \;=\; (g \circ f)^+(\omega), \\ (g^- \circ f)(\omega) \;&=\; g^-\big(f(\omega)\big) \;=\; \big[g\big(f(\omega)\big)\big]^- \;=\; \big[(g \circ f)(\omega)\big]^- \;=\; (g \circ f)^-(\omega). \end{aligned} \tag{6.49}$$

It follows that

$$\begin{aligned} \int_\Omega |g \circ f| \, d\mu \;&=\; \int_\Omega (g \circ f)^+ \, d\mu \;+\; \int_\Omega (g \circ f)^- \, d\mu \\ &\overset{(6.49)}{=} \int_\Omega (g^+ \circ f) \, d\mu \;+\; \int_\Omega (g^- \circ f) \, d\mu \\ &\overset{(6.48)}{=} \int_{\Omega'} g^+ \, d\mu_f \;+\; \int_{\Omega'} g^- \, d\mu_f. \end{aligned}$$

All quantities here are finite since $\int (g \circ f)^+ d\mu < \infty$ and $\int (g \circ f)^- d\mu < \infty$. We thus may subtract and obtain

$$\int_\Omega g \circ f \, d\mu \;=\; \int_{\Omega'} g^+ \, d\mu_f \;-\; \int_{\Omega'} g^- \, d\mu_f. \;\blacksquare$$

We also use the ILMD method to prove the next theorem.

Theorem 6.12. ⋆

Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and let f be a nonnegative, real–valued, Borel–measurable function on $(\Omega, \mathfrak{F}, \mu)$. Let ν be the measure defined by

$$\nu(A) \;:=\; \int_A f \, d\mu \tag{6.50}$$

(see Theorem 6.6 on p.151). Further, let φ be a real–valued, Borel–measurable function on Ω, such that $\varphi \geq 0$ or φ is ν–integrable. Then

$$\int_A \varphi \, d\nu \;=\; \int_A \varphi \cdot f \, d\mu, \;\; \textit{i.e.,} \;\; \int_A \varphi(\omega) \, \nu(d\omega) \;=\; \int_A \varphi(\omega) \, f(\omega) \, \mu(d\omega); \;\; A \in \mathfrak{F}. \tag{6.51}$$

PROOF:

Step 1. We prove formula (6.51) for indicator functions. Assume that $\varphi = \mathbf{1}_B$ for some $B \in \mathfrak{F}$. Then

$$\begin{aligned} \int_A \varphi \, d\nu \;&=\; \int \mathbf{1}_A \mathbf{1}_B \, d\nu \;=\; \int \mathbf{1}_{A \cap B} \, d\nu \;=\; \nu(A \cap B) \\ &=\; \int_{A \cap B} f \, d\mu \;=\; \int \mathbf{1}_A \mathbf{1}_B f \, d\mu \;=\; \int_A \mathbf{1}_B f \, d\mu \;=\; \int_A \varphi f \, d\mu \,. \end{aligned}$$

Thus, (6.51) holds for $\varphi = \mathbf{1}_B$.

Step 2. linearity of the integral allows to extend the formula from indicator functions to simple functions $\varphi = \sum_{j=1}^{n} c_j \mathbf{1}_{A_j}$ ($n \in \mathbb{N}, c_j \geq 0, A_j \in \mathfrak{F}$).

Step 3. Assume that φ is a nonnegative, $\mathfrak{F} - -\mathfrak{B}^1$ measurable function. $0 \leq \varphi_n \uparrow \varphi$ a sequence of simple functions. By the monotone convergence property, (6.51) is true for φ.

Step 4. Since the proof is done for $\varphi \geq 0$, we now assume that $\varphi = \varphi^+ - \varphi^-$ is ν–integrable. By linearity, the integral of a difference is the difference of the integrals. We obtain

$$\begin{aligned}\int_A \varphi\, d\nu &= \int_A \varphi^+ \, d\nu - \int_A \varphi^- \, d\nu \overset{\text{Step 3}}{=} \int_A \varphi^+ \cdot f \, d\mu - \int_A \varphi^- \cdot f \, d\mu \\ &= \int_A (\varphi^+ - \varphi^-) \cdot f \, d\mu = \int_A \varphi \cdot f \, d\mu \,. \quad \blacksquare\end{aligned}$$

6.5 Expectation and Variance as Probability Measure Integrals

Introduction 6.5. ⋆

We have defined the abstract integral $\int Y dP$ for random variables Y defined on any kind of probability space, $(\Omega, \mathfrak{F}, P)$. We will attach some meaning to this expression as an average of sorts, and why it will be called the expected value aka expectation of Y. We will also do this for the variance of Y. This characteristic of Y is defined as the integral $\int g \circ Y dP = \int g\big(Y(\omega)\big) P(d\omega)$, where $g : \mathbb{R} \to \mathbb{R}$ is the function $g(y) = (y - \int Y dP)^2$. □

Definition 6.15 (Expected value of a random variable)**.**

Let Y be a random variable on a probability space $(\Omega, , P)$.

(a) We call

$$E[Y] := \int Y \, dP \tag{6.52}$$

the **expected value**, also **expectation** or **mean** of Y.

(b) We call

$$Var[Y] := E\big[(Y - E[Y])^2\big] = \int (Y - E[Y])^2 \, dP \tag{6.53}$$

the **variance**, of Y.

(c) We call $SD[Y] := \sigma_Y := \sqrt{Var[Y]}$ The **standard deviation** of Y. □

Remark 6.16. ⋆ Some notes on notation.

- It is OK to write $E(Y), Var(Y), SD(Y)$ for $E[Y], Var[Y], SD[Y]$. This could even be a very good idea if to have an expression with nested brackets and use alternating delimiters to make it easier to "unpack" that expression.
- The use of $\sigma(Y)$ and $\sigma[Y]$ rather than subscripting σ_Y is discouraged since this might lead to confusion with $\sigma\{Y\}$, the σ–algebra generated by Y. By the way, this is the reason why this author chose $\sigma\{Y\}$ rather than $\sigma(Y)$, a symbol that is quite popular to denote the σ–algebra generated by Y. $\square$

You may recall from integral calculus the following mean value theorem. If f is (Riemann) integrable on $[a,b]$ and $\alpha, \beta \in \mathbb{R}$ such that $\alpha \le f(x) \le \beta$ for $a \le x \le b$, then there is $\alpha \le \gamma \le \beta$ such that

$$\gamma \;=\; \frac{1}{b-a}\int_a^b f(t)dt \;=\; \frac{1}{\lambda^1[a,b]}\int_{[a,b]} f d\lambda^1\,.$$

The meaning is intuitively clear, at least if $f \ge 0$. We rewrite this equation

$$\int_a^b f(t)dt \;=\; \gamma\cdot(b-a)$$

and we see that γ is determined by having the area between the graph of f, the horizontal axis, and the vertical lines through a and b equal to the area of a rectangle of width $b-a$ and height $\gamma - 0 = \gamma$. In that sense, $\int_{[a,b]} f d\lambda^1/\lambda^1[a,b]$ is a good middle value or mean for the values that f can take.

Remark 6.17. ⋆ We can generalize what we just mentioned to any measure μ instead of λ^1 and set $A \in \Omega$ instead of an interval $[a,b] \subseteq \mathbb{R}$ Assuming that $0 < \mu(A) < \infty$,

$$\frac{1}{\mu(A)}\int_A f d\mu$$

is a good middle value for f. In particular, if we have a probability measure P, a random variable Y, and the event $A = \Omega$, then

$$\frac{1}{P(\Omega)}\int_\Omega Y dP \;=\; E[Y]$$

is a good mean value for the random variable Y.

Remark 6.18. ⋆ Let us assume that Y is a discrete random variable. In other words, there is a countable set $B^* \subseteq \mathbb{R}$ such that $P_Y(B^*) = P\big(Y^{-1}(B^*)\big) = 1)$. See, e.g., Proposition 5.5 on p.127. For $y \in B^*$, let $A_y := \{Y = y\}$. Then $Y(\omega) = \sum\limits_{y\in B^*} y\mathbf{1}_{A_y}(\omega)$, for $\omega \in A^* := \biguplus\limits_{y\in B^*} A_y$.

Since $P\big((A^*)^\complement\big) = 0$, $\int_\Omega \cdots dP = \int_{A^*} \cdots dP$. We just saw that $Y = \sum\limits_{y\in B^*} y\mathbf{1}_{A_y}$ on A^*. Thus,

$$E[Y] \;=\; \int Y dP \;=\; \int\Big(\sum_{y\in B^*} y\mathbf{1}_{A_y}\Big) \;=\; \sum_{y\in B^*} yP\big(Y^{-1}\{y\}\big) \;=\; \sum_{y\in B^*} yP_Y\{y\}\,. \tag{6.54}$$

Let $g : \mathbb{R} \to \mathbb{R}$ be a Borel function which is nonnegative or such that $\int g\circ Y dP < \infty$.
Since $P_Y\big((B^*)^\complement\big) = 0$ and $g(\tilde{y}) = \sum\limits_{y\in B^*} \mathbf{1}_{\{y\}}(\tilde{y})$, for $\tilde{y} \in B^*$, we see that

$$\int g\,dP_Y \;=\; \int_{B^*} g\;dP_Y \;=\; \int \Big(\sum_{y\in B^*} g(y)\mathbf{1}_{\{y\}}\Big)\,dP_Y\,. \tag{6.55}$$

We apply Theorem 6.11 (Integrals under Transforms) on p.163 to We obtain

$$E[g\circ Y] \;=\; \int g\circ Y dP \;=\; \int g\,dP_Y \;=\; \int \Big(\sum_{y\in B^*} g(y)\mathbf{1}_{\{y\}}\Big)\,dP_Y \overset{(6.55)}{=} \sum_{y\in B^*} g(y)P_Y\{y\}\,. \tag{6.56}$$

In particular, if $g(y)=(y-E[Y])^2$,

$$Var[Y] \;=\; \int (Y-E[Y])^2 dP \;=\; \sum_{y\in B^*} (y-E[Y])^2 P_Y\{y\}\,. \;\square \tag{6.57}$$

Remark 6.19. ⋆ In Chapter 10 (Continuous Random Variables), a continuous random variable Y on a probability space $(\Omega,\mathfrak{F},P)$ will be defined as one which possesses a "density". A density for Y is a Borel function

$$f_Y:\mathbb{R}\longrightarrow\mathbb{R};\qquad y\mapsto f_Y(y)\,,$$

which, for all intervals $]a,b]$, satisfies

$$P_Y(]a,b]) \;=\; P\{a<Y\le b\} \;=\; \int_a^b f_Y(y)\,dy \;=\; \int_{]a,b]} f_Y\,d\lambda^1\,. \quad\text{and thus,}\quad P_Y(B) \;=\; \int_{]a,b]} f_Y\,d\lambda^1\,,$$

Since the distribution P_Y is uniquely determined by those values $P_Y(]a,b])$, it is the measure

$$P_Y:\mathfrak{B}^1\to[0,1]\,;\qquad B\mapsto P_Y(B) \;=\; \int_B f_Y\,d\lambda^1\,,\;\; (B\in\mathfrak{B}^1)\,.$$

By Theorem 6.12 on p.164,

$$\int g\,dP_Y \;=\; \int g\cdot f\,d\lambda^1 \;=\; \int_{-\infty}^{\infty} g(y)f_Y(y)\,dy\,. \tag{6.58}$$

We apply Theorem 6.11 (Integrals under Transforms) on p.163 and obtain

$$E[g\circ Y] \;=\; \int g\circ Y dP \;=\; \int g\,dP_Y \overset{(6.58)}{=} \int_{-\infty}^{\infty} g(y)f_Y(y)\,dy\,. \;\square \tag{6.59}$$

7 Combinatorial Analysis

In many important cases we find ourselves in the situation of Example 5.6 on p.107, where we have a finite probability space (Ω, P), in which each outcome $\omega \in \Omega$ as equal probability

$$P\{\omega\} \;=\; \frac{1}{|\Omega|}$$

and thus, for each event $A \subset \Omega$,

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

Hence, all we need to determine $P(A)$, is the knowledge of how to count the elements of Ω and of A. Combinatorial analysis, also called **combinatorics**, , is a branch of mathematics that provides us with tools to accomplish that task.

7.1 The Multiplication Rule

The first result is known under names such as the basic principle of counting ([8] Ross, Sheldon M.: A First Course in Probability, 3rd edition) and the mn rule (WMS text).

Theorem 7.1 (Multiplication rule)**.**

(A) *Assume that two actions A and B are performed such that*

- *the first one has m outcomes, $\{a_1, a_2, \ldots, a_m\}$,*
- *the second one has n outcomes $\{b_1, b_2, \ldots, b_n\}$ for each outcome of the first one.*
- *Then the number of combined outcomes (a_i, b_j) is mn.*

(B) *Generalization. Assume that k actions $A_1, \ldots, A_k$ are performed such that*

- *action A_1 has n_1 outcomes, $\{a_1^{(1)}, a_2^{(1)}, \ldots, a_{n_1}^{(1)}\}$,*
- *action A_2 has n_2 outcomes, $\{a_1^{(2)}, a_2^{(2)}, \ldots, a_{n_2}^{(2)}\}$ for each outcome of A_1,*
- *action A_3 has n_3 outcomes, $\{a_1^{(3)}, a_2^{(3)}, \ldots, a_{n_3}^{(3)}\}$ for each combined outcome (x_1, x_2), where x_1 is one of the A_1–outcomes and x_2 is one of the A_2–outcomes,*

  -

- *action A_k has n_k outcomes, $\{a_1^{(k)}, a_2^{(k)}, \ldots, a_{n_k}^{(k)}\}$ for each combined outcome (x_1, x_2, x_{k-1}), where each x_j is one of the A_j–outcomes, i.e., x_j is one of $a_1^{(j)}, \ldots, a_{n_j}^{(j)}$.*

- *Then there are $n_1 \cdot n_2 \cdots n_k$ combined outcomes $(x_1, x_2, \ldots, x_k)$.*
  Here, each x_j is one of the n_j outcomes $a_1^{(j)}, \ldots, a_{n_j}^{(j)}$ of A_j.

PROOF: We identify the actions with their outcomes, i.e., we define

$$A_j \;=\; \{a_1^{(j)}, \ldots, a_{n_j}^{(j)}\}, \quad \text{for } \; j = 1, 2, \ldots, k.$$

Now, the multiplication rule merely states that $\big|\, A_1 \times A_2 \times \cdots \times A_n \,\big| \;=\; \big|A_1\big| \cdot \big|A_2\big| \;\cdots\; \big|A_n\big|$, and this is true according to (2.61) on p.52. ■

Example 7.1 (Ross-prob-thy-3ed Example 2c)**.** How many 7–digit license plates can be created if the first three are letters (CAPS) and the lst four are digits?

Answer: $26^3 \cdot 10^4 = 175,760,000$ □

Example 7.2 (Ross-prob-thy-3ed Example 2e)**.** How many different 7–digit license plates can be created if the first three are letters (CAPS) and the last four are digits and none of those symbols can be repeated?

Answer: $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$ □

Example 7.3. How many 7–digit license plates can be created if the first three are letters (CAPS) and the last four are digits and none of the letters can be repeated, but the digits can be repeated?

Answer: $26 \cdot 25 \cdot 24 \cdot 10^4 = 26 \cdot 600 \cdot 10^4 = 15,600 \cdot 10^4 = 15,600,000.$ □

Example 7.4 (Ross-prob-thy-3ed Example 2d)**.** If $\big|\Omega\big| = n$, how many different functions $\psi : \Omega \to \{0,1\}$, i.e., how many functions on Ω that can only take the values 0 and 1, do exist?

Answer: If $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, then

- we have 2 choices for the $\psi(\omega_1)$ selection.
- For each choice of $\psi(\omega_1)$, there are 2 choices for the $\psi(\omega_2)$ selection.
- For each choice of those $\psi(\omega_1) and \psi(\omega_2)$, there are 2 choices for the $\psi(\omega_3)$ selection.
- -
- For each choice of $\psi(\omega_1), \ldots, \psi(\omega_{n-1})$, there are 2 choices for the $\psi(\omega_n)$ selection.

So we have a total of $2 \cdot 2 \cdots 2 = 2^n$ selections. □

Example 7.5. If $\big|\Omega\big| = n$, how many subsets of Ω, including $\emptyset$ and Ω, do exist?

Answer: If $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, any subset $A \subseteq \Omega$ can be uniquely represented by an element $\vec{d} = \vec{d}(A) = (d_1, d_2, \ldots, d_n)$ of $\{0,1\}^n$ as follows:

- $d_j = 1 \Leftrightarrow \omega_j \in A$ and $d_j = 0 \Leftrightarrow \omega_j \notin A$.

The assignment $F : A \mapsto \vec{d}(A)$ between the subsets of Ω and $\{0,1\}^n$ is injective:

- If $A' \subseteq \Omega$ such that $\vec{d}(A) = \vec{d}(A')$, then $\omega \in A \Leftrightarrow \omega \in A'$, i.e., $A = A'$.

F also is surjective: if $\vec{d}(d_1, d_2, \ldots, d_n) \in \{0,1\}^n$, then

- $B := \{\omega_j : d_j = 1\}$ (a subset of Ω) which satisfies $F(A) = \vec{d}$.

Thus, F is a bijection. We illustrate this with the following example. Let $\Omega := \{\omega_1, \omega_2, \omega_3, \omega_4\}$.

- $A_1 = \{\omega_2, \omega_3\} \Rightarrow F(A_1) = (0,1,1,0)$. Also, $F^{-1}(0,1,1,0) = \{\omega_j : d_j = 1\} = \{\omega_2, \omega_3\} = A_1$.
- $A_2 = \{\omega_4\} \Rightarrow F(A_2) = (0,0,0,1)$. Also, $F^{-1}(0,0,0,1) = \{\omega_j : d_j = 1\} = \{\omega_4\} = A_2$

Since F is a bijection, there are as many subsets of Ω as there are vectors

$\vec{d}(A) = (d_1, d_2, \ldots, d_n)$ of zeros and ones of length n. And how many are those?

- we have 2 choices for d_1: either $d_1 = 0$ or $d_1 = 1$.
- For each of those choices: either $d_2 = 0$ or $d_2 = 1$.
- -
- For each of those 2^{n-1} choices $\big[\, d_j = 0$ or $d_j = 1$ $(j = 1, 2, \ldots, n-1)\,\big]$: either $d_n = 0$ or $d_n = 1$.

Thus, we have $2 \cdot 2 \cdots 2 = 2^n$ choices. □

7.2 Permutations

Definition 7.1 (WMS Ch.02.6, Definition 2.7 - Permutation)**.**

An ordered arrangement of r distinct objects is called a **permutation** of size r. The number of ways of ordering n distinct objects taken r at a time will be designated by the symbol P_r^n. □

Theorem 7.2 (WMS Ch.02.6, Theorem 2.2)**.**

$$P_r^n = n(n-1)(n-2)\cdot(n-r+1) = \frac{n!}{(n-r)!}. \tag{7.1}$$

Here, $n!$ ("n factorial") is defined as follows.

$$n! = \begin{cases} n(n-1)\cdots 2\cdot 1, & \text{if } n \in \mathbb{N}, \\ 1, & \text{if } n = 0. \end{cases} \tag{7.2}$$

PROOF: We can consider each permutation as the result of the following actions $A_1, \ldots, A_r$.

- A_1 is the selection of the first item. Since all n items are available for selection, A_1 has n outcomes.
- A_2 is the selection of the second item. Since one item was already selected and duplicates are not allowed, only $n-1$ items are available for selection. Thus, A_2 has $n-1$ outcomes.
- -
- A_r is the selection of item r. Since $r-1$ items have been previously selected and duplicates are not allowed, only $n-(r-1) = n-r+1$ items are available for selection. Thus, A_r has $n-r+1$ outcomes.

It follows from the multiplication rule that there are $n(n-1)\cdots(n-r+1)$ different ways to select r items without repeating a selection, i.e., of obtaining a permutation of size r of those n items. ■

Problem 7.1 (WMS Ch.02.8, Example 2.8)**.** The names of 3 employees are to be randomly drawn, without replacement, from a bowl containing the names of 30 employees of a small company. The person whose name is drawn first receives \$100, and the individuals whose names are drawn second and third receive \$50 and \$25, respectively. How many sample points are associated with this experiment?

Solution: Because the prizes awarded are different, the number of sample points is the number of ordered arrangements of $r = 3$ out of the possible $n = 30$ names. Thus, the number of sample points in S is

$$P_3^{30} = \frac{30!}{27!} = (30)(29)(28) = 24,360. \ \square$$

Example 7.6. Jenny has collected 20 post cards, all of them different:

- 4 from France, • 2 from Peru, • 8 from Japan, • 6 from Kenia.

She wants to place them into 4 numbered boxes according to their country of origin.

(A) Jenny considers two arrangements different if, say, Esteban's card takes a different spot in the Peru box, but she does not care whether the Peru cards end up in box #1 or #2 or #3 or #4. How many different arrangements are possible?

Answer:

- 4 choices for France card #1,
- 3 choices for France card #2 (into the same box),
- 2 choices for France card #3 (into the same box),
- 1 choice for France card #4 (into the same box).
- Thus, there are 4! choices for the France cards.
- For each one of those 4! choices we obtain in a similar manner that there are 2! choices for Peru.
- For each one of those $4! \cdot 2!$ choices we obtain in a similar manner that there are 8! choices for Japan.
- For each one of those $4! \cdot 2! \cdot 8!$ choices we obtain in a similar manner that there are 6! choices for Kenia.

Thus, $4! \cdot 2! \cdot 8! \cdot 6!$ different arrangements are possible.

(B) As before, Jenny considers two arrangements different if, say, Esteban's card takes a different spot in the Peru box. But this time it also matters in which box a country's cards are placed.. How many different arrangements are possible now?

Answer: There are 4! permutations of the 4 boxes. This amounts to 4! rearrangements of each choice made in **(A)**. Thus, $4! \cdot 2! \cdot 8! \cdot 6! \cdot 4!$ arrangements are possible. □

7.3 Combinations, Binomial and Multinomial Coefficients

In Example 7.5 on p.169, a simple application of the multiplication rule showed the following:
If Ω is a set of finite size, then its powerset 2^Ω (i.e., the set of all subsets of Ω), has size $\left|2^\Omega\right| = 2^{|\Omega|}$.
A related question would be the following:

- How many subsets of Ω have size k?

Examining how many permutations of size k can be obtained from the elements $\omega_1, \omega_2, \ldots, \omega_n$ might not be a bad idea, since permutations of distinct items remain free of duplicates, just as we require for (sub–)sets. But rearrangements of the order in which the elements $\omega_{n_1}, \omega_{n_2}, \ldots, \omega_{n_k}$ of such a subset lead to different permutations although the subset remains the same, since the order of the elements of a set is disregarded.

Thus, we must divide P_k^n, the number of permutations of size k of the elements of Ω, by the number of rearrangements that one can obtain from a given set of its members. Since that number is $P_k^k = k!$, we have obtained the following result.

Theorem 7.3.

Let $0 \le k \le n$. A set of size n has

$$\frac{n!}{k!(n-k)!}\,.$$

subsets of size k.

PROOF: There are $P_k^n = n(n-1)\cdots,(n-k+1)$ permutations of size k that can be obtained from the n (distinct!) elements $\omega_1, \omega_2, \ldots, \omega_n$ of Ω. Let $A := \{\omega_{n_1}, \omega_{n_2}, \ldots, \omega_{n_k}\}$ be such a permutation.

There are $P_k^k = k!$ rearrangements of $\omega_{n_1}, \omega_{n_2}, \ldots, \omega_{n_k}$. Since order does not matter in sets (and their subsets), each one of those $k!$ permutations forms one and the same set A.

To say this differently, the number P_k^n was obtained by counting each size k subset $k!$ times.

Thus, we must divide P_k^n by P_k^k to obtain the number of subsets of size k. We obtain

$$\frac{P_k^n}{P_k^k} = \frac{n(n-1)\cdots(n+k-1)}{k!} = \frac{n(n-1)\cdots\big(n-(k-1)\big)}{k!}\cdot\frac{(n-k)!}{(n-k)!} = \frac{n!}{k!(n-k)!}\,.$$

This proves the theorem. ■

Selections of size k from a collection of n distinct objects disregarding the order in which those k items were selected (as is the case when selecting a subset of size k from a set of size $n \ge k$,) are so important when counting is involved that they deserve a name of their own. For the following see also WMS Ch.02.6, Definition 2.8.

Definition 7.2 (Number of combinations)**.**

We call the number of selections of size k from a collection of n distinct items when the order in which those k items were selected is ignored, the **number of combinations of n objects taken k at a time**. We write $\binom{n}{k}$ for this number. □

Remark 7.1.

(a) Some texts also use the symbol C_k^n instead of $\binom{n}{k}$. This is considered outdated terminology.

(b) We emphasize that both are true: $\binom{n}{k}$
$=$ number of selections of size k from n distinct items when disregarding order
$=$ number of subsets of size k of a set of size n. □

Theorem 7.4.

Given are n items of which n_1 are alike, n_2 are alike, $\ldots$, n_r are alike ($n_1 + \cdots + n_r = n$).
Then the number of distinguishable arrangements of those n items is

$$\binom{n}{n_1, n_2, \ldots n_r} = \frac{n!}{n_1!\, n_2! \cdots n_r!}\,.$$

PROOF:

- We tag the group 1 items as $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}$,
- the group 2 items as $x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)}$,
- -
- the group r items as $x_1^{(r)}, x_2^{(r)}, \ldots, x_{n_r}^{(r)}$,

to make all n items artificially distinguishable. We have learned that there are $n!$ permutations.

When we only keep the superscripts that indicate the group but we remove the subscripts, since in truth items belonging the same group cannot be distinguished, there will be a lot less arrangements that are distinct.

To fix the ideas, assume that group 2 has 4 members and we have an arrangement

$$\text{Arr \#1:} \quad \star\ \star\ \star\ x_3^{(2)}\ \star\ \star\ \star\ \star\ \star\ x_2^{(2)}\ x_4^{(2)}\ \star\ \star\ \star\ \star\ x_1^{(2)}\ \star\ \star$$

and that we have another arrangement

$$\text{Arr \#2:} \quad \star\ \star\ \star\ x_1^{(2)}\ \star\ \star\ \star\ \star\ \star\ x_4^{(2)}\ x_2^{(2)}\ \star\ \star\ \star\ \star\ x_3^{(2)}\ \star\ \star$$

where all items that do not belong to group 2 (the ones marked "$\star$") occupy the same column in both arrangements. To put it differently, we obtained Arr #2 from Arr #1 by permuting the items in group 2 and leaving all other items in place.

In total there are $n_2! = 4! = 24$ such permutations. Let us consider one of them as special. For example, this one,

$$\text{Arr \#5:} \quad \star\ \star\ \star\ x_1^{(2)}\ \star\ \star\ \star\ \star\ \star\ x_2^{(2)}\ x_3^{(2)}\ \star\ \star\ \star\ \star\ x_4^{(2)}\ \star\ \star$$

where the group 2 items are arranged, left to right, in increasing order of their subscripts.

We go through all $n!$ permutations and discard all those where the group 2 items are ordered differently from $x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, x_4^{(2)}$.

$$\text{Then only} \quad \frac{n!}{n_2!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in group 2.

We repeat the above procedure to those survivors, but for group 1. We discard all those where the group 1 items are not ordered $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}$.

$$\text{Then only} \quad \frac{n!}{n_2!\, n_1!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in groups 1 and 2.

We keep going with the remaining groups.

$$\text{Then only} \quad \frac{n!}{n_1!\, n_2! \cdots n_r!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in all r groups.

It follows that there are $n!\,/(n_1!\, n_2! \cdots n_r!)$ different arrangements if we cannot distinguish the items belonging to the same group. ■

Example 7.7. How many distinct permutations are there of the word SHANANANANA

Answer: We designate Groups 1–4 according to the letters S, H, A, N.

Then $n_1 = n_2 = 1, n_3 = 5, n_4 = 4$. Further, $n = 1 + 1 + 5 + 4 = 11$. Thus, there are

$$\frac{11!}{5! \cdot 4! \cdot 1! \cdot 1!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{4 \cdot 3 \cdot 2} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{3\,(4 \cdot 2)} = 11 \cdot 10 \cdot 9 \cdot 7 \cdot 2 = 13,860$$

distinguishable arrangements of the word SHANANANANA. □

Definition 7.3 (Multinomial coefficients)**.**

The numbers

$$\binom{n}{n_1\, n_2 \cdots n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}\,. \tag{7.3}$$

that appear in Theorem 7.4 are called **multinomial coefficients.** If $r = 2$, then there is some integer $0 \le k \le n$ such that $n_1 = k$ and $n_2 = n - k$. We write

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \qquad \text{for} \qquad \binom{n}{k, n-k} \tag{7.4}$$

and speak of **binomial coefficients**. Convention: We define $\binom{n}{k} := 0$ for $k > n$. □

Example 7.8. For the variables A, H, N, S, what is the coefficient of $(A + H + N + S)^{11}$ for

$$5 \text{ factors } A, \quad 1 \text{ factor } H, \quad 4 \text{ factors } N, \quad 1 \text{ factor } S?$$

Answer: For $n_A = 5, n_H = 1, n_N = 4, n_S = 1$ the coefficient is

$$\binom{11}{5, 1, 4, 1} = \frac{11!}{5! \cdot 1! \cdot 4! \cdot 1!} = 13,860\,. \tag{7.5}$$

There is a connection to Example 7.7 on p.173. One of the $13,860$ products obtained by multiplying the factors listed in (7.5) is $S \cdot H \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A$.

- It has the following in common with the other $13,859$ products: They all consist of 5 symbols A, 1 symbol H, 4 symbols N, 1 symbol S.
- The other $13,859$ products differ from $S \cdot H \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A$ as follows: At least one of the 11 positions contains a different symbol

Thus, if we identify $S \cdot H \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A \cdot N \cdot A$ with the word "SHANANANANA", we found out that there are exactly $\binom{11}{5,1,4,1} = 13,860$ different words that can be formed from the letters found in "SHANANANANA". That is the same result as that in Example 7.7! □

The next theorem explains the appropriateness of the previous definition.

Theorem 7.5.

Let $r, n \in \mathbb{N}$ such $r \le n$ and $x_1, x_2, \ldots x_r \in \mathbb{R}$. Then

$$\big(x_1 + x_2 + \cdots + x_r\big)^n = \sum_{\substack{n_1,\ldots,n_r \ge 0 \\ n_1+\cdots+n_r = n}} \binom{n}{n_1, n_2, \ldots n_r} x_1^{n_1}\, x_2^{n_2} \cdots x_r^{n_r}\,. \tag{7.6}$$

*In particular, if $r = 2$, we obtain the **binomial theorem**:*

$$\big(x_1 + x_2\big)^n = \sum_{j=0}^{n} \binom{n}{j} x_1^j\, x_2^{n-j}\,.$$

PROOF:

First, we show that the case $n = 2$ follows from 7.6.

Since $n_1, n_2 \geq 0$ and $n_1 + n_2 = n \Rightarrow 0 \leq n_1 \leq n$ and $n_2 = n - n_1$, writing j for n_1 yields the binomial theorem formula.

To prove the first formula, We start by "multiplying out" the product

$$\big(x_1 + x_2 + \cdots + x_r\big)^n \;=\; \big(x_1 + x_2 + \cdots + x_r\big)\big(x_1 + x_2 + \cdots + x_r\big)\cdots\big(x_1 + x_2 + \cdots + x_r\big)$$

and obtain in the resulting expansion terms of the form

$$a_1 \cdot a_2 \cdots a_n \quad \text{such that each factor } a_j \text{ is either } x_1 \text{ or } x_2 \text{ ... or } x_r.$$

In the following we consider the sizes $n_1, n_2, \ldots, n_r$ as fixed

Note that it is not possible to obtain two selections

$$\vec{a} = (a_1, a_2, \ldots, a_n) \quad \text{and} \quad \vec{b} = (b_1, b_2, \ldots, b_n) \quad \text{such that} \quad a_j = b_j \;\text{ for all } j.$$

The reason: We multiply out the n factors $(x_1 + \cdots x_r)$ in such a way that for no two of the resulting products we picked the same variable x_i in each one of those n factors $(x_1 + \cdots x_r)$

But then the following is true if we consider such a selection as a word $a_1 a_2 \ldots a_n$ where each lettter is one of x_1 or x_2 ... or x_r. Any two of those words are distinguishable even though some or all of the letters x_i can occur multiple times.

> For example, if $n = 7, n_1 = 2, n_2 = 3, n_3 = 2$ and we write X for x_1, Y for x_2, Z for x_3, we have this situation.
>
> The word $YXZZYYX$ is formed only once. But of course, we obtain other words with the same sizes n_j, e.g. the rearrangement $ZYXZYXY$ which is distinguishable from the first word.

Thus, in the general case, there are as many terms in the expansion of $\big(x_1 + x_2 + \cdots + x_r\big)^n$ containing each symbol x_j exactly n_j times as there are distinguishable "words" that contain each x_j exactly n_j times. According to Theorem 7.4, there are

$$\binom{n}{n_1, n_2, \ldots n_r} \;=\; \frac{n!}{n_1!\, n_2! \cdots n_r!}\,.$$

such terms. Since this is the number of times the product $x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$ occurs in the expansion of $\big(x_1 + x_2 + \cdots + x_r\big)^n$, it follows that

$$\big(x_1 + x_2 + \cdots + x_r\big)^n \;=\; \sum_{\substack{n_1,\ldots,n_r \geq 0 \\ n_1 + \cdots + n_r = n}} \binom{n}{n_1, n_2, \ldots n_r}\, x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}\,. \;\blacksquare$$

Theorem 7.6.

Given are n distinct items and r distinct bins of fixed sizes $n_1, n_2, \ldots, n_r$ such that $n_1 + \cdots + n_r = n$. Then the number of distinguishable placements of the n items into those r bins, when disregarding the order in which the items were placed into any one of those bins, is

$$\binom{n}{n_1, n_2, \ldots n_r} \;=\; \frac{n!}{n_1!\, n_2! \cdots n_r!}\,.$$

The proof is given after the following example which will help clarify how to interpret Theorem 7.6.

Example 7.9. Given are a list of $n = 7$ items and $r = 3$ bins as follows.

- The 7 items are a, b, c, d, e, f, g.
- Bin 1 has size 2, bin 2 has size 3, bin 3 has size 2 (thus $n = 2 + 3 + 3 = 7$).
- Arr #1: bin 1 has b, c, bin 2 has e, a, g, bin 3 has f, d
- Arr #2: bin 1 has c, b, bin 2 has a, g, e, bin 3 has d, f
- Arr #3: bin 1 has b, d, bin 2 has a, g, e, bin 3 has c, f
- Then Arr #1 and Arr #2 are considered the same since each bin contains the same items. Only their order is different.
- On the other hand, both Arr #1 and Arr #2 both are considered different from Arr #3 since, e.g., bin 1 contains item d for #3, but bin 1 does not contain item d for the other two arrangements. □

PROOF of Theorem 7.6:

The proof is very similar to that of Theorem 7.4, so we keep the discussion brief.

- For each one of the $n!$ permutations of all n items, there are $n_1! - 1$ others which possess the same n_1 elements in bin 1, only differently ordered, but have exactly the same item at each other of the remaining $n - n_1$ spots. Removing those duplicates leaves us with $n!/n_1!$ arrangements.
- Of those $n!/n_1!$ arrangements, there are $n_2! - 1$ others which possess the same n_2 elements in bin 2, only differently ordered, but have exactly the same item at each other of the remaining $n - n_1 - n_2$ spots. Removing those duplicates leaves us with $n!/(n_1!n_2!)$ arrangements.
- -
- Having removed the duplicates from bins 1 through $k-1$, we are left with $\frac{n!}{n_1!\cdots n_{k-1}}$ arrangements. For each one of those there are $n_k! - 1$ others which possess the same n_k elements in bin k, only differently ordered. Removing those duplicates leaves us with $\frac{n!}{n_1!\cdots n_k}$ arrangements.
- For any two surviving arrangements the following is true: There is at least one bin that does not contain the same elements (possibly rearranged) for both those arrangements.

to make all n items artificially distinguishable. We have learned that there are $n!$ permutations.

When we only keep the superscripts that indicate the group but we remove the subscripts, since in truth items belonging the same group cannot be distinguished, there will be a lot less arrangements that are distinct.

To fix the ideas, assume that group 2 has 4 members and we have an arrangement

$$\text{Arr \#1:} \quad \star\ \star\ \star\ x_3^{(2)}\ \star\ \star\ \star\ \star\ \star\ x_2^{(2)}\ x_4^{(2)}\ \star\ \star\ \star\ \star\ x_1^{(2)}\ \star\ \star$$

and that we have another arrangement

$$\text{Arr \#2:} \quad \star\ \star\ \star\ x_1^{(2)}\ \star\ \star\ \star\ \star\ \star\ x_4^{(2)}\ x_2^{(2)}\ \star\ \star\ \star\ \star\ x_3^{(2)}\ \star\ \star$$

where all items that do not belong to group 2 (the ones marked "$\star$") occupy the same column in both arrangements. To put it differently, we obtained Arr #2 from Arr #1 by permuting the items in group 2 and leaving all other items in place.

In total there are $n_2! = 4! = 24$ such permutations. Let us consider one of them as special. For example, this one,

$$\text{Arr \#5:} \quad \star\,\star\,\star\; x_1^{(2)}\; \star\,\star\,\star\,\star\,\star\; x_2^{(2)}\; x_3^{(2)}\; \star\,\star\,\star\,\star\; x_4^{(2)}\; \star\,\star$$

where the group 2 items are arranged, left to right, in increasing order of their subscripts.
We go through all $n!$ permutations and discard all those where the group 2 items are ordered differently from $x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, x_4^{(2)}$.

$$\text{Then only} \quad \frac{n!}{n_2!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in group 2.
We repeat the above procedure to those survivors, but for group 1. We discard all those where the group 1 items are not ordered $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}$.

$$\text{Then only} \quad \frac{n!}{n_2!\, n_1!} \quad \text{arrangements remain,}$$

but for those the artificial distinction coming from the subscipts is gone in groups 1 and 2.
We keep going with the remaining groups....

$$\text{In the end only} \quad \frac{n!}{n_1!\, n_2! \cdots n_r!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in all r groups.
It follows that there are $n!\,/(n_1!\, n_2! \cdots n_r!)$ different arrangements if we cannot distinguish the items belonging to the same group. ■

Proposition 7.1.

(A) *There are* $\binom{n-1}{r-1}$ *distinct integer–valued vectors* $\vec{x} = (x_1, x_2, \ldots, x_r)$ *such that*

$$x_1 + x_2 + \cdots + x_r = n \qquad \textit{and} \quad x_i > 0,\ i = 1, \ldots, r\,.$$

(B) *There are* $\binom{n+r-1}{r-1}$ *distinct integer–valued vectors* $\vec{y} = (y_1, y_2, \ldots, y_r)$ *such that*

$$y_1 + y_2 + \cdots + y_r = n \qquad \textit{and} \quad y_i \geq 0,\ i = 1, \ldots, r\,.$$

PROOF of **(A):**
Each such equation corresponds to an arrangement of n symbols $\otimes$ which denote the numbers $1, 2, \ldots, n$ in sequence, and $r-1$ bars | which are places in-between those symbols, in such a way, that no two bars are adjacent. For example, the arrangement

$$\bullet\,\bullet \;|\; \bullet\,\bullet\,\bullet\,\bullet \;|\; \bullet\,\bullet\,\bullet$$

expresses the equation $2+4+3=9$. In the general case, one or zero bars can be placed in the $n-1$ gaps between the n bullets:

(A) $\bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \cdots \otimes \bullet \otimes \bullet$

Thus, there are as many different integer equations as there are ways to select $r-1$ of those $n-1$ gaps for the $r-1$ bars. This number is $\binom{n-1}{r-1}$.

FIRST PROOF of **(B):**

An equation $\sum_{j=1}^{r} y_j = n; y_j \geq 0$ of part **(B)** becomes an equation $\sum_{j=1}^{r} x_j = n+r; x_j > 0$ of part **(A)**, by setting $x_j := y_j + 1$.

In reverse, equation $\sum_{j=1}^{r} x_j = n+r; x_j > 0$ of part **(A)** becomes an equation $\sum_{j=1}^{r} y_j = n; y_j \geq 0$ of part **(B)**, by setting $y_j := x_j - 1$.

We have shown in **(A)** that there are $\binom{n+r-1}{r-1}$ different equations of the form $\sum_{j=1}^{r} x_j = n+r; x_j > 0$.

Thus, there also that many of the form $\sum_{j=1}^{r} y_j = n; y_j \geq 0$. This proves **(B)**.

ALTERNATE PROOF of **(B):** We add two more placeholders $\otimes$ for the separating bars. One to the left of the leftmost bullet and another to the right of the rightmost bullet. The condition $y_j \geq 0$ instead of $x_j > 0$ implies that each one of those placeholders can be occupied by as few as zero bars and as many as all $r-1$ bars. To put it differently, any combination of bullets and bars is admissible.

We create a tagged list of $n+r-1$ distinct placeholders for both bullets and bars and select $r-1$ of them for the bars. Obviously, the order of the bars does not matter. Thus there are $\binom{n+r-1}{r-1}$ such selections. ■

Consider the issue of distributing n indistinguishable items into r distinct bins where bin_j contains $0 \leq n_j \leq n$ items and the n_j are allowed to vary for different selections. (Of course, $n_1+\cdots+n_r = n$.) Then each such selection corresponds to an integer vector $\vec{n} = (n_1, \ldots, n_r)$ which is a solution of the equation $\sum_{j=1}^{r} n_j = n; n_j \geq 0$.

If we demand in addition that each bin contains at least one item, then each such selection corresponds to an integer vector $\vec{n} = (n_1, \ldots, n_r)$ which is a solution of the equation $\sum_{j=1}^{r} n_j = n; n_j > 0$.

We obtain from Proposition 7.1 the following.

Proposition 7.2.

(A) *There are* $\binom{n-1}{r-1}$ *ways to select* n *indistinguishable items into* r *distinct bins such that each bin contains at least one item.*

(B) *There are* $\binom{n+r-1}{r-1}$ *ways to select* n *indistinguishable items into* r *distinct bins.*

PROOF: This follows from from Proposition 7.1. ■

Example 7.10. Mother Jones' cookies and the stars & bars examples:

- How many ways are there to give 10 cookies to 4 kids if each one gets at least one cookie? **A:** There are $\binom{10-1}{4-1} = (9 \cdot 8 \cdot 7)/(3 \cdot 2 \cdot 1) = 84$ ways.
- How many ways are there to separate 6 stars by two bars into three parts, if one or more of those parts may contain zero stars? **A:** There are $\binom{6+3-1}{3-1} = (8 \cdot 7)/(2 \cdot 1) = 28$ ways. □

Here is another example that employs binomial coefficients.

Example 7.11 (Ross-prob-thy-3ed Example 4c)**.** Given are n antennas of which d are defective. They will be arranged in a linear order and will relay signals. This chain will not function if two or more defective items are placed next to each other.

How many ways are there to arrange the antennas so that we obtain a functioning arrangement?

Answer: We denote the $n - d$ working antennas by the $\otimes$ symbol, separate them by bullets $\bullet$ and add one $\bullet$ each to the left of the leftmost and to the right of the rightmost.

$$\bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \cdots \otimes \bullet \otimes \bullet$$

Then the functioning relays are precisely those where one or zero defective antennas are placed at each one of those $\bullet$ spots. Each such placement corresponds to a selection of size d of those $n-d+1$ bullets: The selected spots will get a defective antenna and nothing will happen to the others.

$$\text{Thus, there are } \binom{n-d+1}{d} \text{ functioning arrangements. } \square$$

Problem 7.2. A lottery is held among N participants. There are K drawings in which a prize is given away. $(K < N)$. In each drawing, each participant has an equal chance of obtaining the prize. (Thus, it is possible, though unlikely, that one single person walks away with all K prizes.) Amanda is one of the participants. What is the probability that she will walk away with exactly k prizes? Of course, $(k \le K)$.

Solution:

(a) There are N different selections for drawing #1.

(b) Each one of those has N selections for drawing #2. Thus, there are N^2 different ways to distribute the first two prizes

(c) Each one of those N^2 has N selections for drawing #3. Thus, there are N^3 different ways to distribute the first 3 prizes .....

(d) ..... Thus, there are N^K different ways to distribute all K prizes .....

It follows that the sample space Ω has size N^K. Since all drawings are done at random, all outcomes $\omega \in \Omega$ are equally likely. Thus, $P\{\omega\} = 1/(N^K)$ for all ω. Note that an outcome $\omega \in \Omega$ is of the form

$$(\star) \qquad \omega = (i_1, i_2, \ldots, i_K) : \quad \text{prize } 1 \text{ goes to person } i_1, \ \ldots \ \text{prize } K \text{ goes to person } i_K$$

- Let $A := \{$ Jane gets exactly k prizes $\}$.

Assume that the outcomes ω and ω' are as follows:

- ω: participant i_1 gets prize j_1 and i_2 gets prize j_2
- ω': participant i_1 gets prize j_2 and i_2 gets prize j_1
- There is no difference how other $K - 2$ prizes were awarded.

Even though order matters, we only are able to distinguish the outcomes ω and ω' if j_1 and j_2 are given to different persons. Otherwise all K slots of both ω and ω' are identical, i.e., $\omega = \omega'$.

Thus, there are (only) as many different ways to give k of the K prizes to Jane as there are ways to select k of K items DISREGARDING ORDER. That number is $\binom{K}{k}$.

Next, consider that each one of those $\binom{K}{k}$ ways of designing k of the K slots of an outcome ω to Jane must be complemented by filling each one of the remaining $K-k$ slots with one of the other $N-1$ participants. This time we CANNOT DISREGARD ORDER. See the discussion above concerning the outcomes ω and ω'.

- We repeat the reasoning of **(a) – (d)** to $N-1$ instead of N choices for those $K-k$ instead of k drawings and see that there are $(N-1)^{K-k}$ possible selections.
- The event A consists all outcomes obtained by matching any one of those $(N-1)^{K-k}$ selections wih any one of the $\binom{K}{k}$ ways of allocating k prizes to Jane.
- By the multiplication rule, $|A| = \binom{K}{k} (N-1)^{K-k}$.
- Since all outcomes are equally likely, $P(A) = \dfrac{|A|}{|\Omega|} = \dfrac{\binom{K}{k}(N-1)^{K-k}}{N^K}$. □

We summarize the results of Theorem 7.4, Theorem 7.6, Proposition 7.1, and Proposition 7.2.

Remark 7.2. The multinomial coefficients

$$\binom{n}{n_1\, n_2 \cdots n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

of Definition 7.3 appear in the following settings:

- Distinct selections of n items of which n_1 are alike, n_2 are alike, ..., n_k are alike. Example: different rearrangements of the word "BANANA".
- They are coefficients in the expansion of $\left(x_1 + x_2 + \cdots x_k\right)^n$.
- Selections of n distinct items into k distinct bins of fixed sizes $n_1, \ldots, n_k$, disregarding order within each bin. That is the WMS definition in their Theorem 2.3 of Ch.02.6. □

8 More on Probability

This chapter corresponds to material found in WMS ch.2

8.1 Total Probability and Bayes Formula

Theorem 8.1 (Total Probability and Bayes Formula [86]).

Assume that $\{B_1, B_2, \dots\}$ is a partition of Ω and that $A \subseteq \Omega$. such that $P(B_j) > 0$ for all j. Then

$$P(A) = \sum_{j=1}^{\infty} P(A \mid B_j)\, P(B_j)\,. \tag{8.1}$$

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{\infty} P(A \mid B_i)\, P(B_i)}\,. \qquad \textbf{(Bayes formula)} \tag{8.2}$$

Note that the above also covers finite partitions $\{B_1, B_2, \dots, B_k\}$ of Ω: apply the formulas with

$$B_{k+1} := B_{k+2} := \cdots := 0\,.$$

PROOF: Since $(B_j)_j$ partitions Ω $(A \cap B_j)_j$ partitions A. Thus, $A = \biguplus_j (A \cap B_j)$. Thus,

$$P(A) = \sum_{j=1}^{\infty} P(A \cap B_j)\,. = \sum_{j=1}^{\infty} P(A \mid B_j)\, P(B_j)\,.$$

This proves (8.1). To prove (8.2), we apply to its right–hand side the already proven (8.1). We obtain

$$\frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{\infty} P(A \mid B_i)\, P(B_i)} = \frac{P(A \mid B_j)P(B_j)}{P(A)} = \frac{P(A \cap B_j)}{P(A)} = P(B_j \mid A)\,. \quad \blacksquare$$

When working with conditional probabilities, in particular when one wants to apply the Bayes formula, it often is convenient to work with tree diagrams. This is demonstrated in the next example.

Problem 8.1. It has been established that 40% of all jobs for college graduates are in the technology sector. Of those college graduates who work in technology, one quarter enjoys listening to classical music. Of those college graduates who hold other kinds of jobs, one out of three enjoys listening to classical music.

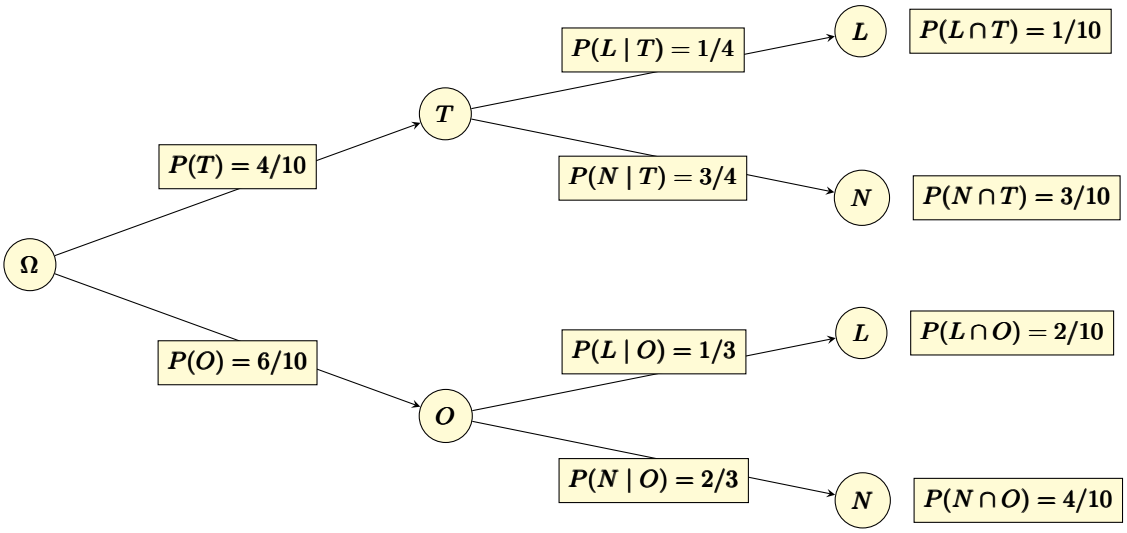
(a) What is the probability that Pedro neither works in technology, nor listens to classical music?
(b) Harry works in technology. How likely is it that he does not listen to classical music?
(c) Jane says that she likes classical music. What is the probability that she works in technology?

Solution: We use the following abbreviations:

T:	Works in technology	O:	"Other": does not work in technology
L:	Listens to classical music	N:	Does not listen to classical music

[86] Thomas Bayes (1702 - 1761) was an English clergyman and mathematician.

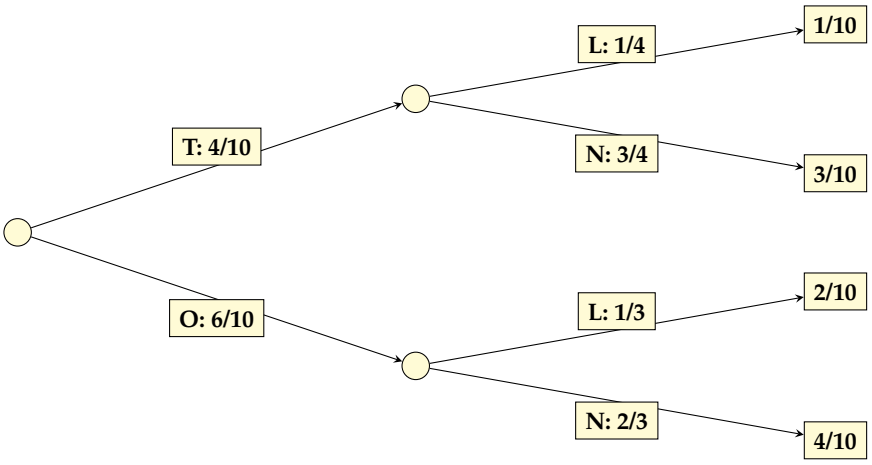
The information available to us is sufficient to draw the following tree diagram:



A line segment that connects two nodes indicates conditioning of the right side on the left side. For example, the node that connects T and N signifies that the event N is conditioned on the event T. $P(L \mid T)$, the corresponding conditional probability, is attached to the line segment. Note that this is also true for the two line segments that emanate from Ω, since $P(T) = P(T \mid \Omega)$ and $P(O) = P(O \mid \Omega)$ Note that T and O partition Ω and the same is true for L and N.

Tree diagrams can be very convenient because the probability of an intersection is obtained by multiplying the two probabilities to the left. For example, $P(T \cap N) = (4/10)(3/4) = 3/10$.

Not all the notation is necessary to work with such a diagram. Here is a pared down version:



Let us now discuss the answers to the three problems posed above

(a) What is the probability that Pedro neither works in technology, nor listens to classical music?
• This is the ordinary (no conditioning) probability $P(O \cap N) = 4/10$.

(b) Harry works in technology. How likely is it that he does not listen to classical music?
• We are conditioning on the event T and want to compute $P(N \mid T$. The diagram shows that $P(N \mid T) = 3/4$.

(c) Jane says that she likes classical music. What is the probability that she works in technology?
• We are asking for the conditional probability $P(T \mid L)$.

This is a reverse conditioning (Bayes formula problem. The tree diagram makes it easy to find all the probabilities involved:

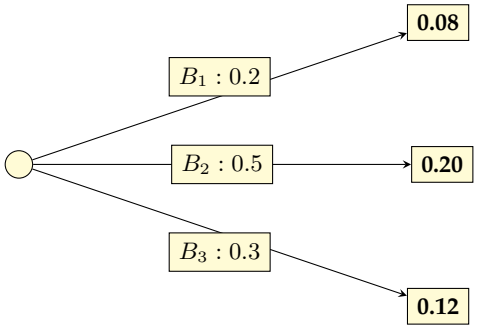
- $P(T \mid L) = P(T \cap L)/P(L)$.
- $P(T \cap L) = 1/10$ and $P(L) = P(O \cap L) + P(T \cap L) = (2+1)/10 = 3/10$.
- Thus, $P(T \mid L) = (1/10)/(3/10) = 1/3$.

We continue with some general remarks concerning tree diagrams.

It should be clear how to generalize such diagrams. One can condition at each stage on more than just two events. For example, Let us assume the following.
In stage 1, we "condition" Ω on $\Omega = A_1 \uplus A_2 \uplus A_3$,
In stage 2, we condition A_2 on $\Omega = B_1 \uplus B_2 \uplus B_3 \uplus B_4$.
If

$$P(A_2) \;=\; 0.4\,,$$

then the resulting tree fragment is to the right.



Because $\Omega = \biguplus_j B_j$, it is always true that

$$\sum_j P(B_j \mid A_k) \;=\; \frac{\sum_j P(B_j \cap A_k)}{P(A_k)} \;=\; \frac{P(A_k)}{P(A_k)} \;=\; 1$$

Thus, the sum of the conditional probabilities over all line segment that emanate from a given node is 1. In the tree excerpt above: that node is $A_k = A_2$ and the sum of the conditional probabilities is

$$P(B_1 \mid A_2) \;+\; P(B_2 \mid A_2) \;+\; P(B_3 \mid A_2) = 0.2 + 0.5 + 0.3 \;=\; 1\,. \;\square$$

8.2 Sampling and Urn Models With and Without Replacement

The following definition is **PRELIMINARY** and will be amended in Definition 8.2 (Sampling as a Random element) below (see p.185).

Definition 8.1.

(a) We call the action of picking n items $x_1, x_2, \ldots, x_n$ from a collection of N items a **sampling action of size** n. Aternatively, we also use the phrases **sampling process** and **sampling procedure**. Here, $n \in \mathbb{N}$ and $N \in \mathbb{N}$ or $N = \infty$.

(b) We call the specific outcome of such a sampling action (the list $x_1, x_2, \ldots, x_n$) a **realization** of that sampling action. $\square$

(c) In yet another instance of notational abuse, both the sampling action and an outcome of this action (a realization) will be referred to as a **sample** of size n if this does not lead to any confusion. Note that we had mentioned this previously in Example 1.5 on p.12. $\square$

Example 8.1. Each of the following can be considered samples.

- **(a)** Drawing blindfolded a ball from an urn that contains N balls $n = 5$ times in a row recording each time the outcome and then replacing the ball (putting it back).
- **(b)** Drawing blindfolded $n = 5$ balls from an urn that contains N balls in one fell swoop, i.e., not replacing any of the balls
- **(c)** Rolling a die twice in a row and recording the outcome.
- **(d)** Selecting in a random fashion $n = 2,000$ persons from all persons eligible to vote without replacement, i.e., we want a sample of n distinct voters. Note that N is huge when compared to n.
- **(e)** Same as **(d)**, but we only record their voting preference, their annual income and their age and discard all other data.
- **(f)** Same as **(e)**, but we only record their annual income.
- **(g)** The random numbers generator of a computer creates a sample of n numbers such that they are uniformly distributed on the interval $[0, 1]$. [87] (Computers can do that!)
  Since there are infinitely many such numbers and the computer can generate any one of them, [88] $N = \infty$.
- **(h)** A factory mass–produces an item, e.g., screws, at a huge rate per hour. Quality control randomly picks $n = 50$ every hour and checks for defective items. Since the number N of screws from which the sample is obtained is so huge, one can, for all practical purposes, act as if $N = \infty$. (This will considerably simplify the mathematics involved in computing, e.g., the probability that such a sample contains 5 or more defective items) if the rate of defectives is supposed to be 3.5%.
- **(i)** We write down the numbers $1, 2, \ldots, 10$. Such **deterministic sampling** is very boring for a course called "Probability Theory", because no randomness is involved. Nevertheless, Definition 8.1 encompasses determininistic sampling. □

Remark 8.1.

- **(a)** We only are interested in samples that involve randomness. In other words, if there is a set U such that $x_j \in U$ for all j, our sample can be modeled, for fixed n, as a random element $\vec{X} : (\Omega, P) \to U^n$. Since deterministic samples can be interpreted as functions of ω which do not vary with ω, i.e., as constant random elements, they too are covered by Definition 8.1.
- **(b)** Since the "population" from which each item $x_j = X_j(\omega)$ is sampled is the set U from **(a)**, it is possible to implement $\Omega := U^N$ as the carrier set of the probability space (Ω, P). In other words, we could narrow things down to $\vec{X} : (U^N, P) \to U^n$. Matter of fact, you will be as specific as you can when trying to find the formula or just the particular number that solves a given problem.
- **(c)** But there are advantages to refer to an unspecified probability space (Ω, P) when dealing with the general theory. A good example are the theorems and definitions about expectation and variance in MF Chapter 9 (Discrete Random Variables and Random Elements) where going into specific settings would hinder rather than help the understanding. □

Here is the promised amended version of Definition 8.1.

[87] "uniformly distributed" means that the proportion of numbers x_j that fall within the interval $0 \le a < b \le 1$ is (approximately) $b - a$.

[88] in theory, since there is no such thing as "infinitely many") in our physical reality

Definition 8.2 (Sampling as a Random element)**.** Let (Ω, P) be a probability space. Let $U \neq \emptyset$ be a collection of N items ($N \in \mathbb{N}$ or $N = \infty$), which we can think of as the "population of interest". Let $n \in \mathbb{N}$ (so $n < \infty$), such that $n \leq N$.

(a) Let $\vec{X} : (\Omega, P) \longrightarrow U^n$ be a random element with codomain U^n. If we interpret $\vec{X}$ as the action of picking n items

$$\vec{x} = x_1, x_2, \ldots, x_n = \vec{X(\omega)} = X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$$

from U, then we call $\vec{X}$ a **sampling action of size** n. Aternatively, we also use the phrases **sampling process** and **sampling procedure**.

(b) We call a specific outcome (the list $\vec{x} = (x_1, x_2, \ldots, x_n)$) a **realization** of that sampling action. See Example 1.5 on p.12.

(c) Both the sampling action and an outcome of this action (a realization) are called a **sample** of size n if the context makes it clear what is being discussed.

(d) If there is a specific $\vec{x}^* \in U^n$ such that $P\{\vec{X} = \vec{x}^*\} = 1$, (this certainly is the case if $\vec{X}(\omega) = \vec{x}^*$ for all $\omega \in \Omega$), then we call both the sampling action $\vec{X}$ and the realization $\vec{x}^*$ a **deterministic sample**. □

Remark 8.2.

(a) You may wonder about the difference between a U^n–valued random element and a sample of n items which are picked from a population U. The answer: Mathematically speaking, there is no difference whatsoever. It is the interpretation that matters!

(b) Going back to using the terms probability space and sample space interchangeably, this author likes to think not of (Ω, P), but only of $(U^n, P_{\vec{X}})$ as a sample space. The reason is that the latter hosts the potential outcomes of the sampling action $\vec{X}$. (And yes, the probability measure $P_{\vec{X}}$ on that sample space is the distribution of $\vec{X}$).

(c) Do those individual sample picks X_j happen with or without replacement? In other words, can the same $x \in U$ be picked more than once or are for a fixed ω all outcomes distinct? The answer: The definition does not say. This must always be explicitly stated or known from the context.

(d) Consider items **(d)** and **(h)** of Example 8.1. If $N \gg n$, then the computational differences between selecting the sample with or without replacement are so small that we can assume sampling with replacement even if the sampled items are not returned to the population after each pick. This often simplifies the computational effort involved. □

Remark 8.3. We switch focus to the role of proper randomization when picking a sample.

(a) Picking a small size sample that allows us to make inferences to the population from which it was drawn can require a lot of thought. The budget available for collecting that sample is often limited and will limit the methods available. Of course, a smaller sample will cost less than a bigger one if the procedure to collect the data is the same in both cases.

We fix $n \in \mathbb{N}$. What will make the sample representative of the population, i.e.

• what guarantees that the composition of the sample mirrors that of the population?

It certainly will not help if the sample has, e.g., 90% students, whereas the population of interest only has 20%. So we fix that by establishing quota and requiring the proportion of students to be 20%. Of course, there is also the ethnic composition of the population that we want mirrored in the sample. And there is income distribution, gender and $5,000$ or more attributes for which we want to maintain close to identical proportions in the sample.

(b) Clearly, a practical limit to the number of ways a (hopefully small) sample can be partitioned into "strata" is reached quickly. So we must look for an alternative way to obtain a sample that is not biased in favor of value $\boldsymbol{v}$, say "is male" of attribute $\boldsymbol{A}$ (here: gender), when compared to the proportion in the population. And we need this for all important $\boldsymbol{v}$ and $\boldsymbol{A}$.

(c) The solution is to make the sample selection as random as possible:

• We pick the first item at random, i.e., with the same chance $\frac{1}{N}$,
• Then we pick $\#2$ at random from the remaining $N-1$,
• Then we pick $\#3$ at random from the remaining $N-2$,
• .... Finally, we pick $\#n$ at random from the remaining $N-n+1$ items.

Doing so ensures that any collection $\vec{x} = (x_1, \ldots, x_n)$ has the same chance of being selected as any other collection $\vec{x}' = (x'_1, \ldots, x'_n)$. By the way, we know that probability:

• If we do not worry about the order in which the n distinct items were selected, then there are $\binom{N}{n}$ different selections and that probability is $1/\binom{N}{n}$.
• If order does matter and we deal with permutations, then the answer is $1/P_n^N$.

The degree of randomness obtained by following this procedure prevents any kind of gross distortion (bias) in the sample.

(d) Would the requirement of **(c)** that each collection of n items have the same chance to be drawn as any other such collection be the same as simply asking that each item in the population have the same probability, $1/N$, of being selected? **The answer is NO** as Example 8.2 below will show. □

Example 8.2. We have a population of $N = 600$ students. 100 of them are freshmen, 100 of them are sophomores, 100 of them are juniors, 100 of them are seniors, 100 of them are first year graduate students, the others are second year graduate students.

A sample of $n = 100$ will be selected as follows. A fair die is rolled. If the outcome is 1, all freshmen will be selected, On a 2, all sophomores will be selected, ... On a 6, all second year graduate students will be selected.

- In the resulting sample each student has the same probability $1/6$ of being selected.
- But only 6 of the possible $\binom{600}{100}$ possible outcomes have a non–zero chance (of $1/6$ each) of being selected: Those where each student belongs to the same group as all the others! □

There is a special name for samples which are collected as outlined in Remark 8.3**(c)**.

Definition 8.3 (Simple Random Sample)**.**

(a) We call a sampling action of size n ($n \in \mathbb{N}$) from a population of size $N < \infty$ a **simple random sampling action**, in brief, an **SRS action**, if there are no duplicates allowed (i.e., we sample without replacement) and each of the potential outcomes has equal chance of being selected.

(b) As in Definition 8.2 (Sampling as a Random element), we call both an SRS action and a realization of this action a **simple random sample of size** n. (Briefly, an **SRS**.) □

The generic sounding term "random sample" has a very specific meaning in statistics.

Definition 8.4 (Random Sample)**.**

(a) We call a sampling action of size n ($n \in \mathbb{N}$) from a population of size $N < \infty$ a **random sampling action**, if the picks are independent of each other. See Chapter 5.4 (Independence of Random Elements) for the definition of independent random elements.

(b) As in Definition 8.2 (Sampling as a Random element), we call both a random sampling action and a realization of this action a **random sample of size** n. □

SRS amounts to sampling according to Remark 8.3**(c)**. When abstracting from the specifics, this boils down to being blindfolded and selecting, **WITHOUT REPLACEMENT**, n well shuffled balls from an urn containing N numbered balls.

On the other hand, random samples are obtained when those balls are drawn from the urn **WITH REPLACEMENT**.

Some authors use the scenario of tickets in a box rather than balls in an urn.

Definition 8.5 (Urn models)**.**

(a) An **urn model without replacement** describes a mechanism by which a blindfolded person selects a fixed number of balls from an urn in which the balls have been well mixed. Note that the resulting realizations will contain no duplicates.

(b) An **urn model with replacement** describes a mechanism by which a blindfolded person selects a fixed number of balls from an urn as follows.
(1) The balls are well mixed.
(2) A ball is picked and the outcome is recorded.
(3) The ball is put back into the urn.
(4) Steps **(1)** through **(3)** are repeated until all n balls have been selected. □

More material may be added to this section at a later time.

9 Discrete Random Variables and Random Elements

This chapter corresponds to material found in WMS ch.3

Remark 9.1. There was no need for a specific arrangement $x_1, x_2, \ldots$ of the countably many x that satisfy $p_X(x) > 0$ in the series $\sum_{x \text{ s.t. } p_X(x)>0} p_X(x)$:

- $p_X(x_j) \geq 0$ for all j ensures that the value of the series $\sum p_X(x_j)$ does not depend on the particular order in which the numbers $p_X(x_j)$ are added. See Theorem 3.2 on p.58. See also Remark 5.3 on p.107, in which this issue was addressed.

However, going forward, there will be series $\sum a_j$ that do not necessarily satisfy $a_j \geq 0$ for all j. An important example for this will be the expected value, $E[Y] = \sum_{y:p_Y(y)>0} y \cdot P_Y(y)$, of a discrete random variable Y. See p.190 below, Definition 9.2 (Expected value of a discrete random variable). Accordingly, the blanket assumption that follows this remark will prove very convenient. □

Assumption 9.1 (All series are absolutely convergent)**.**

We assume the following for the entire remainder of these lecture notes.

- Unless explicitly stated otherwise, all sequences are either known to be absolutely convergent or assumed to be absolutely convergent.

In particular, if $p_X(x)$ is the probability mass function of a discrete random element X which takes values in a set Ω', $g : \Omega' \to \mathbb{R}$ is a real–valued function, and ω'_n is a sequence in Ω', then we assume that the series $\sum g(\omega'_n)p_X(\omega'_n)$ is absolutely convergent. □

9.1 Probability Mass Function and Expectation

We start with a trivial observation.

Proposition 9.1. *A real–valued function of a random element is a random variable.*

PROOF: Let $X : (\Omega, P) \to \Omega'$ be a random element on a probability space (Ω, P) and $g : \Omega' \to \mathbb{R}$ be a real–valued function. Then $\omega \mapsto g\big(X(\omega)\big)$ is a real–valued function of ω, hence it is a random variable. ■

Definition 9.1 (Probability mass function)**.**

For a discrete random element X on (Ω, P), define

$$p(x) := p_X(x) := P_X\{x\} = P\{X = x\}. \tag{9.1}$$

We call p_X the **probability mass function** (WMS: **probability function**) for X. We also write **PMF** for probability mass function. □

Theorem 9.1.

If p_X is the probability mass function of a discrete random element X, then

$$0 \le p_X(x) \le 1; \quad \text{for all } x \tag{9.2}$$

$$\sum_{x \text{ s.t. } p_X(x)>0} p_X(x) = 1 \tag{9.3}$$

Proof: See WMS ch.3. ■

Remark 9.2. Assume that $p_X(x)$ is the probability mass function of a discrete random element X with values in a set Ω'. Then there exists a countable set $\Omega^* \subseteq \Omega'$ such that $P_X(\Omega^*) = 1$. Thus, the probability mass function $p_X(\cdot)$ of X satisfies

$$p_X(x) = 0 \qquad \text{for all } x \in (\Omega^*)^{\complement}.$$

Let $g : \Omega' \to \mathbb{R}$ be a real–valued function. Clearly,

$$g(x) \cdot p_X(x) = 0 \qquad \text{for all } x \in (\Omega^*)^{\complement}.$$

Ω^* being countable means that $\Omega^* = \{x_1, x_2, \dots\}$ for some finite or infinite sequence x_j. The following is trivial in the finite case, so we confine ourselves to the infinite case, $\Omega^* = \{x_j : j \in \mathbb{N}\}$. For $j \in \mathbb{N}$, let $a_j := g(x_j)p_X(x_j)$. By Assumption 9.1 on p.188, the series $\sum a_j$ is absolutely convergent. Hence, its value does not depend on the ordering of the elements of Ω^*. Thus, we are justified to write

$$\sum_{x\in\Omega^*} g(x)p_X(x) \qquad \text{rather than} \qquad \sum_{j=1}^{\infty} g(x_j)p_X(x_j)\,.^{89}$$

We go a step further. Since $g(x)p_X(x) = 0$ for $x \notin \Omega^*$, we can omit "$x \in \Omega^*$" and write either of the following:

$$\begin{aligned} \sum_x g(x)p_X(x) &= \sum_{x\in\Omega'} g(x)p_X(x) = \sum_{x\in\Omega^*} g(x)p_X(x) \\ &= \sum_{x:\, p_X(x)>0} g(x)p_X(x) = \sum_{p_X(x)>0} g(x)p_X(x) = \sum_{j=1}^{\infty} g(x_j)p_X(x_j)\,. \end{aligned} \tag{9.4}$$

Choosing $g(x) = 1$, we can express probabilities involving X as follows. If $B \subseteq \Omega'$, then

$$P\{X \in B\} = P_X(B) = \sum_{x\in B} p_X(x) = \sum_{x\in\Omega^*\cap B} p_X(x) = \sum_{x\in B, p_X(x)>0} p_X(x)\,. \ \square \tag{9.5}$$

Problem 9.1. Johnny may choose 2 cookies from a plate with 4 chocolate cookies and 3 oatmeal cookies We write CC for chocolate cookies and OC for oatmeal cookies. Johnny has no preference and picks two cookies at random.

[89] See Remark 3.13 on p.78.

Let $Y :=$ number of CC chosen by Johnny. Find the PMF $p_Y(y)$ for Y.

Solution:

Note that you were not given the domain (sample space) (S, P) of the random variable Y. There is no need to specify it completely. It suffices to know that, since Johnny can choose 2 of the 7 cookies in $\binom{7}{2}$ ways,

(1) $|S| = \binom{7}{2} = \frac{7 \cdot 6}{2!} = 21$. Since selection was at random, $P\{s\} = \frac{1}{21}$ for all $s \in S$.

The codomain can be any set of numbers that contains $0, 1, 2$, because $p_Y(y) = P\{Y = y\} = 0$ for all other numbers y. Thus, our task is to compute $p_Y(0)$, $p_Y(1)$, $p_Y(2)$.

(2) Each selection of y CCs comes with a selection of $2 - y$ OCs

Thus, there are $\binom{4}{y} \cdot \binom{3}{2-y}$ ways to select y CCs and $2 - y$ OCs. $(y = 0, 1, 2.)$

(3)
$$p_Y(0) = \frac{\binom{4}{0} \cdot \binom{3}{2}}{21} = \frac{3}{3 \cdot 7} = \frac{1}{7},$$
$$p_Y(1) = \frac{\binom{4}{1} \cdot \binom{3}{1}}{21} = \frac{4 \cdot 3}{3 \cdot 7} = \frac{4}{7},$$
$$p_Y(2) = \frac{\binom{4}{2} \cdot \binom{3}{0}}{21} = \frac{(4 \cdot 3)/2}{3 \cdot 7} = \frac{2}{7}. \quad \square$$

Whereas a PMF is defined for any discrete random element Y, the next definition needs that the values of Y are numbers.

Definition 9.2 (Expected value of a discrete random variable)**.**

Let Y be a discrete random variable with probability mass function $p_Y(y)$. Then

$$E[Y] := \sum_y y\, p_Y(y) = \sum_y y\, P\{Y = y\},$$

is called the **expected value**, also **expectation** or **mean** of Y. $\square$

Remark 9.3.

A strict definition of $E[Y]$ would explicitly require that the sum $\sum_y y \cdot p_Y(y)$ is absolutely convergent, i.e.,

$$\sum_y |y| p_Y(y) < \infty.$$

The reason: Only absolute convergence of a series guarantees that its value does not depend on the order in which the terms are added. As in WMS and according to Assumption 9.1 on p.188, we will quietly asssume that absolute convergence is satisfied for all random variables for which the expected value is used. $\square$

Proposition 9.2. ★ *Let $A_1, A_2, \ldots, A_n$ a list of mutually disjoint events in a probability space (Ω, P). Let $y_1, y_2, \ldots, y_n \in \mathbb{R}$. Then*

$$E\left[\sum_{j=1}^{n} y_j \mathbf{1}_{A_j}\right] = \sum_{j=1}^{n} y_j P(A_j)\,. \tag{9.6}$$

PROOF: Let $Y := \sum_{j=1}^{n} y_j \mathbf{1}_{A_j}$; let $A := \biguplus_{j=1}^{n}$. We may assume that $A = \Omega$, since we can add the zero term $0 \cdot \mathbf{1}_{A^{\complement}}$ to Y if $A^{\complement} \neq \emptyset$.

We further may assume that all numbers $y_1, \ldots, y_n$ are distinct for the following reason. Assume for example, that $y_{n_1} = y_{n_2} = y_{n_k} = y'$ and that this is the complete list of indices n_j such that $y_{n_j} = y'$. We define $A' := A_{n_1} \uplus A_{n_2} \uplus \cdot \uplus A_{n_k}$. Since

$$\sum_{j=1}^{k} y_{n_j} \mathbf{1}_{A_{n_j}} = \sum_{j=1}^{k} y' \cdot \mathbf{1}_{A_{n_j}} = y' \sum_{j=1}^{k} \mathbf{1}_{A_{n_j}} = y' \cdot \mathbf{1}_{A_{n_j} \uplus \cdots \uplus A_{n_k}} = y' \cdot \mathbf{1}_{A'}\,,$$

we can replace those terms with duplicate y'–values with the single term $y' \cdot \mathbf{1}_{A'}$.

We repeat this procedure with all y–values, even if they occur even once. This way we can write

$$Y = \sum_{j=1}^{m} y_i' \mathbf{1}_{A_i'}\,, \quad \text{where } \Omega = \biguplus_{i=1}^{m} A_i' \text{ and all } y_i' \text{ are distinct.} \tag{9.7}$$

In such a representation of Y, the distinctness of the y_i' implies that

$$Y(\omega) = y_i' \;\Leftrightarrow\; \omega \in A_i' \;\Leftrightarrow\; \{Y = y_i'\} = A_i'\,.$$

In particular, $P\{Y = y_i'\} = P(A_i')$. Thus,

$$E[Y] = E\left[\sum_{i=1}^{m} y_i' \mathbf{1}_{A_i'}\right] = \sum_{y} y' P\{Y = y'\} = \sum_{i=1}^{m} y_i' P\{Y = y_i'\} = \sum_{i=1}^{m} y_i' P(A_i')\,. \tag{9.8}$$

In the last step of the proof we bring back the duplicate y–values. As above, we assume that $y_{n_1} = y_{n_2} = y_{n_k} = y_i'$ and $A_i' := A_{n_1} \uplus A_{n_2} \uplus \cdot \uplus A_{n_k}$. Then

$$y_i' P(A_i') = y_i' P\left(\biguplus_{j=1}^{k} A_{n_j}\right) = y_i' \sum_{j=1}^{k} P(A_{n_j}) = \sum_{j=1}^{k} y_{n_j} P(A_{n_j})\,.$$

We substitute this result in (9.8) and obtain $E[Y] = \sum_{i=1}^{m} \sum_{j=1}^{k} y_{n_j} P(A_{n_j})\,.$

Since $\sum_{i=1}^{m}$ is the summation over all complete groups of equal y–values and each $\sum_{j=1}^{k}$ sums over all items in that group, that double sum equals $\sum_{j=1}^{n} y_j P(A_{n_j})$. Thus, $E[Y] = \sum_{j=1}^{n} y_j P(A_{n_j})\,.$

This proves the proposition. ■

Theorem 9.2.

Let Y be a discrete random variable and $g : \mathbb{R} \to \mathbb{R};\ y \mapsto g(y)$ be a real-valued function. Then the random variable $g \circ Y : \omega \mapsto g\big(Y(\omega)\big)$ has the following expected value:

$$E[g(Y)] \ = \sum_{\text{all } y} g(y)\, p_Y(y) \ = \sum_{\text{all } y} g(y)\, P\{Y = y\}\,. \tag{9.9}$$

PROOF: We give the proof asuming that Y takes only finitely many distinct values $y_1, y_2, \ldots, y_n$. [90]
Let $\{z_1, z_2, \ldots, z_m\}$ denote the set of all distinct function values $g(y_i)$, $i = 1, \ldots, n$. In general, $m \le n$ rather than $m = n$, because is possible for one or more of the arguments y to have the same function value $g(y)$.
For $j = 1, \ldots, m$, let

$$I_j \ := \ \{\, i \in [1, n] \ : \ g(y_i) = z_j \,\}$$

denote the set of all those indices i such that g assigns y_i to the same function value z_j. Note that

(1) each I_j contains at least one index.
(2) The index sets I_j form a partition of the indices i for the arguments y_i of g:

$$[1, n] \ = \ I_1 \uplus I_2 \uplus \cdots \uplus I_m\,. \tag{A}$$

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let

$$B_i \ := \ \{Y = y_i\} \ = \ \{\omega \in \Omega : Y(\omega) = y_i\}\,; \quad C_j \ := \ \{Z = z_j\} \ = \ \{\omega \in \Omega : Z(\omega) = z_j\}\,. \tag{B}$$

Since $\omega \in C_j \ \Leftrightarrow \ Z(\omega) = z_j \overset{\text{(B)}}{\Leftrightarrow} Y(\omega) = y_i$ for some $i \in I_j \ \Leftrightarrow \ \omega \in \biguplus\limits_{i \in I_j} B_i$, it follows that

$$C_j \ = \ \biguplus_{i \in I_j} B_i\,. \tag{C}$$

We have for Y and Z the representations

$$Z(\omega) \ = \sum_{j=1}^{m} z_j \mathbf{1}_{\{Z = z_j\}}(\omega) \ = \sum_{j=1}^{m} z_j \mathbf{1}_{C_j}(\omega) \overset{\text{(C)}}{=} \sum_{j=1}^{m} z_j \mathbf{1}_{\biguplus_{i \in I_j} B_i}(\omega) \ = \ \sum_{j=1}^{m} z_j \sum_{i \in I_j} \mathbf{1}_{B_i}(\omega)\,. \tag{D}$$

Here the last equation holds because the indicator function of a disjoint union is the sum of the indicator functions. That is a triviality which has been noted in (2.67) on p.54.
Since $g(y_i) = \text{const} = z_j$ for all $i \in I_j$, we can rewrite that last sum as

$$Z(\omega) \ = \sum_{j=1}^{m} \sum_{i \in I_j} g(y_i) \mathbf{1}_{B_i}(\omega) \overset{\text{(A)}}{=} \sum_{i=1}^{n} g(y_i) \mathbf{1}_{B_i}(\omega)\,. \tag{E}$$

[90] As an aside, note that $y \mapsto g(y)$ need not be defined for all $y \in \mathbb{R}$. It suffices that the domain of g contains $Y(\Omega) = \{Y(\omega) : \omega \in Y(\Omega)\}$. (The range of the function Y; see Definition 2.17 on p.33.)

We conclude from **(D)** and **(E)** that $E[Y] = E\left[\sum_{i=1}^{n} g(y_i)\mathbf{1}_{B_i}\right]$.

Finally, we apply Proposition 9.2 on p.191 and obtain, since $B_i = \{Y = y_i\}$,

$$E[Y] = \sum_{i=1}^{n} g(y_i)P(B_i) = \sum_{i=1}^{n} g(y_i)P\{Y = y_i\}. \blacksquare$$

ALTERNATE PROOF – based on Ch. 6 (Advanced Topics – Measure and Probability):
(9.9) is formula (6.54) of Remark 6.18 on p.166. ■

The following corresponds to WMS Theorems 3.4 and 3.5.

Theorem 9.3.

Let $c \in \mathbb{R}$, Y be a discrete random variable and $g_1, g_2, g_n : \mathbb{R} \to \mathbb{R}$ be a list of n real-valued functions. Then

$$E[c] = c \quad \text{and} \quad E[cY] = cE[Y], \tag{9.10}$$

$$E[cg_j(Y)] = cE[g_j(Y)]. \tag{9.11}$$

Further, the random variable

$$\sum_{j=1}^{n} g_j \circ Y : \Omega \longrightarrow \mathbb{R}; \qquad \omega \mapsto \sum_{j=1}^{n} g_j\big(Y(\omega)\big)$$

has the following expected value:

$$E\left[\sum_{j=1}^{n} g_j \circ Y\right] = \sum_{j=1}^{n} E[g_j \circ Y]. \tag{9.12}$$

PROOF: Let Z denote the random variable $Z = c : \omega \mapsto c$, then

$$P\{Z = z\} = \begin{cases} 1, & \text{if } y = c, \\ 0, & \text{if } y \neq c. \end{cases}$$

Thus, $E[Z] = \sum_{z:P_Z\{z\}>0} z \cdot P_Z\{z\} = c \cdot 1 = c$. This proves the first half of (9.10).

For the proof of the second half, note that $c = 0$ implies $cY = 0$. Thus, $E[cY] = cE[Y]$ becomes $E[0] = 0$, and we covered that case already. So we may assume that $c \neq 0$.

Let $Y' := cY$ and $y' := cy$. Then $Y'(\omega) = y' \Leftrightarrow Y(\omega) = \frac{y'}{c}$. Thus, $P\{Y' = y'\} = P\{Y = \frac{y'}{c}\}$. Thus,

$$E[cY] = E[Y'] = \sum_{y'} y' \cdot P\{Y' = y'\} = \sum_{y'} y' \cdot P\{Y = \frac{y'}{c}\}$$

$$= \sum_{y} c \cdot \frac{y'}{c} \cdot P\{Y = \frac{y'}{c}\} = c \cdot \sum_{y} y \cdot P\{Y = y\} = c \cdot E[Y].$$

This proves the second half of (9.10). We apply this formula with $g_j(Y)$ in place of Y and (9.11) follows.

Finally, we apply Theorem 9.4 with $g_j \circ Y$ in place of Y_j. [91] This results in (9.12). ■

ALTERNATE PROOF – based on Ch. 6 (Advanced Topics – Measure and Probability):

Since expectations $E[Y]$ are abstract integrals $\int Y dP$ (see Definition 6.15 (Expected value of a random variable) on p.165, all assertions follow from Theorem 6.9 on p.159. ■

The following cannot be found in the WMS text.

Theorem 9.4.

Let $Y_1, Y_2, \ldots, Y_n : \Omega \to \mathbb{R}$ be discrete random variables which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$). Then the random variable

$$\sum_{j=1}^{n} Y_j : \Omega \longrightarrow \mathbb{R}; \qquad \omega \mapsto \sum_{j=1}^{n} Y_j(\omega)$$

has the following expected value:

$$E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j]\,. \tag{9.13}$$

In other words, the expectation of the sum is the sum of the expectations.

PROOF: ⋆ There are finite or infinite sequences $x_i, y_j \in \mathbb{R}$ as follows. Let $A_i := \{X = x_i\}$ and $B_j := \{Y = y_j\}$. Then the A_i are disjoint, the B_j are disjoint, and $A_* := (\biguplus_i A_i)^{\complement}$, , $B_* := \left(\biguplus_j B_j\right)^{\complement}$ have probability zero. We may assume that $X(\omega) = 0$ for $\omega \in A_*$ and $Y(\omega) = 0$ for $\omega \in B_*$, since that does not change any assertions that are based on probabilities, such as the taking of expected values: Being able to discard the expressions $\biguplus_i A_i$ and $\biguplus_j B_j$ considerably simplifies the proof.

For example, this assumption allows us to write, without having to exclude any $\omega \in \Omega$,

$$\textbf{(A)} \qquad X(\omega) \;=\; \sum_i x_i \mathbf{1}_{\{X=x_i\}}(\omega)\,, \qquad Y(\omega) \;=\; \sum_j y_j \mathbf{1}_{\{Y=y_j\}}(\omega)\,.$$

If $P\{X = 0\} > 0$, then we include 0 as one of the x_i and if $P\{Y = 0\} > 0$, then we include 0 as one of the y_j. We do so even though $0 \cdot \mathbf{1}_{\{X=0\}} = 0 \cdot \mathbf{1}_{\{Y=0\}} = 0$ contributes nothing to those sums, since then

$$A_i \;:=\; \{X = x_i\}\,, \qquad B_j \;:=\; \{Y = y_j\}_j\,, \qquad C_{i,j} \;:=\; A_i \cap B_j$$

form partitions $\biguplus_i A_i \;=\; \biguplus_j B_j \;=\; \biguplus_{i,j} C_{i,j} \;=\; \Omega$ of Ω. Moreover, for each i, j,

$$\textbf{(B)} \qquad \begin{aligned} A_i \;&=\; \biguplus_k C_{i,k} \quad \text{and} \quad B_j \;=\; \biguplus_k C_{k,j}\,, \\ \text{which implies} \qquad \mathbf{1}_{A_i} \;&=\; \sum_k \mathbf{1}_{C_{i,k}} \quad \text{and} \quad \mathbf{1}_{B_j} \;=\; \sum_k \mathbf{1}_{C_{k,j}}\,. \end{aligned}$$

[91] The proof of that theorem does not make use of this current one.

$$\text{Since} \quad X \overset{\text{(A)}}{=} \sum_i x_i \mathbf{1}_{A_i} \overset{\text{(B)}}{=} \sum_{i,j} x_i \mathbf{1}_{C_{i,j}} \qquad Y \overset{\text{(A)}}{=} \sum_j y_j \mathbf{1}_{B_j} \overset{\text{(B)}}{=} \sum_{i,j} y_j \mathbf{1}_{C_{i,j}}$$
$$\text{and thus,} \quad X+Y = \sum_{i,j} x_i \mathbf{1}_{C_{i,j}} + \sum_{i,j} y_j \mathbf{1}_{C_{i,j}} = \sum_{i,j} (x_i + y_j) \mathbf{1}_{C_{i,j}} ,$$

tt follows from Prop.9.2 on p.191, that

$$\textbf{(C)} \qquad E[X] = \sum_{i,j} x_i P(C_{i,j}) , \quad E[Y] = \sum_{i,j} y_j P(C_{i,j}) , \quad E[X+Y] = \sum_{i,j} (x_i + y_j) P(C_{i,j}) .$$

We conclude the proof as follows:

$$E[X+Y] \overset{\text{(C)}}{=} \sum_{i,j} (x_i + y_j) P(C_{i,j}) = \sum_{i,j} x_i P(C_{i,j}) + \sum_{i,j} y_j P(C_{i,j}) \overset{\text{(C)}}{=} E[X] + E[Y] . \blacksquare$$

ALTERNATE PROOF – based on Ch. 6 (Advanced Topics – Measure and Probability):
Since expectations $E[Y]$ are abstract integrals $\int Y dP$ (see Definition 6.15 (Expected value of a random variable) on p.165, this follows from the linearity of the $\int \dots dP$. See Theorem 6.9 on p.159.

Remark 9.4.

(1) The last theorem encompasses all variants of Theorem 9.3. For example, (9.12) follows with $Y_j = g_j \circ Y$.

(2) The reason that many texts on an undergraduate probability theory do not list this theorem is that the proof, though elementary, is very tedious and requires working with the PMF of the random element $\vec{Y} = (Y_1, \dots, Y_n)$, given by

$$p_{\vec{Y}}(\vec{y}) = P\{Y_1 = y_1, \dots, Y_n = y_n\} \ \square$$

Variance and standard deviation of a random variable indicate how strongly its distribution is concentrated around its expected value.

Definition 9.3 (Variance and standard deviation of a random variable)**.**

Y be a random variable. The **variance** of Y is defined as the expected value of $(Y - E[Y])^2$. In other words,

(9.14) $$Var[Y] := \sigma_Y^2 := E\big[(Y - E[Y])^2\big] .$$

We call $SD(Y) := \sigma_Y := \sqrt{Var[Y]}$ The **standard deviation** of Y. $\square$

Theorem 9.5.

If Y is a discrete random variable, then

$$Var[Y] = E\big[Y^2\big] - \big(E[Y]\big)^2 .$$

PROOF:

$$\begin{aligned} Var[Y] &= E[(Y - E[Y])^2] \;=\; E(Y^2 - 2(E[Y]\cdot Y) + (E[Y])^2 \\ &= E(Y^2) - (2E[Y])E[Y] + (E[Y])^2 \;=\; E(Y^2) - (E[Y])^2\,. \;\blacksquare \end{aligned}$$

Theorem 9.6.

Let Y be a discrete random variable and $a, b \in \mathbb{R}$. Then

$$Var\,[aY + b] \;= a^2 Var[Y]\,. \tag{9.15}$$

In other words, shifting a random variable by b, leaves its variance unchanged and multiplying it by a constant multiplies its variance by the square of that constant.

PROOF: We prove this by first showing that, for random variables Y and Y',

$$Var[aY] \;=\; a^2 Var[Y] \qquad \text{and} \qquad Var[Y' + b] \;=\; Var[Y']$$

The assertion then follows from replacing Y' with aY.
We obtain from (9.10) that

$$Var[aY] \;= E[a^2Y^2] - \big(E[aY]\big)^2 \;=\; a^2E[Y^2] - \big(aE[Y]\big)^2 \;=\; a^2\big(E[Y^2] - (E[Y])^2\big) \;=\; a^2 Var[Y]\,.$$

To prove that $Var[Y' + b] = Var[Y']$, we observe that for any random variable Z and constant a, $E[Z + a] = E[Z] + E[a] = E[Z] + a$. Thus,

$$\begin{aligned} Var[Y' + b] &= E\left[\big((Y' + b) - E[Y' + b]\big)^2\right] \\ &= E\left[\big((Y' + b) - (E[Y'] + b)\big)^2\right] \;=\; E\left[\big(Y' - E[Y']\big)^2\right] \;=\; Var[Y']\,. \;\blacksquare \end{aligned}$$

Remark 9.5. Since $\sqrt{a^2} = -a$ for negative numbers a,

$$\sigma(aY) \;= |a|\sigma(Y)\,. \;\square \tag{9.16}$$

The following cannot be found in the WMS text.

Theorem 9.7 (Bienaymé formula)**.**

Let $Y_1, Y_2, \ldots, Y_n : \Omega \to \mathbb{R}$ be independent discrete random variables which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$). Here we take the naive definition of independence: The outcomes of any Y_k are not influenced by the outcomes of the other Y_j. We will give a formulation of independence in terms of probabilities in a later chapter. Then

$$Var\left[\sum_{j=1}^{n} Y_j\right] \;= \sum_{j=1}^{n} Var[Y_j]\,. \tag{9.17}$$

In other words, for independent random variables, the variance of the sum is the sum of the variances.

PROOF: Will be given later as part of Corollarycor-x:uncorrel-bienayme-447 (Bienaymé formula for uncorrelated variables) on p.261. ■

Remark 9.6. The independence is necessary, otherwise there are counterexamples:
If $Y_1 = Y_2 = Y$ for some random variable Y, then

$$Var[Y+Y] \;=\; Var[2Y] \;=\; 4Var[Y] \;\neq\; Var[Y] + Var[Y]\,. \;\square$$

9.2 Bernoulli Variables and the Binomial Distribution

Definition 9.4 (Bernoulli trials and variables)**.**

A **Bernoulli trial**. is a random element with only two outcomes, such as
⊡ S (success) or F (failure) ⊡ T (true) or F (false) ⊡ Y (Yes) or N (No) ⊡ 1 or 0
- We call $p := P\{X = \text{success}\}$ the **success probability**
and $q := 1 - p = P\{X = \text{failure}\}$ the **failure probability** of the Bernoulli trial.

- If a Bernoulli trial X has outcomes 1 and 0, then we call X a **Bernoulli variable** or a **0–1 encoded Bernoulli trial**.
- A **Bernoulli sequence** is an iid sequence (Def. 5.18 on p.136) of Benoulli trials. □

Remark 9.7.
(a) The entire distribution of a Bernoulli trial is determined by the value of its success probability.
(b) Note that the definition of a Bernoulli sequence $(X_j)_j$ implies that
(1) the X_j are independent
(2) each X_j has the same success and failure probabilities. We write p and q for those numbers.

(c) Unless stated otherwise, we interpret the value 0 of a 0–1 encoded Bernoulli trial as failure and the value 1 as success. □

Theorem 9.8 (Expected value and variance of a 0–1 encoded Bernoulli trial)**.**

Let X be a 0–1 encoded Bernoulli trial with $p := P\{X = 1\}$. Then

(9.18) $$E[X] \;=\; p \qquad \textit{and} \qquad Var[X] \;=\; pq\,.$$

PROOF:
$E[X] \;=\; 0q \,+\, 1\cdot p \;=\; p.$
For the variance, $Var[X] \;=\; E[X^2] \,-\, (E[X])^2 \;=\; E[X^2] \,-\, p^2$. Further,

$$E[X^2] \;=\; 0^2\cdot q \,+\, 1^2\cdot p \;=\; p.$$

Hence, $Var[X] \;=\; p - p^2 \;=\; p(1-p) \;=\; pq.$ ■

Definition 9.5 (Binomial Distribution)**.**

Let $n \in \mathbb{N}$ and $0 \le p \le 1$. Let Y be a random variable with probability mass function

$$p_Y(y) \;=\; \binom{n}{y} p^y\, q^{n-y}\,. \tag{9.19}$$

Then we say that Y has a **binomial distribution**. with parameters n and p or, in short, a **binom(n,p) distribution**. We also say that Y is binom(n,p). □

Remark 9.8. How does one see that p_Y of (9.19) satisfies $p_Y(y) \ge 0$ for all y and $\sum_y p_Y(y) = 1$, i.e., it really is a probability mass function?

- $p_Y(y) \ge 0$ is true, since $p, q, \binom{n}{y} \ge 0$.
- We apply the binomial theorem (see Theorem 7.5) to $(p+q)^n$ and obtain

$$1 \;=\; 1^n \;=\; (p+q)^n \;=\; \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}\,. \quad \square$$

Theorem 9.9.

Let X_1, X_2, X_n be a Bernoulli sequence of size n with success probability p. Let Y be the number of successes in that sequence, i.e., $Y(\omega)$ = number of indices j such that $X_j(\omega) = S$.

- *Then Y is binom(n,p).*

PROOF: Clearly,

$$Y(\omega) \;=\; y \;\Leftrightarrow\; \begin{cases} X_j(\omega) = S & \text{for } y \text{ indices } j, \\ X_j(\omega) = F & \text{for } n-y \text{ indices } j. \end{cases}$$

Let $\vec{x} := (x_1, \ldots, x_n)$ a vector that consists of y components S and $n-y$ components F. For such an arrangement $\vec{x}$ of y successes and $n-y$ failures, let n_1, n_2, n_y denote the indices for which $X_{n_j} = S$ and m_1, m_2, m_{n-y} those indices for which $X_{m_j} = F$. Further, let $A(\vec{x})$ denote the event

$$A(\vec{x}) \;:= \{X_1 = x_1\,, X_2 = x_2, \ldots, X_n = x_n\}\,.$$

Then independence of the Bernoulli trials X_j and thus, of the events $\{X_j = x_j\}$, yields

$$\begin{aligned} P\big(A(\vec{x})\big) \;&= P\big(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\}\big) \;=\; P\{X_1 = x_1\} \cdot P\{X_2 = x_2\} \cdots P\{X_n = x_n\} \\ &= P\{X_{n_1} = S\} \cdots P\{X_{n_y} = S\} \cdot P\{X_{m_1} = F\} \cdots P\{X_{m_{n-y}} = F\} \;=\; p_y \cdot q^{n-y}\,. \end{aligned} \tag{A}$$

There are as many different vectors $\vec{x}$ with y successes and $n-y$ failures as there are ways to form different lists of size n consisting of y items S and $n-y$ items F. That number is $\binom{n}{y}$.

We observe that the events $A(\vec{x})$ and $A(\vec{x}')$ are disjoint for different $\vec{x}$ and $\vec{x}'$, since this means that there is at least one index $j$such that either $x_j = S$ and $x'_j = F$ or the other way around.

Let us assume that $x_j = S$ and $x'_j = F$. If $\omega \in A(\vec{x})$, then $X_j(\omega) = S$ But then $\omega \notin A(\vec{x}')$, since then $X_j(\omega)$ would have to be F. Thus, $A(\vec{x}) \cap A(\vec{x}') = \emptyset$. The case that $x_j = F$ and $x'_j = S$ is handled in

the same fashion. Since

$$\{Y = y\} \;=\; \biguplus_{\vec{x}} A(\vec{x})$$

where $\vec{x}$ assumes all $\binom{n}{y}$ arrangements of y successes and $n-y$ failures, it follows that

$$P\{Y = y\} \;=\; \sum_{\vec{x}} A(\vec{x}) P\big(A(\vec{x}) \overset{\text{(A)}}{=} \binom{n}{y} p^y q^{n-y}\,.$$

This last expression equals the PMF of a binom(n,p) distribution and this concludes the proof. ■

Theorem 9.10 (Expected value and variance of a binom(n,p) variable)**.**

Let Y be a binom(n,p) variable. Then

(9.20) $$E[Y] \;=\; np \qquad \textit{and} \qquad Var[Y] \;=\; npq\,.$$

PROOF: Let $X_1, \ldots, X_n$ be an iid list of 0–1 encoded Bernoulli trials with $p := P\{X = 1\}$. Let $Y' := \sum_{j=1}^{n} X_j$. according to Theorem 9.8, Theorem 9.4 on p.194, and, since the X_j are independent, Theorem 9.7 (Bienaymé formula) on p.196,

$E[Y'] \;=\; \sum_{j=1}^{n} E[X_j] \;=\; np$ and $Var[Y'] \;=\; \sum_{j=1}^{n} Var[X_j] \;=\; npq.$

Further, $Y' = y \;\Leftrightarrow$ exactly y of the X_j have outcome y. Thus, Y' denotes the number of successes of those Bernoulli trials. Acccording to Theorem 9.9 on p.198, Y' has a binom(n,p) distribution. Since expected value and variance of a discrete random variable are determined by its PMF, $E[Y] \;=\; E[Y'] \;=\; np$ and $Var[Y] \;=\; Var[Y] \;=\; npq.$ ■

9.3 Geometric + Negative Binomial + Hypergeometric Distributions

Definition 9.6 (Geometric distribution)**.**

A random variable Y is said to have a **geometric distribution** with parameter $0 \le p \le 1$ or, in short, a **geom(p) distribution**, if its probability mass functions is as follows:

(9.21) $$p_Y(y) \;=\; q^{y-1}\,p\,, \qquad \text{for } y = 1, 2, 3, \ldots\,. \;\square$$

Theorem 9.11. *Let $X_1, X_2, \cdots : (\Omega, P) \to \{S, F\}$ be an infinite Bernoulli sequence with success probability $0 \le p \le 1$.*

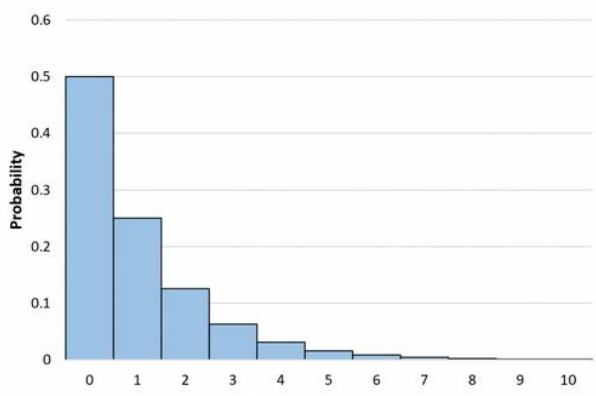
Let $T(\Omega, P) \to \mathbb{N}$ be the random variable

$$T(\omega) := \begin{cases} \text{smallest integer } k > 0 \text{ such that } X_k(\omega) = S & \text{if such a } k \text{ exists,} \\ \infty\,, & \text{else.} \end{cases}$$

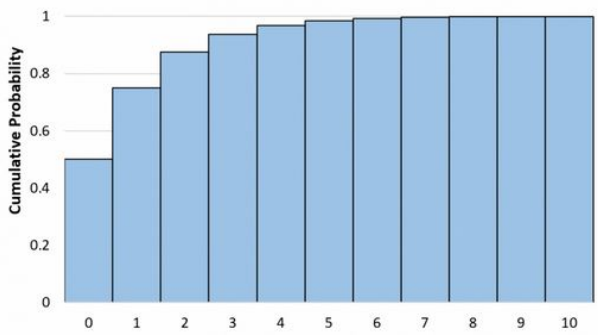
- *Then T is geom(p).*

PROOF: Since $T(\omega) = n \Leftrightarrow X_1(\omega) = X_2(\omega) = X_{n-1}(\omega) = F$ and $X_n(\omega) = S$ and the independence of the X_i implies that the events $\{X_1 = F\}$, $\{X_2 = F\}, \{X_{n-1} = F\}, \{X_n = S\}$, are independent, we obtain

$$\begin{aligned} P\{X_1 = F,\, X_2 = & F, X_{n-1} = F, X_n = S\} \;=\; P\{X_1 = F\} \cap \cdots \{X_{n-1} = F\} \cap \{X_n = S\} \\ & = P\{X_1 = F\} \cdot P\{X_2 = F\} \cdots P\{X_{n-1} = F\} \cdot P\{X_n = S\} \;=\; q^{n-1}\, p\,. \;\blacksquare \end{aligned}$$



9.1 (Figure)**. PMF for geom(**0.5**).**



9.2 (Figure)**. CDF for geom(**0.5**).**

Remark 9.9. In Theorem **??** we wrote $T(\omega)$ rather than the usual $Y(\omega)$ for the following reason. If we interpret the index j of the Bernoulli trial X_j as the point in time when the jth trial takes place, then $\omega \mapsto T(\omega)$ represents a **random time**, the time at which the first success happens. □

Theorem 9.12 (WMS Ch.03.5, Theorem 3.8)**.**

If Y is a geom(p) random variable, then

$$E[Y] \;=\; \frac{1}{p}\,, \qquad \textit{and} \qquad Var[Y] \;=\; \frac{q}{p^2}\,.$$

PROOF:

A: Expectation $E[Y]$:

One can obtain the derivative of the series $\sum_{y=1}^{\infty} q^y$ by differentiating it term–by==term. Since

$$\frac{d}{dq} q^y \;=\; y q^{y-1}\,,$$

it follows that

(**A**)
$$\frac{d}{dq}\left(\sum_{y=1}^{\infty} q^y\right) \;=\; \sum_{y=1}^{\infty} y q^{y-1}\,.$$

We use **(A)** as follows.

$$\begin{aligned} E(Y) \;&=\; \sum_{y=1}^{\infty} y p_Y(y) \;=\; \sum_{y=1}^{\infty} y q^{y-1} p \;=\; p \sum_{y=1}^{\infty} y q^{y-1} \;\overset{\text{(A)}}{=}\; p\,\frac{d}{dq}\left(\sum_{y=1}^{\infty} q^y\right) \\ &= p\,\frac{d}{dq}\left(\frac{q}{1-q}\right) \;=\; p\,\frac{1\cdot(1-q) \;-\; q(-1)}{(1-q)^2} \;=\; p\,\frac{1}{p^2} \;=\; \frac{1}{p}\,. \end{aligned}$$

B: Variance $Var[Y]$: [92]

[6] Kargin, Vladislav: BU Lecture Notes for the Introduction to Probability Course

We compute the variance by again interchanging differentiation and summation. It follows from

$$\frac{d^2}{dq^2} q^y \;=\; y(y-1) q^{y-2}\,,$$

that

(**B**)
$$\frac{d^2}{dq^2}\left(\sum_{y=1}^{\infty} q^y\right) \;=\; \sum_{y=2}^{\infty} y(y-1) q^{y-2} \;=\; \frac{1}{pq} \sum_{y=2}^{\infty} y(y-1) q^{y-1}\cdot p\,.$$

We use **(B)** as follows.

$$\begin{aligned} E[Y(Y-1)] \;&=\; \sum_{y=1}^{\infty} y(y-1) p_Y(y) \;=\; \sum_{y=2}^{\infty} y(y-1) q^{y-1} p \;=\; pq \sum_{y=2}^{\infty} y(y-1) q^{y-2} \\ &\overset{\text{(B)}}{=} pq\,\frac{d^2}{dq^2}\left(\sum_{y=2}^{\infty} q^y\right) \;=\; pq\cdot\frac{d^2}{dq^2}\left(\sum_{y=0}^{\infty} q^y\right) \;=\; pq\cdot\frac{d^2}{dq^2}\left(\frac{1}{1-q}\right) \\ &= pq\cdot\frac{d}{dq}\left(\frac{-1}{(1-q)^2}\right) \;=\; pq\cdot\frac{2}{p^3} \;=\; \frac{2q}{p^2}\,. \;\blacksquare \end{aligned}$$

Since $Var[Y] \;=\; E[Y^2] - \big(E[Y]\big)^2 \;=\; E[Y^2] - (1/p)^2$, we conclude that

$$\begin{aligned} Var[Y] \;&=\; \big(E[Y^2] \;-\; E[Y]\big) \;-\; \left(\frac{1}{p}\right)^2 \;+\; E[Y] \;=\; E[Y(Y-1)] \;-\; \left(\frac{1}{p}\right)^2 \;+\; \frac{1}{p} \\ &= \frac{2q}{p^2} \;-\; \frac{1}{p^2} \;+\; \frac{p}{p^2} \;=\; \frac{2q \;-\; (1-p)}{p^2} \;=\; \frac{q}{p^2} \;\blacksquare \end{aligned}$$

[92] Source: [6] Kargin, Vladislav: BU Lecture Notes for the Introduction to Probability Course

Definition 9.7 (Negative binomial distribution). ⋆

A random variable Y has a **negative binomial distribution** with parameters p and r if

(9.22) $\quad p_Y(y) \;=\; \binom{y-1}{r-1} p^r q^{y-r}\,, \quad$ where $\; r \in \mathbb{N},\;\; y = r, r+1, r+2, \ldots,\;\; 0 \le p \le 1\,.$ □

This last definition has been marked as ⋆, so you are not expected to recall p_Y from memory. In contrast, the next theorem is NOT optional.

Theorem 9.13. *Let $X_1, X_2, \cdots : (\Omega, P) \to \{S, F\}$ be an infinite Bernoulli sequence with success probability $0 \le p \le 1$.*

Let $t_1 < t_2 < \cdots$ be the subsequence of those indices at which a success happens. In other words,

$$X_n(\omega) \;=\; \begin{cases} S \;=\; \textit{success} \;\; \textit{if } n \textit{ is one of } t_1, t_2, \ldots, \\ F \;=\; \textit{failure}\,, \qquad \textit{else.} \end{cases}$$

Two points to note:

- *There will be different subsequences $t_1, t_2, \ldots$ for different arguments $\omega \in \Omega$. In other words, we are dealing with a sequence of random variables(!)*

$$t_1 = T_1(\omega),\; t_2 = T_2(\omega),\; t_3 = T_3(\omega),\; \ldots$$

- *It is possible that we are dealing with an ω for which there are only 18 successes in the entire (infinite) sequence $X_1(\omega), X_2(\omega), \ldots$. In this case, we define $T_{19}(\omega) = T_{20}(\omega) = \cdots = \infty$.*
  More generally, if $r \in \mathbb{N}$ and the sequence $X_1(\omega), X_2(\omega), \ldots$ has less than r successes, we define

$$T_r(\omega) \;:=\; \infty\,.$$

Now that we have defined $T_r = T_r(\omega)$, we are ready to state the theorem.

- *The random variable T_r has a negative binomial distribution with parameters p and r.*

PROOF: We define the following events.

- $A \;:=\; \{T_r = t\} \;=\; \{$ success $\#r$ happens at time $t\}$
- $B \;:=\; \{$ there are $r-1$ successes before $t\}$
- $C \;:=\; \{X_t = \text{success}\} \;=\; \{$ there is a success at $t\}$

Note that B only depends on the random variables $X_1, \ldots, X_{t-1}$ and C only depends on X_t.

Since the X_t are independent, B and C are independent. Thus, $P(B \cap C) = P(B) \cdot P(C)$. Moreover, $A = B \cap C$, since a moment's reflection shows that

- success $\#r$ happens at time $t \;\Leftrightarrow\;$ there are $r-1$ successes before t and a success happens at t.

Thus, $P(A) = P(B) \cdot P(C)$. From all the above, it follows that

(⋆) $\qquad P\{T_r = t\} \;= P(A) \;=\; P(B) \cdot P(C) \;=\; P(B) \cdot P\{X_t = \text{ success}\}\,.$

Since the number of successes up to time $t-1$follows a binom$(t-1,p)$ distribution and X_t is Bernoulli with success probability p, we see that

$$P(B) \;=\; \binom{t-1}{r-1} p^{r-1}\, q^{(t-1)-(r-1)} \;=\; \binom{t-1}{r-1} p^{r-1}\, q^{t-r} \qquad \text{and} \qquad P\{X_t = \text{ success }\} \;=\; p\,.$$

It follows from $(\star)$ that

$$p_{T_r}(t) \;=\; P\{T_r = t\} \;=\; \binom{t-1}{r-1} p^{r-1}\, q^{t-r} \cdot p \;=\; \binom{t-1}{r-1} p^{r}\, q^{t-r}\,.$$

This matches (**??**)f Definition 9.7 (Negative binomial distribution) on p.202 if we replace T_r with Y and t with y. ■

Remark 9.10. If we think of the indices n of the sequence X_n as points in time, we can interpret the random variables $T_1, T_2, \ldots$ as follows.

- T_r is the time of the rth success in the underlying Bernoulli sequence X_n. □

Theorem 9.14. ⋆

If the random variable Y is negative binomial with parameters p and r,

$$E[Y] \;=\; \frac{r}{p} \qquad \textit{and} \qquad Var[Y] \;=\; \frac{r(1-p)}{p^2}\,.$$

PROOF: Not given here. ■

Definition 9.8 (Hypergeometric distribution)**.**

A random variable Y has a **hypergeometric distribution** with parameters N, R and n if its PMF is

$$p_Y(y) \;=\; \frac{\binom{R}{y}\binom{N-R}{n-y}}{\binom{N}{n}}\,, \tag{9.23}$$

where the nonnegative integers N, R, n and y are subject to the following conditions:

- $y \le n$
- $y \le R$
- $n - y \le N - R$ □

Remark 9.11. For the following you should review Section 8.2 (Sampling and Urn Models With and Without Replacement).

The hypergeometric distribution provides the mathematical model for drawing SRS samples of size n from a population of size N where each item in that population is classified as either S (success) or F (failure).

In contrast to the scenarios involving the binomial, geometric and negative binomial distributions, those n picks $X_1, X_2, \ldots, X_n$ do NOT constitute a Bernoulli sequence since SRS sampling is sampling without replacement and the X_j will neither be independent nor have the same success probability across all j.

Rather, we must model this kind of sampling with an urn model without replacement. See Definition 8.5 (Urn models) on p.187. It simplifies matters greatly that we are only interested in success or failure of each sample pick, since this means that we can model our population as N well–mixed balls in an urn, of which R are labeled S and the remaining $N - R$ are labeled F. Picking the SRS sample of size n from the population then is modeled by picking a sample of size n without replacement from that urn. □

Theorem 9.15.

- *Given is an urn wich contains N well–mixed balls of two colors, Red and Black. We assume that R are Red and thus, the remaining $N - R$ are Black.*
- *A sample of size n is drawn without replacement from that urn, according to Definition 8.5**(a)**.*

Let the random variable Y denote the number of Red balls in that sample. Then Y is hypergeometric with parameters N, R and n. In other words, its PMF is

$$p_Y(y) \;=\; \frac{\binom{R}{y}\binom{N-R}{n-y}}{\binom{N}{n}}\,.$$

PROOF: We give here a very skeletal proof. For more detail consult WMS Chapter 3.7.

We are not interested in the order in which those Red balls were picked, so our probability space Ω will be that of all combinations of size n that can be selected from N balls. Thus,

$$|\Omega| \;=\; \binom{N}{n}\,.$$

$p_Y(y)$ is the probability of selecting exactly y Red balls in the sample of size n Such a selection is obtained by partitioning the N balls into the heap of all R red balls, the heap of all $N-R$ Black balls and then proceding as follows.

Conceptually we pick one of the $\binom{R}{y}$ possible selections of y items from the R red balls and then complementing it with one of the $\binom{N-R}{n-y}$ possible selections of the remaining $n - y$ items from the $N - R$ black balls. By Theorem 7.1 (multiplication rule of combinatorial analysis) on p.168, there are $\binom{R}{y} \cdot \binom{N-R}{n-y}$ such selections. It follows that

$$p_Y(y) \;= P\{Y = y\} \;=\; \frac{\binom{R}{y} \cdot \binom{N-R}{n-y}}{\binom{N}{n}}\,.$$

It follows that Y is hypergeometric with parameters N, R and n. ■

Theorem 9.16 (WMS Ch.03.7, Theorem 3.10)**.** ⋆

Let Y be a hypergeometric random variable with parameters N, R and n. Then

$$E[Y] = \frac{nR}{N} \quad \textit{and} \quad Var[Y] = n\left(\frac{R}{N}\right)\left(\frac{N-R}{N}\right)\left(\frac{N-n}{N-1}\right). \tag{9.24}$$

PROOF: We reproduce here the plausibility argument given by WMS in their "proof" of WMS Theorem 3.10.

Since we consider picking an R–item as a success, the above formulas read with $p := \frac{R}{N}$ and $q = 1-p = \frac{N-R}{N}$ as follows:

$$E[Y] = n \cdot p \quad \text{and} \quad Var[Y] = n \cdot p \cdot q\left(\frac{N-n}{N-1}\right).$$

Except for the factor $(N-n)/(N-1)$

those are expectation and variance of the binom$(n, R/n)$ distribution. Note for the

$$\textbf{correction factor} \quad \frac{N-n}{N-1}, \qquad \text{that} \qquad \lim_{N\to\infty}\frac{N-n}{N-1} = 1.$$

This reflects the fact that, if N is huge in comparison to n, drawing from an urn with or without replacement yields, up to a rounding error, the same probabilities. ■

9.4 The Poisson Distribution

We start out with the simple observation that $e^x = \sum_{j=0}^{\infty}\frac{x^j}{j!}$ for any $x \in \mathbb{R}$.

Proposition 9.3. *Let $\lambda > 0$. Then the function $p(y) := e^{-\lambda}\frac{\lambda^y}{y!}$ defines a probability mass function on $[0,\infty[_{\mathbb{Z}} = \{0,1,2,\dots\}$.*

PROOF: Obviously, $p(y) \geq 0$ for all y.

To show that $\sum_y p(y) = 1$, we apply the formula $e^x = \sum_{j=0}^{\infty}\frac{x^j}{j!}$, which is true for any $x \in \mathbb{R}$, with $x = \lambda$ and $j = y$. ■

This simple proposition enables us to make the following definition.

Definition 9.9 (Poisson variable)**.**

Let Y be a random variable and $\lambda > 0$. We say that Y has a **Poisson probability distribution** with parameter λ, in short, Y is **poisson(λ)**, if its probability mass function is

$$p_Y(y) = \frac{\lambda^y}{y!}e^{-\lambda}, \qquad \text{for } y = 0,1,2,\dots, \ \square$$

We follow WMS Chapter 3.8 to show what phenomena can be modeled by a Poisson variables

Proposition 9.4. *Given is some event of interest, E.*

(1) *We define a random variable Y which counts how often E happen in a "unit". We leave it open whether this unit is a time interval (maybe a minute or a year) or a subset of d–dimensional space ($d = 1, 2, 3$). Let us write A for that unit.*

- *Example: Y is the number of car accidents that happen in Binghamton during a day (unit of time),*
- *Example: Y is the number of typos on a randomly picked page of these lecture notes ("page" is a twodimensional unit – square inches).*

(2) *Given $n \in \mathbb{N}$, we subdivide the unit (A) into n parts of equal size. Let*

$$\vec{X}^{(n)} := X_1^{(n)}, X_2^{(n)}, X_n^{(n)}\,,$$

where $X_j^{(n)}$ = the number of times that E happens in subunit j.

- *Assume that for all big enough, FIXED n,*
  - ⊡ *the $X_j^{(n)}$ are independent*
  - ⊡ *for each j, $P\{X_j^{(n)} = 0 \text{ or } 1\} = 1$: E (i.e., the event of interest) happens at most once in such a small subunit*
  - ⊡ *$p_n := P\{X_j^{(n)} = 1\}$ is constant in j ($j = 1, 2, \ldots, n$)*
  - ⊡ *$\lambda := n \cdot p_n$ is constant in n: For large enough k, $kp_k = (k+1)p_{k+1} = (k+2)p_{k+2} = \cdots = \lambda$.*

Given these assumptions, the following is true:

(a) *The random variable $Y^{(n)} := X_1^{(n)} + X_2^{(n)} + \cdots, +X_n^{(n)}$ is binom(n, p_n) for large n.*

(b) *The binom(n, p_n) probability mass functions $p_{Y^{(n)}}$ converge to that of a poisson(λ) variable:*

$$\lim_{n\to\infty} p_{Y^{(n)}}(y) \;=\; \lim_{n\to\infty} \binom{n}{y} p_n^y (1-p_n)^{n-y} \;=\; e^{-\lambda} \cdot \frac{\lambda^y}{y!}\,, \qquad \text{for } y = 0, 1, 2, \ldots, \tag{9.25}$$

PROOF: We follow WMS:

Recall that $\lambda = np_n$. Thus,

$$\begin{aligned}
\binom{n}{y} p_n^y (1-p_n)^{n-y} &= \frac{n(n-1)\cdots(n-y+1)}{y!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\
&= \frac{\lambda^y}{y!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-y+1)}{n^y} \left(1 - \frac{\lambda}{n}\right)^{-y} \qquad (\star) \\
&= \left(\frac{\lambda^y}{y!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-y} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{y-1}{n}\right).
\end{aligned}$$

From calculus we obtain $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$. Further,

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-y} = \lim_{n\to\infty} \left(1 - \frac{1}{n}\right) = \lim_{n\to\infty} \left(1 - \frac{2}{n}\right) = \cdots = \lim_{n\to\infty} \left(1 - \frac{y-1}{n}\right) = 1\,.$$

We take limits in $(\star)$ and obtain

$$\lim_{n\to\infty} \binom{n}{y} p_n^y (1-p_n)^{n-y} \;=\; \left(\frac{\lambda^y}{y!}\right) e^{-\lambda}\,. \quad \blacksquare$$

Theorem 9.17 (WMS Ch.03.8, Theorem 3.11)**.**

A poisson(λ) random variable has expectation and variance λ. In other words,

$$E[Y] \;=\; Var[Y] \;=\; \lambda\,. \tag{9.26}$$

A. PROOF of $E[Y] = \lambda$:

$$E(Y) \;=\; \sum_y y p_Y(y) \;=\; \sum_{y=0}^{\infty} y\,\frac{\lambda^y e^{-\lambda}}{y!}\,. \;=\; \sum_{y=1}^{\infty} y\,\frac{\lambda^y e^{-\lambda}}{y!} \;=\; \lambda \sum_{y=1}^{\infty} \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!}\,.$$

In the last equation we used $y!/y = (y-1)!$. We write $k = y-1$ for the index variable and obtain

$$E(Y) \;=\; \lambda \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \;=\; \lambda \sum_{k=0}^{\infty} p(k)\,,$$

where $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ is the PMF of a poisson(λ) random variable. Thus, $\sum\limits_{k=0}^{\infty} p(k) = 1$ and it follows that $E[Y] = \lambda$.

B. PROOF of $Var[Y] = \lambda$:

$$\text{Observe that} \qquad y^2 e^{-\lambda}\frac{\lambda^y}{y!} \;= e^{-\lambda}\cdot \frac{y^2 \lambda \lambda^{y-1}}{y!} \;=\; (\lambda e^{-\lambda})\frac{y\lambda^{y-1}}{(y-1)!} \;=\; (\lambda e^{-\lambda})\frac{1}{(y-1)!}\frac{d}{d\lambda}(\lambda^y)$$

We interchange summation and differentiation and obtain

$$\begin{aligned} E[Y^2] \;&=\; \sum_{y=0}^{\infty} y^2 e^{-\lambda}\frac{\lambda^y}{y!} \;=\; \sum_{y=1}^{\infty} y^2 e^{-\lambda}\frac{\lambda^y}{y!} \;=\; (\lambda e^{-\lambda}) \sum_{y=1}^{\infty} \frac{d}{d\lambda}\left(\frac{\lambda\cdot\lambda^{y-1}}{(y-1)!}\right) \\ &= (\lambda e^{-\lambda})\frac{d}{d\lambda}\left(\lambda \sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!}\right) \;=\; (\lambda e^{-\lambda})\frac{d}{d\lambda}\left(\lambda \sum_{y=0}^{\infty}\frac{\lambda^y}{y!}\right)\,. \end{aligned}$$

Since $\sum\limits_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{\lambda}$, this implies $E[Y^2] \;=\; (\lambda e^{-\lambda})\frac{d}{d\lambda}\left(\lambda e^{\lambda}\right) \;=\; \lambda e^{-\lambda}(e^{\lambda} + \lambda e^{\lambda}) \;=\; \lambda + \lambda^2$.

We use $E[Y^2] \;=\; \lambda + \lambda^2$ together with $E[Y] = \lambda$, which we proved in part **A**. We obtain

$$Var[Y] \;= E[Y^2] \;-\; (E[Y])^2 \;=\; (\lambda \;+\; \lambda^2) \;-\; \lambda^2 \;=\; \lambda\,. \;\blacksquare$$

We refer to the WMS text for examples of random variables with a Poisson distribution.

9.5 Moments, Central Moments and Moment Generating Functions

Unless something different is stated, Y is a random variable $Y : (\Omega, P) \to \mathbb{R}$ on some probability space (Ω, P).

$$\mu = E[Y], \quad \sigma^2 = Var[Y], \quad \sigma = \sqrt{Var[Y]}\,,$$

denote expectation, variance and standard deviation of Y.

Definition 9.10 (kth Moment)**.**

If Y is a random variable and $k \in \mathbb{N}$,

(9.27) $$\mu'_k \ := \ E[Y^k]$$

is called the kth **moment** of Y. μ'_k also is referred to as the kth **moment of** Y **about the origin**. □

Note in particular that the first moment of Y is the expectation of Y and that

$$\mu'_2 = Var[Y] + E[Y]^2.$$

Another useful moment of a random variable is one taken about its mean.

Definition 9.11 (kth Central Moment)**.**

If Y is a random variable and $k \in \mathbb{N}$,

(9.28) $$\mu_k \ := \ E[(Y - E[Y])^k] \ = \ E[(Y - \mu)^k]$$

is called the kth **central moment** of Y aka the kth **moment of** Y **about its mean**. □

Proposition 9.5 (The moments determine the distribution)**.** ⋆

Under fairly slight assumptions the following is true for two random variables Y_1 and Y_2.

$$\textbf{\textit{If}} \ E[Y_1^k] \ = \ E[Y_2^k] \ \textit{ for } k = 1, 2, 3, \ldots, \quad \textit{then } \ P_{Y_1} \ = \ P_{Y_2} \, .$$

In other words, the distribution of a random variable is uniquely determined by its moments.

PROOF: Beyond the scope of these lecture notes. ■

Next we associate with a random variable Y which is a function $\omega \mapsto Y(\omega)$ a function $t \mapsto m_Y(t)$ of a real variable t. It allows us to generate all moments μ'_k of Y by computing its kth derivative at $t = 0$. Since $m_Y(t)$ determines in this way all moments of Y and since those in turn determine P_Y, [93] $m_Y(t)$ uniquely determines the entire distribution of Y.

Definition 9.12 (Moment–generating function)**.**

Let Y be a random variable for which one can find $\delta > 0$ (no matter how small), such that

(9.29) $$m(t) \ := m_Y(t) \ := \ E\left[e^{tY}\right] \qquad \text{is finite for } \ |t| < \delta \, .$$

Then we say that Y has **moment–generating function**, in short, **MGF**, $m_Y(t)$. □

[93] See Proposition 9.5

Theorem 9.18. *The following is WMS Ch.03.9, Theorem 3.12.*

Let Y be a random variable with MGF $m_Y(t)$ and $k \in \mathbb{N}$. Then its kth moment is obtained as the kth derivative of $m_Y(\cdot)$, evaluated at $t = 0$:

(9.30)
$$\mu'_k = m^{(k)}(0) = \frac{d^k m(t)}{dt^k}\Big|_{t=0}.$$

PROOF: We write $m(t)$ for $m_Y(t)$. From the series expansion $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, we obtain

$$\begin{aligned} m(t) &= E\left[e^{tY}\right] = E\left[\sum_{k=0}^{\infty} \frac{t^k Y^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} E\left[Y^k\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mu'_k \\ &= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots. \end{aligned}$$

Taking derivatives repeatedly,

$$m^{(1)}(t) = \mu'_1 + \frac{2t}{2!}\mu'_2 + \frac{3t^2}{3!}\mu'_3 + \cdots \qquad \Rightarrow\ m^{(1)}(0) = \mu'_1 + 0 + 0 + \cdots,$$
$$m^{(2)}(t) = \mu'_2 + \frac{2t}{2!}\mu'_3 + \frac{3t^2}{3!}\mu'_4 + \cdots \qquad \Rightarrow\ m^{(2)}(0) = \mu'_2 + 0 + 0 + \cdots,$$

....

$$m^{(k)}(t) = \mu'_k + \frac{2t}{2!}\mu'_{k+1} + \frac{3t^2}{3!}\mu'_{k+2} + \cdots \qquad \Rightarrow\ m^{(k)}(0) = \mu'_k + 0 + 0 + \cdots$$

In summary,
$$m^{(1)}(0) = \mu'_1, \quad m^{(2)}(0) = \mu'_2, \quad \ldots, \quad m^{(k)}(0) = \mu'_k. \ \blacksquare$$

Technical note: The existence of the MGF of Y allowed us to compute the derivative of a series as the sum of the derivatives.

You find the next proposition as Example 3.23 in WMS Ch.3.9.

Proposition 9.6. ⋆ *If Y is a poisson(λ) random variable ($\lambda > 0$), its MGF is*

(9.31)
$$m_Y(t) = e^{\lambda(e^t - 1)}.$$

PROOF: For this proof, we abbreviate **(A)** $\tilde{\lambda} := \lambda e^t$.

Note that the Taylor expansion $e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$ yields, with x and j replaced by $\tilde{\lambda}$ and y,

(B) $e^{\tilde{\lambda}} = \sum_{y=0}^{\infty} \frac{\tilde{\lambda}^y}{y!}$.

$$\text{Then,} \quad m_Y(t) = E\big(e^{tY}\big) = \sum_{y=0}^{\infty} e^{ty} p(y) = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{(-\lambda)}}{y!}$$
$$= \sum_{y=0}^{\infty} (e^t)^y \lambda^y \frac{e^{-\lambda}}{y!} = \sum_{y=0}^{\infty} \frac{\big(\lambda e^t\big)^y e^{-\lambda}}{y!} \overset{\text{(A)}}{=} e^{-\lambda} \sum_{y=0}^{\infty} \frac{\tilde{\lambda}^y}{y!}$$
$$\overset{\text{(B)}}{=} e^{-\lambda} e^{\tilde{\lambda}} \overset{\text{(A)}}{=} e^{(-1)\lambda} e^{\lambda e^t} = e^{\lambda(-1+e^t)} = e^{\lambda(e^t-1)}. \quad \blacksquare$$

- The subsection titled "The Tchebysheff Inequality" which was at this location has been integrated into subsection 10.8 (Inequalities for Probabililities)

9.6 Exercises for Ch.9

Exercise 9.1. If the random variable Y has expectation $E[Y] = -2$ and standard deviation $\sigma_Y = 2$, what is $E\big[(Y+3)^2\big]$?

Answer: Since $E[Y^2] = Var[Y] + (E[Y])^2 = (\sigma_Y)^2 + (-2)^2 = 8$,

$$E\big[(Y+3)^2\big] = E[Y^2] + 6E[Y] + 9 = 8 - 12 + 9 = \boxed{5} \quad \blacksquare$$

Exercise 9.2. If the random variable Y has the PMF

$$p_Y(-2) = 0.13,\ p_Y(0) = 0.24,\ p_Y(1) = 0.18, p_Y(2) = 0.45\,,$$

(a) compute $E[Y]$
(b) compute $Var[Y]$
(c) compute σ_Y

Answer (the numeric computations might have errors):

(a) $E[Y] = \sum_y y \cdot p_Y(y) = (-2)(0.13) + 0(0.24) + 1(0.18) + 2(0.45) = 0.82$
(b) $Var[Y] = \sum_y (y - E[Y])^2 \cdot p_Y(y)$
$= (-2-0.82)^2(0.13) + (0-0.82)^2(0.24) + (1-0.82)^2(0.18) + (2-0.82)^2(0.45) = 1.8276$
(c) $\sigma_Y = \sqrt{Var[Y]} = \sqrt{1.8276} \approx 1.3513888$

Exercise 9.3. Let Y be a 0–1 encoded Bernoulli variable with $P\{Y=1\} = p$.

(a) Compute its MGF
(b) Use the MGF method to compute the nth moment about the origin, $E[X^n]$

Answer:

(a) $$M_Y(t) \,=E\big[e^{tY}\big] \,= \qquad e^{0t}\cdot q \,+\, e^{1t}\cdot p \,=\, \boxed{q+pe^t}$$

(b) The derivatives of $M_Y(t)$ are

$$M_Y'(t) \,=(q+pe^t)' \,=\, pe^t\,, M_Y'(t) \,=\, (pe^t)' \,=\, pe^t,\,\ldots,\, M_Y^{(n)}(t) \,=\, pe^t,\,\ldots,$$

Thus, $E[Y^n] \,=\, \mu_n' \,=\, M_Y^{(n)}(0) \,=\, pe^0 \,=\, \boxed{p}$ for all n.
(c) We use the results of **(b)** to compute the variance:

$$Var[Y] \,=\, E[Y^2] \,-\, (E[Y])^2 \,=\, \mu_2' - (\mu_1')2 \,=\, p \,-\, p^2 \,=\, (1-p)p \,=\, \boxed{pq} \;\blacksquare$$

Exercise 9.4. Let Y be a binom(n,p) variable. Use the MGF method to verify that $E[Y] = np$ and $Var[Y] = npq$.

Answer: Since the PMF of Y is $p_Y(y) = \binom{n}{y}p^y q^{n-y}$,

$$\begin{aligned} M_Y(t) \,&=E\big[e^{tY}\big] \,=\, \sum_{y=0}^{n} e^{ty}\binom{n}{y}p^y q^{n-y} \,=\, \sum_{y=0}^{n}\binom{n}{y}\big(e^t\big)^y p^y q^{n-y} \\ &= \sum_{y=0}^{n}\binom{n}{y}\big(pe^t\big)^y q^{n-y} \,=\, \big(pe^t+q\big)^n \end{aligned}$$

Here we obtained the last equation by applying the binomial theorem,

$$(a+b)^n \,=\, \sum_{j=0}^{n}\binom{n}{j}a^j q^{n-j}\,,$$

with $a = pe^t$ and $b = q$.

$$\begin{aligned} M_Y(t)' \,&=npe^t(pe^t+q)^{n-1}\,, \\ M_Y(t)'' \,&=npe^t(pe^t+q)^{n-1} \,+\, n(n-1)(pe^t)^2(pe^t+q)^{n-2}\,. \end{aligned}$$

Thus,

$$\begin{aligned} E[Y] \,=\, M_Y(0)' \,&=np\,, \\ E[Y^2] \,&= M_Y(0)'' \,=\, np+n(n-1)p^2\,. \end{aligned}$$

It follows that

$$\begin{aligned} Var[Y] \,&= E[Y^2] \,-\, (E[Y])^2 \,= \\ E[Y^2] \,&= M_Y(0)'' \,=\, np+n(n-1)p^2 \,-\, n^2p^2 \,=\, npq\,. \;\blacksquare \end{aligned}$$

10 Continuous Random Variables

10.1 Cumulative Distribution Function of a Random Variable

The material found in this section does not make any references to continuous random variables.

Definition 10.1 (Cumulative Distribution Function)**.**

Let Y denote any random variable (it need not be discrete). The **distribution function** of Y, also called its **cumulative distribution function** or **CDF (cumulative distribution function)**, is defined as follows.

(10.1) $$F(y) := F_Y(y) := P\{Y \le y\} \qquad \text{for } y \in \mathbb{R}\,. \ \square$$

Problem 10.1. Let Y be a binom$(2, 1/4)$ random variable, i.e., $n = 2$ and $p = 1/4$. Compute $F_Y(y)$.
Solution: The probability mass function for Y is

$$p_Y(y) = \binom{2}{y}\left(\frac{1}{4}\right)^y\left(\frac{3}{4}\right)^{2-y}.$$

Thus,

$$p_Y(0) = \frac{9}{16}\,, \qquad p_Y(1) = 2\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{6}{16}\,. \qquad p_Y(2) = \frac{1}{16}\,.$$

It follows that

- $y < 0 \Rightarrow F_Y(y) = P_Y(\emptyset) = 0$.
- $0 \le y < 1 \Rightarrow F_Y(y) = p_Y(0) = 9/16$.
- $1 \le y < 2 \Rightarrow F_Y(y) = p_Y(0) + p_Y(1) = 15/16$.
- $y \ge 2 \Rightarrow F_Y(y) = p_Y(0) + p_Y(1) + p_Y(2) = 1$.

Note that F_Y is constant on intervals A of $\mathbb{R}$ if $p_Y(a) = 0$ for all $a \in A$. $\square$

Theorem 10.1 (Properties of a Cumulative Distribution Function)**.**

If $F_Y(y)$ is the cumulative distribution function of a random variable Y, then

(1) $F_Y(-\infty) = \lim_{y\to-\infty} P(Y \le y) = 0$.

(2) $F_Y(\infty) = \lim_{y\to\infty} P(Y \le y) = 1$.

(3) *$F_Y(y)$ s a nondecreasing function of y. In other words, if $y_1 < y_2$, then $F_Y(y_1) \le F_Y(y_2)$*
See Definition 2.23 on p.36.

(4) *$y \mapsto F_Y(y)$ is **right continuous** at all arguments y, i.e., $F(y) = F(y+)$ for all y.*

PROOF:
The proof of **(1)** and **(2)** follows from

It follows from $-\infty < Y(\omega) < \infty$ that

$$\bigcap_{y\in\mathbb{R}} \{Y \le y\} \;=\; \bigcap_{n\in\mathbb{N}} \{Y \le -n\} \;=\; \emptyset$$
$$\bigcup_{y\in\mathbb{R}} \{Y \le y\} \;=\; \bigcup_{n\in\mathbb{N}} \{Y \le n\} \;=\; \Omega$$

We apply Theorem 5.1 (Continuity property of probability measures) on p.109 and obtain

$$F_Y(-\infty) \;=\; \lim_{n\to\infty} P\left(\bigcap_{y\in\mathbb{R}} \{Y \le y\}\right) \;=\; P(\emptyset) \;=\; 0\,,$$
$$F_Y(\infty) \;=\; \lim_{n\to\infty} P\left(\bigcup_{y\in\mathbb{R}} \{Y \le y\}\right) \;=\; P(\Omega) \;=\; 1\,.$$

Obvious from $P \ge 0$ and $y_1 < y_2 \;\Rightarrow\; \{Y \le y_2\} = \{Y \le y_2\} \uplus \{y_1 < Y \le y_2\}$, since this implies

$$F(y_2) \;=\; P\{Y \le y_2\} \;=\; P\{Y \le y_1\} + P\{y_1 < Y \le y_2\} \;\ge\; P\{Y \le y_1\} \;=\; F(y_1)\,. \;\blacksquare$$

Remark 10.1. Right continuity of F, i.e., $F(y) = F(y+)$ for all y, means the following: If y is approached from the right by a sequence y_n such as $y_n = y + \frac{1}{n}$ or $y_n = y(1 + e^{-n})$, then

$$\lim_{n\to\infty} F(y_n) \;=\; F(y)\,. \;\square$$

10.2 Continuous Random Variables and Probability Density Functions

Definition 10.2 (Continuous random variable)**.**

We call a random variable Y with distribution function $F_Y(y)$ **continuous**, if $F_Y(y)$ is continuous, for all arguments y. $\square$

Proposition 10.1. *Let Y be a continuous random variable with CDF $F_Y(y)$. Then its distribution gives zero probability to all singletons $\{a\}(a \in \mathbb{R})$. Also, it gives the same probability to an interval with endpoints $-\infty < a < b < \infty$, regardless whether a and/or b do or do not belong to that interval. In other words,*

(10.2) $$a \in \mathbb{R} \;\Rightarrow P\{Y = a\} \;=\; P_Y\{a\} \;=\; 0\,,$$

(10.3) $$-\infty < a < b < \infty \;\Rightarrow P\{a < Y < b\} \;=\; P\{a \le Y < b\} = P\{a < Y \le b\} \;=\; P\{a \le Y \le b\}\,.$$

PROOF: Since $\{a\} \subseteq]a - \frac{1}{n}, a]$ and $]a - \frac{1}{n}, a] \;=\;]-\infty, a] \setminus]-\infty, a - \frac{1}{n}]$ (set difference),

$$P\{Y = a\} \;\le\; P\{a - \frac{1}{n} < Y \;\le a\} \;=\; P\{Y \le a\} \;-\; P\{Y \le a - \frac{1}{n}\} \;=\; F_Y(a) \;-\; F_Y\left(a - \frac{1}{n}\right)\,.$$

F_Y is continuous at a. In particular, F_Y is continuous from the left at a. Thus,

$$\lim_{n\to\infty} F_Y\left(a - \frac{1}{n}\right) \;=\; F_Y(a)\,.$$

It follows that $P\{Y = a\} = F_Y(a) - F_Y(a) = 0$. This proves (10.2).
This result, plus additivity of probability measures, plus

$$[a,b] \;=]a,b[\uplus \{a\} \uplus \{b\}\,, \quad [a,b] \;= [a,b[\uplus \{b\}\,, \quad [a,b] \;=]a,b] \uplus \{a\}\,,$$

show that (10.3) holds. ■

A lot more can be done with a CDF that is not only continuous but has a continous derivative. We make the following blanket assumption.

Assumption 10.1 (All continuous random variables have a differentiable CDF)**.** Unless explicitly stated otherwise, all continuous random variables are assumed to satisfy the following:

> The first derivative $\frac{dF_Y}{dy}$ of F_Y exists and is continuous except for, at most, a finite number of points in any finite interval.

All cumulative distribution functions for continuous random variables that we deal with in this course satisfy this assumtion. □

This last assumption allows us to make the following definition.

Definition 10.3 (Probability density function)**.**

> Let Y be a continuous random variable with CDF $F_Y(y)$. For all arguments y where the derivative $F'_Y(y) = \frac{dF_Y(y)}{dy}$ exists, we define
>
> $$f(y) \;:=\; f_Y(y) \;:=\; \frac{dF_Y(y)}{dy}\,.$$
>
> We call f_Y the **probability density function** or, in short, the **PDF** of the continuous random variable Y. □

Theorem 10.2.

> *Let Y be a continuous random variable with CDF $F_Y(y)$ and PDF $f_Y(y)$.*
>
> ***(1)*** *If $a, b \in \mathbb{R}$ and $a < b$, then*
>
> (10.4) $$P\{a < Y \le b\} \;=\; F_Y(b) - F_Y(a) \;=\; \int_a^b f(y)dy\,.$$
>
> ***(2)*** $f_Y(y) \;\ge\; 0$ *for* $-\infty < y < \infty$.
>
> ***(3)*** $\int_{-\infty}^{\infty} f_Y(y)dy \;=\; 1$.

PROOF: **(1)** is the fundamental theorem of calculus. Of course, we interpret $\int_a^b f(y)dy$ as follows. Assume that some of the points y at which $f'_Y(y)$ does not exist fall within the interval $[a,b]$. Our assumption guarantee that there are only finitely many such y, say,

$$a \leq y_1 < y_2 < \cdots y_k \leq b\,.$$

Then, by the definition of integrals,

$$\int_a^b f(y)dy = \int_a^{y_1} f(y)dy + \int_{y_1}^{y_2} f(y)dy + \cdots + \int_{y_k}^b f(y)dy\,.$$

(2) and **(3)** are obvious. ■

The following is the reverse of Theorem 10.2.

Theorem 10.3. *Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy the following:*

(1) *ψ is integrable: $\int_a^b \psi(x)dx$ exists for $a < b$.*

(2) *$\psi(x) \geq 0$ for $-\infty < x < \infty$.*

(3) *$\int_{-\infty}^{\infty} \psi(x)dx = 1$.*

- *Then, $Q\{a < Y \leq b\} := \int_a^b \psi(x)dx$ defines a probability measure Q on Ω.*

PROOF: ⋆

The only property that is not immediate is the σ–additivity of Q. That property is satisfied according to Theorem 3.5 on p.77. (Also, from Corollary 4.2 on p.98). ■

Remark 10.2. We combine (10.3) and (10.4) and obtain the following for a continuous random variable Y with PDF $f_Y(y)$: If $a, b \in \mathbb{R}$ and $a < b$, then

$$\begin{aligned} P\{a < Y < b\} &= P\{a \leq Y \leq b\} = P\{a \leq Y < b\} \\ &= P\{a < Y \leq b\} = \int_a^b f(y)dy\,. \ \square \end{aligned} \tag{10.5}$$

The next definition applies to any random variable, be it continuous or discrete or neither. It is based on the following elementary observation.

Remark 10.3. ⋆ Assume that Y is a random variable with CDF $F_Y(y)$. For $0 < p < 1$, let

$$A_p := \{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\}\,.$$

Note that the function $y \mapsto F_Y(y)$ is nondecreasing.

- It is obvious that $\big[\,\alpha < \alpha' \text{ and } F_Y(\alpha) \geq p\,\big] \Rightarrow F_Y(\alpha') \geq p$.
- In other words, $\big[\,\alpha < \alpha' \text{ and } \alpha \in A_p\,\big] \Rightarrow \alpha' \in A_p$.
- In other words, A_p is an interval that stretches all the way to $+\infty$: There must be some real number β such that $A_p =]\beta, \infty[$ or $A_p = [\beta, \infty[$. [94]

[94] and that number is $\beta = \inf(A_p)$ See Definition 2.29 (Minimum, maximum, infimum, supremum) on p.48

We see that $\beta \in A_p$ and thus, $A_p = [\beta, \infty[$, as follows. Let $\beta_n := \beta + \frac{1}{n}$.

- Since $\beta_n \in A_p$, $F_Y(\beta_n) \geq p$. Since F_Y is right continuous, [95] $F_Y(\beta) = \lim_{n\to\infty} F_Y(\beta_n)$.
- Thus, $F_Y(\beta) \geq p$. Thus, $\beta \in A_p$ Thus, $A_p = [\beta, \infty[$.
- Since $A_p = \{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\}$ and $A_p = [\beta, \infty[$, β is the smallest element of A_p, i.e.,

$$\beta \;=\; \min\{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\}\,.$$

The number β is uniquely determined by p. This allows us to denote it by the symbol ϕ_p. □

Definition 10.4 (pth quantile)**.**

Let Y denote any random variable and $0 < p < 1$. Let ϕ_p be the number derived in the previous remark, i.e.,

(10.6) $$\phi_p \;=\; \min\{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\}$$

We call ϕ_p the pth **quantile** and also the $100p$th **percentile** of Y.
Moreover, we call $\phi_{0.25}$ the **first quartile**, $\phi_{0.5}$ the **median**, and $\phi_{0.75}$ the **third quartile**, of the random variable Y. □

Remark 10.4. How does the definition of the $100p$th percentile given above correspond to the one experienced in everyday life: the number y_p that divides a list of numeric observations into $100p\%$ of the data being $\leq y_p$ and the remaining data being above y_p? The connection is as follows.

- Assume that $\vec{y} = (y_1, y_2, \ldots, y_K)$ is the list of observations. It may contain duplicates.
- We remove the duplicates and $N \leq K$ distinct values $\omega_1, \omega_2, \ldots, \omega_N$ remain.
- We define $\Omega := \{\omega_1, \omega_2, \ldots, \omega_N\}$ and $P\{\omega_j\} := \dfrac{n_j}{K}$ (we divide by K, **not** by N!), where
  n_j = number of times that ω_j occurs in the original list, $\vec{y}$.
- σ–additivity extends P from the simple events $\{\omega_j\}$ to all events of Ω.
- Since ϕ_p is defined in terms of the CDF F_Y of a random variable Y, we define the following "dummy" random variable on (Ω, P): $\omega \mapsto Y(\omega) := \omega$ [96]

For example, if the sorted' list of observations is $\vec{y} = (0, 2, 2, 2, 3, 4, 4, 6, 6, 6, 6, 7, 8, 8, 8)$, then

- $K = 15$, $\Omega = \{0, 2, 3, 4, 6, 7, 8\}$, $N = 7$,
- $P\{0\} = \frac{1}{15}$, $P\{2\} = \frac{3}{15}$, $P\{3\} = \frac{1}{15}$, $P\{4\} = \frac{2}{15}$, $P\{6\} = \frac{4}{15}$, $P\{7\} = \frac{1}{15}$, $P\{8\} = \frac{3}{15}$.
- Thus, $F_Y(3) = (1 + 3 + 1)/15 = 5/15$, and $F_Y(4) = (1 + 3 + 1 + 2)/15 = 7/15$
  Thus, $\phi_{7/15} \;=\; \min\{y : \phi(y) \geq 7/15\} \;= 4$.
- Also, the percentage of observations with a score of 4 or less is $700/15 \approx 46.667\%$.
  Hence, a score of 4 corresponds to the 46.667th percentile of $\vec{y}$. □

[95] See Remark 10.1 on p.213.

[96] This method is more frequently employed in reverse: Given is a function $y \mapsto F(y)$ on the real numbers which satisfies the assumptions of Theorem 10.1 (Properties of a Cumulative Distribution Function) on p.212 and the subsequent Remark 10.1: F is nondecreasing, right–continuous, $F(-\infty) = 0$, $F(\infty) = 1$. We then define $\Omega := \mathbb{R}$ and, for $]a, b] \subseteq \Omega$, $P(]a, b]) := F(b) - F(a)$. σ–additivity extends this to a probability measure on all Borel sets of Ω (i.e., of $\mathbb{R}$). Now we define the random variable Y on (Ω, P) via $Y(y) := y$. Its CDF F_Y matches F, since,

$$F_Y(y) \;=\; P\{Y \leq y\} \;=\; P(]-\infty, y]) \;=\; F(y) \,-\, F(-\infty) \;=\; F(y)\,.$$

In other words, Any function F that conforms to Theorem 10.1 and Remark 10.1 can be represented as the CDF F_Y of an appropriate random variable Y.

Example 10.1. Given the toss of a fair coin, let $Y(\omega) = 1$ if Heads and $Y(\omega) = 0$ if Tails come up. Then Y has PMF $p_Y(0) = p_Y(1) = 1/2$ and its CDF is as follows:

$\bullet F_Y(y) = 0$ for $y < 0$, $\bullet F_Y(y) = 0.5$ for $0 \le y < 1$, $\bullet F_Y(y) = 1$ for $y \ge 1$.

We now easily compute ϕ_p for any $0 < p < 1$ by separately considering the cases

$0 < p < \frac{1}{2}$: $F_Y(\alpha) \ge p \Leftrightarrow \alpha \ge 0$. Thus, $\phi_p = 0$.
$p = \frac{1}{2}$: $F_Y(\alpha) \ge \frac{1}{2} \Leftrightarrow \alpha \ge 0$. Thus, $\phi_{1/2} = 0$.
$\frac{1}{2} < p < 1$: $F_Y(\alpha) \ge p \Leftrightarrow \alpha \ge 1$. Thus, $\phi_p = 1$.

Note that there are only two different ϕ_p values across all $0 < p < 1$: Either $\phi_p = 0$ or $\phi_p = 1$
This example also demonstrates that

$$\min\{\alpha \in \mathbb{R} : F_Y(\alpha) \ge p\}$$

cannot be replaced with the simpler expression

$$\min\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\} \ :$$

The set $\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}$ is empty for $0 < p < 1$ unless $p = 0.5$, meaning that the minimum does not even exist! □

The issues encountered in that last example do not occur if $F_Y(y)$ is a continuous function of y.

Proposition 10.2.

Let Y be a continuous random variable with CDF $F_Y(y)$. Then

$$\phi_p \ = \ \min\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}\,. \tag{10.7}$$

PROOF: The continuity of F_Y ensures that the sets

$$B_p \ := \ \{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}$$

are not empty. The result follows from the fact that the function F_Y is nondecreasing. Further details are omitted. ■

Remark 10.5. For a continuous random variable Y with PMF $p_Y(y)$, quantiles have the following geometric meaning:

- The pth quantile is that value on the horizontal(!) axis which splits the area under the PMF into $100 \cdot p\%$ to the left and $100(1-p)\%$ to the right. In particular,
- the median splits the area under the PMF into two halves.
- the first quartile splits the area under the PMF into 25% to the left and 75% to the right.
- the third quartile splits the area under the PMF into 75% to the left and 25% to the right. □

We also use functional notation $\phi(p)$ for ϕ_p, since this makes what follows easier to understand.

Proposition 10.3. [97]

[97] Formula (10.8) of this proposition states that ϕ is a left inverse of the injective function F_Y.

Let Y be a random variable with an injective CDF $F_Y(y)$. (Note that it is not assumed that F_Y is continuous.) Then

$$\phi\big(F_Y(y)\big) \;=\; y \qquad \textit{for all } y \in \mathbb{R} \tag{10.8}$$

PROOF:

Let $p := F_Y(y)$. Since F_Y is nondecreasing, its injectivity means that

$$y_1 \;<\; y \;<\; y_2 \;\Rightarrow\; F_Y(y_1) \;<\; F_Y(y) \;<\; F_Y(y_2) \tag{10.9}$$

We infer that $\alpha < y$ does not satisfy $F_Y(\alpha) \geq F_Y(y) = p$. Since (see 10.6 on p.216)

$$\phi\big(F_Y(y)\big) \;=\; \min\{\alpha \in \mathbb{R} : F_Y(\alpha) \geq \phi\big(F_Y(y)\big)\}\,, \tag{10.10}$$

it follows from (10.10) that $\phi\big(F_Y(y)\big) < y$ is not possible. Thus, $\phi\big(F_Y(y)\big) \geq y$.

On the other hand, $\alpha = y$ does satisfy $F_Y(\alpha) \geq F_Y(y) = p$ and we just have seen that y is the smallest possible of those α. We apply (10.10) once more and conclude that $\phi\big(F_Y(y)\big) = y$. ■

Proposition 10.4.

Let Y be a random variable with a bijective CDF $F_Y : \mathbb{R} \xrightarrow{\sim}]0,1[$. Then $F_Y(y)$ and $\phi(p)$ are inverse to each other, i.e.,

$$\phi\big(F_Y(y)\big) \;=\; y \;\textit{ for all } y \in \mathbb{R} \quad \textit{and} \quad F_Y\big(\phi(p)\big) \;=\; p \;\textit{ for all } 0 < p < 1\,. \tag{10.11}$$

PROOF:

The equation $\phi\big(F_Y(y)\big) \;=\; y$ was shown in Proposition 10.3. Thus, it only remains to be shown that

$$F_Y\big(\phi(p)\big) \;=\; p \;\text{ for all } 0 < p < 1\,. \tag{10.12}$$

We observe that the bijective and nondecreasing function F_Y is strictly increasing and continuous.

It is easy to see that F_Y is strictly increasing: Note that $y_1 < y_2 \;\Rightarrow F_Y(y_1) \leq F_Y(y_2)$ because F_Y is nondecreasing. Injectivity prohibits $F_Y(y_1) = F_Y(y_2)$. Thus, F_Y is strictly increasing.

It is harder to see that F_Y is continuous:

- If there was a point of discontinuity $y_0 \in \mathbb{R}$ for F_Y, then F_Y being nondecreasing and right–continuous would mean that $F_Y(y_0-) = \lim\limits_{y<y_0, y\to y_0} F_Y(y) \;<\; F_Y(y_0)$.
- Also, F_Y nondecreasing $\Rightarrow\; F_Y(y) \leq F_Y(y_0-)$ for $y < y_0$ and $F_Y(y) \geq F_Y(y_0)$ for $y \geq y_0$.
- Thus, no $y \in \mathbb{R}$ and $p \in]F_Y(y_0-), F_Y(y_0)[$ satisfies $F_Y(y) = p$, contradicting surjectivity of F_Y.

Since F_Y is continuous, we obtain from Proposition 10.2 on p.217 that

$$\phi(p) \;=\; \min\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}\,. \tag{10.13}$$

In particular, $\phi(p)$ is an element of the set $\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}$. Thus, $\phi(p)$ satisfies $F_Y\big(\phi(p)\big) = p$. We have shown (10.12). We noted previously that the proposition follows. ■

10.3 Expected Value, Variance and MGF of a Continuous Random Variable

Assumption 10.2 (All continuous random variables have Expectations)**.**

A. Unless explicitly stated otherwise, all continuous random variables are assumed to to possess a probability density function $f_Y(y)$ that satisfies

$$\int_{-\infty}^{\infty} |y| f(y)\, dy| < \infty\,.$$

This technical condition guarantees the existence of $\int\limits_{-\infty}^{\infty} yf(y)dy$ which is needed to define the expected value of Y.

B. We further assume that, unless specifically stated otherwise, there is a common probability space (Ω, P) for all random variables. In other words, all random variables Y, be they discrete, continuous or neither, are of the form $Y : (\Omega, P) \to \mathbb{R}$. □

Definition 10.5 (Expected value of a continuous random variable)**.**

Let Y be a continuous random variable with PDF $f_Y(y)$. We call

$$E(Y) := \int_{-\infty}^{\infty} y f_Y(y)\, dy \tag{10.14}$$

the **expected value**, also **expectation** or **mean** of Y. □

Remark 10.6. ⋆ We recall that expectations $E[Y]$ are abstract integrals $\int Y dP$ (see Definition 6.15 (Expected value of a random variable) on p.165. The connection with (10.14) is established in formulas

$$\int g\, dP_Y = \int g \cdot f\, d\lambda^1 = \int_{-\infty}^{\infty} g(y) f_Y(y)\, dy\,. \tag{6.58}$$

and

$$E[g \circ Y] = \int g \circ Y dP = \int g\, dP_Y = \int_{-\infty}^{\infty} g(y) f_Y(y)\, dy\,. \tag{6.59}$$

of Remark 6.19 on p.167, when setting $g(y) = y$.

As we reviously noticed for the expectations of discrete random variables, all assertions made in Theorem 6.9 on p.159 for general abstract integrals apply to expectations of any kind of random variables Y, since they all can be written as $E[Y] = \int Y dP$. □

We will use the next theorem in the proof of Theorem 10.5 on p.221. The presentation given here follows [4] Ghahramani, Saeed.

Theorem 10.4. ⋆

Let Y be a continuous random variable with CDF F_Y and PDF f_Y.
Then

$$E[Y] = \int_0^\infty \big(1 - F_Y(y)\big)dy - \int_0^\infty F_Y(-y)dy \tag{10.15}$$

$$= \int_0^\infty P\{Y > y\}dy - \int_0^\infty P\{Y \le -y\}dy\,. \tag{10.16}$$

PROOF: We only need to prove (10.15), since (10.16) follows from the definition of a CDF.

Let $A_1 := \{(u', y') : y' < 0,\, 0 < u' < -y'\}\,, \quad B_1 := \{(u', y') : u' > 0,\, y' < -u'\}\,.$

Then $u' < -y' \Leftrightarrow y' < -u'$ implies $A_1 = B_1 = \{(u', y') : u' > 0,\, y' < 0, u' < -y'\}$. Thus,

$$\int_{-\infty}^0 \left(\int_0^{-y} du\right) f(y)\, dy = \iint_{A_1} f_Y(y)\, d(u, y)$$

(a)
$$= \iint_{B_1} f_Y(y)\, d(u, y) = \int_0^\infty \left(\int_{-\infty}^{-u} f_Y(y)\, dy\right) du\,.$$

Let $A_2 := \{(u', y') : y' > 0,\, 0 < u' < y'\}\,, \quad B_2 := \{(u', y') : u' > 0,\, y' > u'\}\,.$

Then $A_2 = B_2$, because both denote the set $\{(u', y') : u' > 0,\, y' > 0, u' < y'\}$. It follows that

$$\int_0^\infty \left(\int_0^{y} du\right) f(y)\, dy = \iint_{A_2} f_Y(y)\, d(u, y)$$

(b)
$$= \iint_{B_2} f_Y(y)\, d(u, y) = \int_0^\infty \left(\int_{u}^{\infty} f_Y(y)\, dy\right) du\,.$$

We use **(a)** and **(b)** in the following chain of equations:

$$\begin{aligned} E[Y] &= \int_{-\infty}^\infty y f_Y(y)\, dy = \int_{-\infty}^0 y f_Y(y)\, dy + \int_0^\infty y f_Y(y)\, dy \\ &= -\int_{-\infty}^0 \left(\int_0^{-y} du\right) f_Y(y)\, dy + \int_0^\infty \left(\int_0^{y} du\right) f_Y(y)\, dy \\ &\overset{\text{(a),(b)}}{=} -\int_0^\infty \left(\int_{-\infty}^{-u} f_Y(y)\, dy\right) du + \int_0^\infty \left(\int_{u}^{\infty} f_Y(y)\, dy\right) du\,. \\ &= -\int_0^\infty F_Y(-u) du + \int_0^\infty \big(1 - F_Y(u)\big)\, du\,. \end{aligned}$$

The last equation follows from $\int_\alpha^\beta f_Y(y)dy = F_Y(\beta) - F_Y(\alpha)$. ■

Corollary 10.1. ⋆

Let Y be a nonnegative, continuous random variable with CDF F_Y and PDF f_Y. Then

$$E[Y] = \int_0^\infty \big(1 - F_Y(y)\big)dy = \int_0^\infty P\{Y > y\}dy\,. \tag{10.17}$$

PROOF: $Y \geq 0$ implies $P\{Y \leq -y\} = 0$ for $0 \leq y < \infty$. Thus, (10.17) follows from (10.15) and (10.16). ■

Quite a few theorems about discrete random variables have continuous counterparts when one replaces probability mass function $p(y)$ with probability density function $f(y)$ and summation over the countably many y for which $p(y) > 0$ with integration over all y. The following theorem corresponds to Theorem 9.2 on p.192. Note that the continuous random variable $\omega \mapsto g\big(Y(\omega)\big)$ of that theorem is covered by Assumption 10.2 on p.219, i.e., $E\big[g \circ Y\big]$ exists.

Theorem 10.5.

Let Y be a continuous random variable with PDF f_Y and $g : \mathbb{R} \to \mathbb{R};\ y \mapsto g(y)$ be a real-valued function. Then the random variable $g \circ Y : \omega \mapsto g\big(Y(\omega)\big)$ has expectation

$$E[g(Y)] = \int_{\infty}^{\infty} g(y) f_Y(y)\, dy\,. \tag{10.18}$$

PROOF: As we mentioned in the remark following Definition 10.5 (Expected value of a continuous random variable) on p.219, (10.18) was derived as formula (6.59) of Remark 6.19 on p.167. ■

ALTERNATE PROOF ⋆ – Doing it the hard way:

The proof of Theorem 9.2 on p.192 handles the discrete case. So we may assume that Y is a continuous random variable.

According to Proposition 2.8 (Preimages of function composition) on p.45,

$$\begin{aligned}\{g \circ Y > u\} &= (g \circ Y)^{-1}(]u, \infty[) = Y^{-1}\big(g^{-1}(]u, \infty[)\big) = \{Y \in g^{-1}(]u, \infty[)\}\,.\\ \{g \circ Y \leq -u\} &= (g \circ Y)^{-1}(]-\infty, -u[) = Y^{-1}\big(g^{-1}(]-\infty, -u[)\big) = \{Y \in g^{-1}(]-\infty, -u[)\}\,.\end{aligned}$$

Thus,

(a)
$$\begin{aligned}P\{g \circ Y > u\} &= P\{Y \in g^{-1}(]u, \infty[)\} = P_Y\{g^{-1}(]u, \infty[)\} = P_Y\{y : g(y) > u\}\\ P\{g \circ Y \leq -u\} &= P\{Y \in g^{-1}(]-\infty, -u[)\}\,. = P_Y\{g^{-1}(]-\infty, -u[)\} = P_Y\{y : g(y) \leq -u\}\,.\end{aligned}$$

Next, we show that $A_1 = B_1$. Here, we define A_1 and B_1 as follows:

(b1) $A_1 := \{(u', y') : 0 < u' < \infty,\ g(y') > u'\}\,, \quad B_1 := \{(u', y') : g(y') > 0,\ 0 < u' < g(y')\}\,,$

To show $A_1 \subseteq B_1$, let $(u, y) \in A_1$, i.e., $(u, y) \in \{(u', y') : 0 < u' < \infty,\ g(y') > u'\}$.

- $0 < u$ and $u < g(y)$ yields $g(y) > 0$ and $0 < u < g(y)$. Thus, $(u, y) \in B_1$.

To see that $B_1 \subseteq A_1$, let $(u, y) \in B_1$, i.e., $(u, y) \in \{(u', y') : g(y') > 0,\ 0 < u' < g(y')\}$.

- Since $0 < u < g(y)$, it follows that $0 < u < \infty$ and $u < g(y)$. Thus, $(u, y) \in A_1$.

(c1) We proved that $A_1 = B_1$. It follows that $\iint_{A_1} f_Y(y)\, d(t, y) = \iint_{B_1} f_Y(y)\, d(t, y)\,.$

On a parallel track, we show that $A_2 = B_2$, where we define A_2 and B_2 as follows:

(b2) $A_2 := \{(u', y') : 0 < u' < \infty,\ g(y') \leq -u'\} \quad B_2 := \{(u', y') : g(y') < 0,\ 0 < u' \leq -g(y')\}\,.$

To show $A_2 \subseteq B_2$, let $(u,y) \in A_2$, i.e., $(u,y) \in \{(u',y') : 0 < u' < \infty,\, g(y') \le -u'\}$.

- Since $g(y) \le -u \;\Leftrightarrow\; u \le -g(y)$ and we also have $0 < u < \infty$, $(u,y) \in A_2$ implies $0 < u \le -g(y)$.
- To show that also $g(y) < 0$ we observe that $g(y) \le -u < -0 = 0$.

Finally, to show $B_2 \subseteq A_2$, let $(u,y) \in B_2 = \{(u',y') : g(y') < 0,\, 0 < u' \le -g(y')\}$.

- $0 < u < \infty$ is immediate from $0 < u \le -g(y)$. We still must show that $g(y') \le -u$.
- To show that also $g(y) < 0$ we observe that $g(y) \le -u < -0 = 0$. But this is immediate from $0 < u \le -g(y)$.

(c2) We proved that $A_2 = B_2$. It follows that $\displaystyle\iint_{A_2} f_Y(y)\, d(t,y) = \iint_{B_2} f_Y(y)\, d(t,y)\,.$

We apply (c1) and (c2) to the integrals $\int_0^\infty P\{g\circ Y > u\}du$ and $\int_0^\infty P\{g\circ Y \le -u\}du$ as follows.

$$\begin{aligned}\int_0^\infty P\{g\circ Y > u\}du &\overset{\text{(a)}}{=} \int_0^\infty P\{Y \in g^{-1}(]u,\infty[)\}du = \int_0^\infty P_Y\{g^{-1}(]u,\infty[)\}du \\ &= \int_0^\infty P_Y\{y : u < g(y) < \infty\}du = \int_0^\infty \left(\int_{\{y:u<g(y)<\infty\}} f_Y(y)\,dy\right) du \\ &\overset{\text{(b1)}}{=} \iint_{A_1} f_Y(y)\, d(t,y) \overset{\text{(c1)}}{=} \iint_{B_1} f_Y(y)\, d(t,y) \overset{\text{(b1)}}{=} \int_{\{y:g(y)>0\}} \left(\int_0^{g(y)} du\right) f_Y(y)\,dy\end{aligned}$$

Hence, since $\int_0^{g(y)} du = g(y)$,

(d1) $$\int_0^\infty P\{g\circ Y > u\}du = \int_{\{y:g(y)>0\}} g(y)\, f_Y(y)\, dy$$

$$\begin{aligned}\int_0^\infty P\{g\circ Y \le -u\}du &\overset{\text{(a)}}{=} \int_0^\infty P\{Y \in g^{-1}(]-\infty,-u[)\}du = \int_0^\infty P_Y\{g^{-1}(]-\infty,-u[)\}du \\ &= \int_0^\infty P_Y\{y : -\infty < g(y) < -u\}du = \int_0^\infty \left(\int_{\{y:-\infty<g(y)<-u\}} f_Y(y)\,dy\right) du \\ &\overset{\text{(b2)}}{=} \iint_{A_2} f_Y(y)\, d(t,y) \overset{\text{(c2)}}{=} \iint_{B_2} f_Y(y)\, d(t,y) \overset{\text{(b2)}}{=} \int_{\{y:g(y)<0\}} \left(\int_0^{-g(y)} du\right) f_Y(y)\,dy\end{aligned}$$

Hence, since $\int_0^{-g(y)} du = -g(y)$,

(d2) $$\int_0^\infty P\{g\circ Y \le -u\}du = -\int_{\{y:g(y)<0\}} g(y)\, f_Y(y)\, dy$$

It follows from (d1) and (d2) and Theorem 10.4 on p.219 and

$$\int_{\{y:g(y)=0\}} g(y) f_Y(y)\,dy = \int_{\{y:g(y)=0\}} 0 f(y) dy = 0\,,$$

that

$$\begin{aligned} E[g\circ Y] &= \int_0^\infty P\{g\circ Y > u\}du - \int_0^\infty P\{g\circ Y \le -u\}du \\ &= \int_{\{y:g(y)>0\}} g(y)\, f_Y(y)\, dy + \int_{\{y:g(y)<0\}} g(y)\, f_Y(y)\, dy \\ &= \int_{\{y:g(y)>0\}} g(y)\, f_Y(y)\, dy + \int_{\{y:g(y)<0\}} g(y)\, f_Y(y)\, dy + \int_{\{y:g(y)=0\}} g(y) f_Y(y)\, dy \\ &= \int_{\mathbb{R}} g(y)\, f_Y(y)\, dy = \int_{-\infty}^{\infty} g(y)\, f_Y(y)\, dy \ \blacksquare \end{aligned}$$

The following corresponds to WMS Theorem 4.5.

Theorem 10.6.

Let $c \in \mathbb{R}$, Y be a discrete or continuous random variable and $g_1, g_2, g_n : \mathbb{R} \to \mathbb{R};\ y \mapsto g(y)$ be a list of n real-valued functions. Then

$$E[c] = c\,, \tag{10.19}$$

$$E[cg_j(Y)] = cE[g_j(Y)]\,. \tag{10.20}$$

Further, the random variable

$$\sum_{j=1}^n g_j\circ Y : \Omega \longrightarrow \mathbb{R}; \qquad \omega \mapsto \sum_{j=1}^n g_j\big(Y(\omega)\big)$$

has the following expected value:

$$E\left[\sum_{j=1}^n g_j\circ Y\right] = \sum_{j=1}^n E[g_j\circ Y]\,. \tag{10.21}$$

PROOF: ■

We will not deal in this course with the sums of continuous and discrete random variables, so the next definition is only included for completeness' sake and to allow the formulation of theorems 10.7 and 10.8 below.

Definition 10.6. ⋆

If $Y_1, Y_2, \ldots, Y_m$ is a list of discrete random variables and $Y_1', Y_2', \ldots, Y_n'$ is a list of continuous random variables, all of which are defined on the same probability space (Ω, P), then we define

$$E\left[\sum_{i=1}^m Y_i + \sum_{j=1}^n Y_j'\right] := \sum_{i=1}^m E[Y_i] + \sum_{j=1}^n E[Y_j']\,p\,. \ \square \tag{10.22}$$

The following is the continuous random variables version of Theorem 9.4 on p.194.

Theorem 10.7.

*Let $Y_1, Y_2, \ldots, Y_n : \Omega \to \mathbb{R}$ be random variables. (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$ by Assumption 10.2.**B**). Some may be continuous, others may be discrete. Then the random variable*

$$\sum_{j=1}^{n} Y_j : \Omega \longrightarrow \mathbb{R}; \qquad \omega \mapsto \sum_{j=1}^{n} Y_j(\omega)$$

has the following expected value:

$$E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j]\,. \tag{10.23}$$

In other words, the expectation of the sum is the sum of the expectations.

PROOF: Not given here. ■

We extend Definition 9.3 on p.195 of the variance and standard deviation of a discrete random variable to the continuous case without modification, i.e.,

$$Var[Y] := \sigma_Y^2 := E\big[(Y - E[Y])^2\big]\,, \tag{10.24}$$

$$\sigma_Y := \sqrt{Var[Y]}\,. \tag{10.25}$$

Theorems 9.5, 9.6 9.7 about the variances of discrete random variables have the following counterpart.

Theorem 10.8. *Let Y be a discrete or continuous random variable. Let $Y_1, Y_2, \ldots, Y_n : \Omega \to \mathbb{R}$ be independent random variables (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$ by Assumption 10.2.**B**). Some may be continuous, others may be discrete. Further, let $a, b \in \mathbb{R}$. Then*

$$Var[Y] = E\big[Y^2\big] - \big(E[Y]\big)^2\,, \tag{10.26}$$

$$Var\,[aY + b] = a^2 Var[Y]\,, \tag{10.27}$$

$$Var\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var[Y_j]\,. \tag{10.28}$$

PROOF: The proof of (10.26) is the same as for Theorem 9.5 on p.195. The proof of the other formulas is not given here. ■

Remark 10.7. Note that independence of $Y_1, \ldots, Y_n$ is required for the validity of (10.28)! □

Example 10.2. [98] A business has daily revenues R and costs C of which it is known that

[98] This a corrected version of WMS Exercise 5.113.

- $R \sim \mathcal{N}(\mu = 50, \sigma^2 = 9)$ • $C \sim chi^2(\text{df} = 8)$ • R and C are independent.

Assuming that R and C are given in thousands of dollars,

a What are expected value and variance of the daily profit?
b Is it likely that tomorrow's profit will exceed $70,000$ dollars?

Solution:

Let Y denote the daily profit. Note that

- $E[R] = \mu = 50$ • $E[C] = \text{df} = 8)$ • $Var[R] = \sigma^2 = 9$ • $Var[C] = 2\,\text{df} = 16)$.

Since $Y = R - C$, we obtain $E[Y] = E[R] - E[C] = 42$.

Also, by independence, $Var[Y] = Var[R] + Var[C] = 25$.

Since $(70 - 48)/5 = 28/5 = 5.6$, tomorrow's profit would have to rise above 5.6 SDs [99] to exceed $70,000$ dollars. That seems extremely unlikely. □

The moments about the origin μ'_k, the moments about the mean μ_k and the MGF $m_Y(t)$ of a discrete random variable Y, all were defined as expected values. This allows us to use those same definitions for continuous random variables.

Unless something different is stated, Y is a random variable $Y : (\Omega, P) \to \mathbb{R}$ on some probability space (Ω, P). Further, $\mu = E[Y]$, $\sigma^2 = Var[Y]$ and $\sigma = \sqrt{Var[Y]}$ denote expectation, variance and standard deviation of Y.

Definition 10.7. For $k \in \mathbb{N}$, we define

$$\mu'_k := E[Y^k] \qquad (k\text{th } \textbf{moment of } Y \textbf{ about the origin}) \tag{10.29}$$

$$\mu_k := E[(Y - E[Y])^k] = E[(Y - \mu)^k] \qquad (k\text{th } \textbf{central moment of } Y) \tag{10.30}$$

$$m(t) := m_Y(t) := E\left[e^{tY}\right] \qquad (\textbf{moment–generating function of } Y) \tag{10.31}$$

As in the discrete case we assume that the expectations defining μ'_k and μ_k exist and that there is some $\delta > 0$ such that $m_Y(t)$ is defined (i.e., finite) for $|t| < \delta$. □

Theorem 9.18 on p.209 remains valid for continuous random variables:

Theorem 10.9.

Let Y be a random variable with MGF $m_Y(t)$ and $k \in \mathbb{N}$. Then its kth moment is obtained as the kth derivative of $m_Y(\cdot)$, evaluated at $t = 0$:

$$\mu'_k = m^{(k)}(0) = \frac{d^k m(t)}{dt^k}\Big|_{t=0}. \tag{10.32}$$

PROOF: The proof of Theorem 9.18 can be used without any alterations. ■

[99] WMS erroneously states this figure as 7.2 SDs

Proposition 10.5.

Let Y be a random variable with MGF $m_Y(t)$. Let $a, b \in \mathbb{R}, Y' := Y + a, Y'' := bY$. Then

$$m_{Y'}(t) = e^{ta} m_Y(t)\,, \tag{10.33}$$

$$m_{Y''}(t) = m_Y(bt)\,. \tag{10.34}$$

PROOF: To prove (10.33), we note that e^{ta} is constant in ω. Thus, $E[e^{ta}W] = e^{ta}E[W]$ for any random variable W. Thus,

$$m_{Y'}(t) = E[e^{t(Y+a)}] = E[e^{tY}e^{ta}] = e^{ta}E[e^{tY}] = e^{ta}m_Y(t)\,.$$

Formula (10.34) follows from

$$m_{Y''}(t) = E[e^{t(bY)}] = E[e^{(tb)Y}] = m_Y(tb)\,. \quad \blacksquare$$

10.4 The Uniform Probability Distribution

Given two real numbers $\theta_1 < \theta_2$, we consider a random variable $Y(\omega)$ that "lives" in the interval $[\theta_1, \theta_2]$, i.e., $P\{\theta_1 \le Y \le \theta_2\} = 1$ and has the same likelyhood of occurring in any subinterval of same length:

Definition 10.8 (Continuous, uniform random variable)**.**

Let Y be a random variable and $-\infty < \theta_1 < \theta_2 < \infty$. We say that Y has a **continuous uniform probability distribution** with parameters θ_1 and θ_2 — also, that Y **is uniform on** $[\theta_1, \theta_2]$ or $Y \sim$ **uniform**(θ_1, θ_2) — if Y has probability density function

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}\,, & \text{if } \theta_1 \le y \le \theta_2, \\ 0\,, & \text{else.} \end{cases} \tag{10.35}$$

□

Remark 10.8 (uniform and equiprobable probability measures)**.** Uniform distributions are the equivalent of the distribution of discrete rancom variables Y that satisfy equiprobability, i.e., their PMF $p_Y(y) = P\{Y = y\}$ is strictly positive only for finitely many numbers $y_1, y_2, \ldots, y_n$ and $p_Y(y_j) = 1/n$ for all $j \in [1, n]_{\mathbb{Z}}$. See Definition 5.3 on p.109. □

Theorem 10.10 (WMS Ch.04.4, Theorem 4.6)**.**

If $\theta_1 < \theta_2$ and Y is a uniform random variable with parameters θ_1, θ_2, then

$$E[Y] = \frac{\theta_1 + \theta_2}{2} \qquad \textit{and} \qquad Var[Y] = \frac{(\theta_2 - \theta_1)^2}{12}\,.$$

PROOF: A simple exercise in integrating $\int_{\theta_1}^{\theta_2} y\,dy$ and $\int_{\theta_1}^{\theta_2} y^2\,dy$. ■

Theorem 10.11.

Assume that Y is a continuous random variable with CDF $F_Y(y)$. Let $U := F_Y(Y)$. Then $U \sim$ uniform$(0,1)$.

SIMPLIFIED PROOF under the assumption that the CDF F_Y is a bijection $F_Y : \mathbb{R} \xrightarrow{\sim}]0,1[$.
The inverse F_Y^{-1} of F_Y satisfies $F_Y^{-1}\big(F_Y(y)\big) = y$ for all $y \in \mathbb{R}$. Thus, for $0 < u < 1$,

$$\begin{aligned} F_U(u) &= P\{U \le u\} \;=\; P\{F_Y \circ Y \le u\} \;=\; P\{F_Y^{-1} \circ F_Y \circ Y \le F_Y^{-1}(u)\} \\ &= P\{Y \le F_Y^{-1}(u)\} \;=\; F_Y\big(F_Y^{-1}(u)\big) \;=\; u\,. \end{aligned}$$

We still must handle the cases $u \le 0$ and $u \ge 1$. We assumed that the codomain of F_Y is $]0,1[$.

- Thus, $y \in \mathbb{R} \;\Rightarrow\; 0 < F_Y(y) < 1$.
- Thus, $\omega \in \Omega \;\Rightarrow\; 0 < U(\omega) = F_Y\big(Y(\omega)\big) < 1$
  $\Rightarrow \big[P\{U \le 0\} = 0 \text{ and } P\{U \le 1\} = 1\big] \Rightarrow \big[F_U(0) = 0 \text{ and } F_U(1) = 1\big]$.
- Thus, $\big[u \le 0 \;\Rightarrow\; F_U(u) \le F_U(0) = 0\big]$ and $\big[u \ge 1 \;\Rightarrow\; F_U(u) \ge F_U(1) = 1\big]$.

It follows that F_U is the CDF of a uniform$(0,1)$ random variable. Thus, $U \sim$ uniform$(0,1)$. ■

GENERAL PROOF $\boxed{\star}$ (We drop the assumption that F_Y is a bijection $\mathbb{R} \xrightarrow{\sim}]0,1[$.):
This proof follows the one of Theorem 2.1.10 in Casella, Berger [3], but it gives additional detail.
Let $0 < p < 1$ and let

(A) $$G(p) \;:=\; \min\{y \in \mathbb{R} : F_Y(y) \ge p\}\,.$$

In other words, $G(p)$ is the pth quantile ϕ_p for the random variable Y. Since G is nondecreasing,

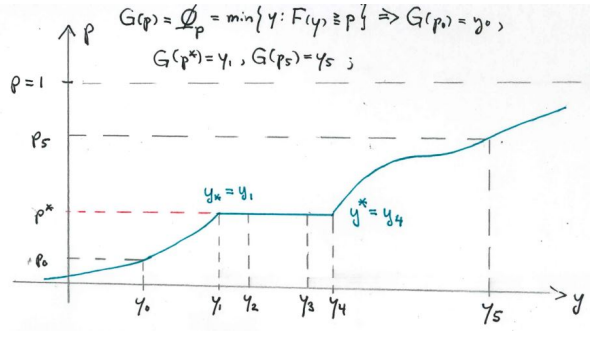
(B) $$F_U(p) \;= P\{U \le p\} \;=\; P\{F_Y(Y) \le p\} \;=\; P\{G\big(F_Y(Y)\big) \le G(p)\}\,.$$

The most difficult part of the proof is to show that

(C) $$P\{G\big(F_Y(Y)\big) \le G(p)\} \;=\; P\{Y \le G(p)\}\,.$$

We consider two different cases.

- **Case 1:** There is a unique y such that $G(p) = y$. In the picture, that would be y_0 for p_0 and y_5 for p_5
  (a) Observe that $G(p) = y \;\Leftrightarrow\; p = F_Y(y)$.
  (b) $G(p') < G(p) < G(p'') \;\Leftrightarrow\; p' < p < p''$.
- **Case 2:** There are $y_* < y^*$, determined by $G(p) = y \;\Leftrightarrow\; y_* < y < y^*$. In the picture, that would be $y_* = y_1$ and $y^* = y_4$ for $F(y) = p$.



10.1 (Figure). **non–injective, continuous CDF.**

We now show that **(C)** is true for **Case 1**.
We deduce from **(a)** and **(b)** that

$$\begin{aligned}\omega \in \{G\big(F_Y(Y)\big) \le G(p)\} &\Leftrightarrow F_Y\big(Y(\omega)\big) \le G(p) \ \big(= F_Y(y)\big)\\ &\Leftrightarrow Y(\omega) \le y \ \big(= G(p)\big) \Leftrightarrow \omega \in \{Y \le G(p)\}\,.\end{aligned}$$

Taking probabilities shows that **(C)** is valid, since we obtain

$$P\{G\big(F_Y(Y)\big) \le G(p) = P\{Y \le G(p)\}\,.$$

Next, we show that **(C)** is true for **Case 2**.
The picture shows that, if $F_Y(y') = p'$ and $F_Y(y) = p \Leftrightarrow y_* \le y \le y^*$, then

(c) $G(p') < G(p) \Leftrightarrow y' < y_*$; $G(p') = G(p) \Leftrightarrow$; $y_* \le y' \le y^*$;
(d) Thus, $G(p') \le G(p) \Leftrightarrow y' \le y^* \Leftrightarrow$; $\big[y' \le y_*$ or $y_* < y' \le y^*$.
Clearly,

$$\omega \in \{G\big(F_Y(Y)\big) \le G(p)\} \Leftrightarrow G\big(F_Y(Y(\omega))\big) \le G(p)(= y_*)\}\,.$$

We apply **(d)** with $y' = Y(\omega)$ and $p' = F_Y\big(Y(\omega)\big)$ and obtain

$$G\big(F_Y(Y(\omega))\big) \le G(p) \Leftrightarrow \big[\,Y(\omega) \le y_* \text{ or } y_* < Y(\omega) \le y^*\,\big]\,.$$

Thus, $\{G\big(F_Y(Y)\big) \le G(p)\} = \{Y \le y_*\} \uplus \{y_* < Y \le y_*\}$. Taking probabilities,

$$\begin{aligned}P\{G\big(F_Y(Y)\big) \le G(p)\} &= P\{Y \le y_*\} + P\{y_* < Y \le y_*\}\\ &= F_Y(y_*) + \big(F_Y(y^*) - F_Y(y_*)\big) = F_Y\big(G(p)\big) = P\{Y \le G(p)\}\,.\end{aligned}$$

Here, the equation next to the last follows from $G(p) = y_*$ and $F_Y(y_*) = G(p) = F_Y(y^*)$.
We have shown that **(C)** also is true for **Case 2**.
We combine **(B)** and **(C)** and obtain

(**D**) $$F_U(p) = P\{F_Y(Y) \le p\} = P\{Y \le G(p)\} = F_Y\big(G(p)\big)\,.$$

Our next goal is to show that $F_Y\big(G(p)\big) = p$. We break this down into the following steps.

(1) By **(A)**, $F_Y\big(G(p)\big) \ge p$. We now show that also $F_Y\big(G(p)\big) \le p$.
(2) Let $y_n := G(p) - 1/n$. Then $G(p) = \lim\limits_{n\to\infty} y_n$.
(3) $G(p)$ being the smallest y such that $F_Y(y) \ge p$ implies that $F_Y(y_n) < p$.
(4) Since Y is continuous, $F(y)$ is continuous. Thus, $F_Y\big(G(p)\big) = \lim\limits_{n\to\infty} F_Y(y_n)$.
(5) Since $F_Y(y_n) < p$ by **(3)**, $\lim\limits_{n\to\infty} F_Y(y_n) \le p$, i.e., $F_Y\big(G(p)\big) \le p$. (See **(4)**.)
(6) We have shown **(1)** and it follows that $F_Y\big(G(p)\big) = p$.

It now follows from **(D)** that $P\{U \le p\} = p$ for any $0 < p < 1$.

The boundary cases $p = 0$ and $p = 1$ are taken into account by extending the definition of $G(p)$ given in **(A)**, which is $G(p) = \min\{y \in \mathbb{R} : F_Y(y) \ge p\}$, as follows.

- Since $F_Y(y) \ge 0$ for all y, it is natural to define $G(0) := -\infty$.
- If there is some y_* such that $F_Y(y_*) = 1$, then **(A)** remains in force for $G(1)$.
- Otherwise, (if $F_Y(y) < 1$ for all y), we define $G(1) := \infty$. ■

Theorem 10.12.

Given are a uniform$(0,1)$ random variable U and a continuous function $F:\mathbb{R}\to[0,1]$ that satisfies the conditions of Theorem 10.1 (Properties of a Cumulative Distribution Function) on p.212:
• *F is nondecreasing* • $F(-\infty) := \lim_{y\to-\infty} F(y) = 0$ • $F(\infty) := \lim_{y\to\infty} F(y) = 1$

(10.36) *Let* $G:[0,1]\to\mathbb{R};\quad p\mapsto G(p) := \min\{y\in\mathbb{R} : F(y)\geq p\}$.

Let $Z := G(U)$ be the random variable $\omega\mapsto Z(\omega) := G\big(U(\omega)\big)$.
Then its CDF matches F. In other words, $F_Z(y) = F(y)$ for all $y\in\mathbb{R}$.

SIMPLIFIED PROOF under the assumption that the F is a bijection $F:\mathbb{R}\xrightarrow{\sim}]0,1[$.
We first show that G is the inverse of F.

- Since F is both nondecreasing and injective, F is strictly increasing.
- Let $0<p_0<1$ and $y_0 := F^{-1}(p_0)$ or, equivalently, $p_0 = F(y_0)$.
- Let $A := \{y\in\mathbb{R} : F(y)\geq p_0\}$. Since $F(y_0) = p_0 \geq p_0$, it follows that $y_0\in A$.
- Since F is strictly increasing, $y<y_0 \;\Rightarrow\; F(y)<F(y_0)=p_0 \;\Rightarrow\; y\notin A$
- Since $y_0\in A$ and $y<y_0 \;\Rightarrow\; y\notin A$, we conclude that $y_0 = \min(A)$.
- By (10.36), $G(p_0) = \min(A)$. We have shown $G(p_0) = y_0 = F^{-1}(p_0)$ for each $0<p_0<1$.

Let $y\in\mathbb{R}$. Since $G = F^{-1}$, we obtain

$$F_Z(y) = P\{Z\leq y\} = P\{G\circ U\leq y\} = P\{F^{-1}\circ U\leq y\} = P\{U\leq F(y)\} = F(y)\,.$$

The last equation follows from $0<F(y)<1$ and $U\sim$ uniform$(0,1)$. It follows that $F_Z(y) = F(y)$ for all y, i.e., $F_Y = F$ ■

GENERAL PROOF ⋆ (We drop the assumption that F_Y is a bijection $\mathbb{R}\xrightarrow{\sim}]0,1[$.):
Let $I := F_Y(\mathbb{R}) = \{F_Y(y) : y\in\mathbb{R}\}$ be the range of F_Y.

- Note that $G(p)$ equals the pth quantile ϕ_p of a random variable with CDF $F(y)$. (See Definition 10.4 on p.216.)
- Further, the continuity of F guarantees that for each $0<p<1$ one can find $y\in\mathbb{R}$ such that $F(y)=p$ (and thus, $p\mapsto G(p)$ is injective).
- Thus, I is one of the following intervals: ⊡ If $0<F(y)<1$ for all y, then $I=]0,1[$
  ⊡ If $0\leq F(y)<1$ for all y, then $I=[0,1[$ ⊡ If $0<F(y)\leq 1$ for all y, then $I=]0,1]$
  ⊡ If $0\leq F(y)\leq 1$ for all y, then $I=[0,1]$
- We will refer in this proof to Figure 10.1 on p.227 (non–injective, continuous CDF) in the proof of Theorem 10.11.

We fix $y\in\mathbb{R}$. Let $p := F(y)$. Then

(a) Since F is continuous and nondecreasing, there are numbers $y_*\leq y_*$ such that $F(\tilde{y}) = p \;\Leftrightarrow\; y_*\leq\tilde{y}\leq y_*$.
(b) Either F is strictly increasing at y and then $y_* = y = y_*$, or F is "flat around y" and $y_*<y_*$.
(c) For $p'\in I$, choose y' such that $F(y') = p'$. Then, since $F(y_*) = p$,
$p'<p \;\Leftrightarrow\; F(y')<p \;\Leftrightarrow\; y'<y_*$ and $p'\leq p \;\Leftrightarrow\; F(y')\leq p \;\Leftrightarrow\; y'\leq y^* \;\Leftrightarrow\; G(p')\leq y^*$.
(d) Further, since F is nondecreasing, G also is nondecreasing. Thus, $p'\leq p \;\Leftrightarrow\; G(p')\leq G(p)$.
It follows from **(c)** that $p'\leq p \;\Leftrightarrow\; G(p')\leq G(p) \;\Leftrightarrow\; y'\leq y^* \;\Leftrightarrow\; G(p')\leq y^*$.

Let $\omega \in \Omega$ and $p' := U(\omega)$. Recall that $p = F(y)$. Then

$$G\big(U(\omega)\big) \;\le\; y \;\Leftrightarrow\; \big[G(p') \le G(p)\big] \overset{\text{(d)}}{\Leftrightarrow} \big[p' \le p\big] \;\Leftrightarrow\; \big[U(\omega) \le F(y)\big]\,.$$

We take probabilities and obtain, since $U \sim$ uniform$(0,1)$ implies $P\{U \le \tilde{p}\} = \tilde{p}$ for $0 \le \tilde{p} \le 1$,

$$F_Z(y) \;= P\{G(U) \;\le\; y\} \;=\; P\{U \;\le\; F(y)\} \;=\; F(y)\,.$$

To summarize, we have shown that $F_Z(y) = F(y)$ for all $y \in \mathbb{R}$. ■

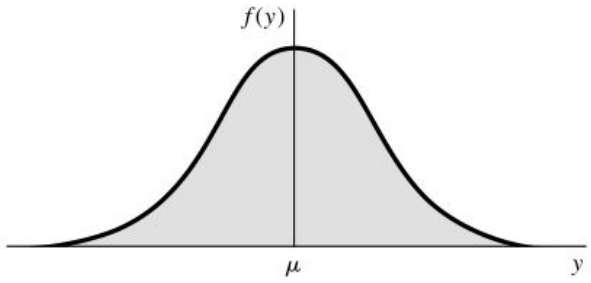
Remark 10.9. A special case of Theorem 10.12 can be found in WMS Ch.06.3, Example 6.5, which shows how to solve the following problem: Let U be a uniform random variable on the interval $(0,1)$. Find a transformation $G(U)$ such that $G(U)$ possesses an exponential distribution with mean β. □

10.5 The Normal Probability Distribution

Many numerical random phenomena yield histograms which are approximately unimodal (a single highest value) and symmetric around the mean μ, like the picture to the right, and they adhere to the **empirical rule**: Approximately

- 68% of the data fall between $\mu \pm 1 \cdot \sigma$
- 95% of the data fall between $\mu \pm 2 \cdot \sigma$
- 99.7% of the data fall between $\mu \pm 3 \cdot \sigma$

Such data are adequately modeled by the normal distribution.



Source: WMS Ch.4.5

The empirical rule is also known as the **68%–95%–99.7% rule**.

Definition 10.9 (Normal random variable)**.**

Let $\sigma > 0$ and $-\infty < \mu < \infty$. We say that a random variable Y has a **normal probability distribution** with mean μ and variance σ^2 if its probability density function is

$$f_Y(y) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, \qquad (y \in \mathbb{R})\,. \;\square \tag{10.37}$$

We also express that by saying that Y is $\mathcal{N}(\mu, \sigma^2)$. Moreover, we call Y **standard normal** if Y is $\mathcal{N}(0,1)$.

We will see that $E[Y] = \mu$ and $Var[Y] = \sigma^2$. This justifies calling the parameters μ and σ^2 the mean and variance of the distribution.

Lemma 10.1.

$$(y-\mu)^2 \;-\; 2yt\sigma^2 \;= \big[y \;-\; (\mu + t\sigma^2)\big]^2 \;-\; 2\mu t\sigma^2 \;-\; t^2\sigma^4\,. \tag{10.38}$$

PROOF: We multiply out the right–hand expression and obtain

$$\begin{aligned}\text{R.S.} &= \big[y - (\mu + t\sigma^2)\big]^2 - 2\mu t\sigma^2 - t^2\sigma^4 \\ &= y^2 - 2y(\mu + t\sigma^2) + (\mu^2 + 2\mu t\sigma^2 + t^2\sigma^4) - 2\mu t\sigma^2 - t^2\sigma^4 \\ &= y^2 - 2\mu y - 2yt\sigma^2 + \mu^2 \\ &= (y-\mu)^2 - 2yt\sigma^2 = \text{L.S.} \ \blacksquare\end{aligned}$$

Proposition 10.6.

Let the random variable Y be $\mathcal{N}(\mu, \sigma^2)$. Then

$$m_Y(t) = e^{\mu t + (\sigma^2 t^2)/2}. \tag{10.39}$$

PROOF:

$$\begin{aligned}m_Y(t) &= \int_{-\infty}^{\infty} e^{yt} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(yt)(2\sigma^2)}{2\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}\left[(y-\mu)^2 - 2yt\sigma^2\right]} dy\,.\end{aligned}$$

We apply Lemma 10.1 and obtain for the exponent the following.

$$\begin{aligned}-\frac{1}{2\sigma^2}\left[(y-\mu)^2 - 2yt\sigma^2\right] &= -\frac{1}{2\sigma^2}\left\{\big[y - (\mu + t\sigma^2)\big]^2 - 2\mu t\sigma^2 - t^2\sigma^4\right\} \\ &= -\frac{\big[y - (\mu + t\sigma^2)\big]^2}{2\sigma^2} + \frac{1}{2\sigma^2}\big[2\mu t\sigma^2 + t^2\sigma^4\big] \\ &= \mu t + \frac{t^2\sigma^2}{2} - \frac{1}{2}\left[\frac{y - (\mu + t\sigma^2)}{\sigma}\right]^2\end{aligned}$$

It follows that

$$\begin{aligned}m_Y(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\mu t + \frac{t^2\sigma^2}{2}} e^{-\frac{1}{2}\left[\frac{y-(\mu+t\sigma^2)}{\sigma}\right]^2} dy \\ &= e^{\mu t + \frac{t^2\sigma^2}{2}} \left[\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{y-(\mu+t\sigma^2)}{\sigma}\right)^2} dy\right].\end{aligned}$$

The expression in square brackets is the integral $\int_{-\infty}^{\infty} \varphi(y)dy$, where $\varphi(y)$ is the PDF of a normal variable with mean $\mu + t\sigma^2$ and variance σ^2. Thus, this integral evaluates to 1 and it follows that

$$m_Y(t) = e^{\mu t + \frac{t^2\sigma^2}{2}}. \ \blacksquare$$

Theorem 10.13 (WMS Ch.04.5, Theorem 4.7)**.**

If Y is a normally distributed random variable with parameters μ and σ, then

$$E[Y] \;=\; \mu \quad \textit{and} \quad Var[Y] \;=\; \sigma^2\,.$$

PROOF: We differentiate $m_Y(t) \;=\; \exp\{\mu t \,+\, \frac{t^2\sigma^2}{2}\}$ twice and obtain

$$\begin{aligned} m_Y'(t) \;&=\; \big(\mu + t\sigma^2\big)\,\exp\left\{\mu t \,+\, \frac{t^2\sigma^2}{2}\right\}\,, \\ m_Y''(t) \;&=\; \big(\mu + t\sigma^2\big)^2\,\exp\left\{\mu t \,+\, \frac{t^2\sigma^2}{2}\right\} \,+\, \sigma^2\,\exp\left\{\mu t \,+\, \frac{t^2\sigma^2}{2}\right\}\,. \end{aligned}$$

Thus,the first and second moment about the origin are

$$\begin{aligned} E[Y] \;&=\mu_1' \;=\; m_Y'(0) \;=\; (\mu+0)e^0 \;=\; \mu\,, \\ E[Y^2] \;&=\mu_2' \;=\; m_Y''(0) \;=\; (\mu+0)^2e^0 \,+\, \sigma^2e^0 \;=\; \mu^2 \,+\, \sigma^2\,. \end{aligned}$$

Finally,

$$Var[Y] \;=\; E[Y^2] \,-\, \big(E[Y]\big)^2 \;=\; \mu^2 \,+\, \sigma^2 \,-\, \mu^2 \;=\; \sigma^2\,. \;\blacksquare$$

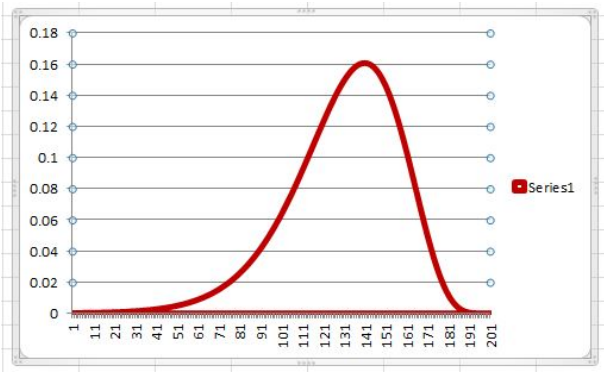
Remark 10.10. The importance of the normal distribution stems from the so called Central Limit Theorem (Theorem 13.13 on p.331), which we will discuss in Chapter 13 (Limit Theorems). It states the following.

- Given is an iid sequence of random variables $Y_1, Y_2, \ldots$ with common expectation $\mu := E[Y_j]$ and finite standard deviation $\sigma := \sqrt{Var[Y_j]} < \infty$ and a standard normal variable Z.
- For $n \in \mathbb{N}$, we define $\bar{Y}_n \;:=\; \frac{1}{n}\sum\limits_{j=1}^{n} Y_j \;=\; \dfrac{Y_1 + \cdots + Y_n}{n}$ and $Z_n \;:=\; \dfrac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$.
- An aside: One easily sees from Theorems 10.7 (p.224) and 10.8 that $E[\bar{Y}_n] = \mu,\ \sigma_{\bar{Y}_n} = \sigma/\sqrt{n}$ and thus, $E[Z_n] = 0,\ Var[Z_n] = 1$.
- The Central Limit Theorem states that for each fixed $z \in \mathbb{R}$, $F_{Z_n}(z)$ converges to $F_Z(z)$.
- In other words, $\lim\limits_{n\to\infty} P\{Z_n \le z\} \;=\; \lim\limits_{n\to\infty} F_{Z_n}(z) \;=\; F_Z(z) \;=\; \displaystyle\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt$ for all z.

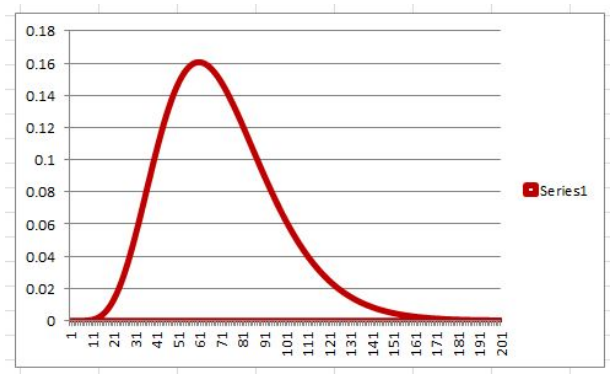
□

10.6 The Gamma Distribution

Whereas the normal distribution is a good fit for histograms which are symmetric, many random phenomena yield **left skewed** (also referred to as **left tailed**) or **right skewed** (also referred to as **right tailed**) histograms which are more appropriately modeled by distributions which themselves also are left or right skewed.



Left skewed distribution



Right skewed distribution

The gamma distribution which we discuss here can be used to generate all kinds of right skewed distributions.

Definition 10.10 (Gamma random variable)**.**

Let $\sigma > 0$ and $-\infty < \mu < \infty$. We say that a random variable Y has a **gamma distribution** with **shape parameter** $\alpha > 0$ and **scale parameter** $\beta > 0$ if its probability density function is

$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}, & \text{if } 0 \le y < \infty, \\ 0, & \text{else}, \end{cases} \tag{10.40}$$

where $\Gamma(\alpha)$ is the **gamma function**

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy. \tag{10.41}$$

We also express that by saying that Y is gamma(α, β). □

Proposition 10.7. *The gamma function satisfies the following:*

$$\Gamma(1) = 1, \tag{10.42}$$

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \quad \textit{for all } \alpha > 1, \tag{10.43}$$

$$\Gamma(n) = (n-1)! \quad \textit{for all } n \in \mathbb{N}. \tag{10.44}$$

PROOF: (10.42) is immediate from $\int_0^\infty e^{-y}dy = -e^{-y}\Big|_0^\infty = 0-(-1) = 1.$

We obtain (10.43) from integration by parts of $\Gamma(\alpha)$:

$$\begin{aligned} \Gamma(\alpha) &= y^{\alpha-1}\big(-e^{-y}\big)\Big|_0^\infty + \int_0^\infty (\alpha-1)y^{\alpha-2}e^{-y}\,dy \\ &= 0 + (\alpha-1)\int_0^\infty y^{(\alpha-1)-1}e^{-y}\,dy \\ &= (\alpha-1)\Gamma(\alpha-1). \end{aligned}$$

To show (10.44) we observe that repeated application of (10.43) yields

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1)\\ &= (n-1)(n-2)\Gamma(n-2)\\ &= (n-1)(n-2)(n-3)\cdots 2\Gamma(2)\\ &= (n-1)(n-2)(n-3)\cdots 2\cdot 1\Gamma(1)\,.\end{aligned}$$

Since $\Gamma(1) = 1$ by (10.42), it follows that

$$\Gamma(n) \;=\; (n-1)(n-2)(n-3)\cdots 2\cdot 1 \;=\; (n-1)!\,.$$

Proposition 10.8.

If the random variable Y is gamma(α,β) it has MGF

(10.45) $$m_Y(t) \;=\; \frac{1}{(1-\beta t)^\alpha} \qquad \textit{for } t \;<\; \frac{1}{\beta}\,.$$

PROOF: ⋆ We define

(A) $$\tilde\beta \;:=\; \frac{\beta}{1-t\beta}$$

and observe that $\tilde\beta > 0$ for $1 - t\beta > 0$, i.e., for $t < 1/\beta$. Further,

(B) $$ty \;-\; \frac{y}{\beta} \;=\; \frac{(-y+ty\beta)}{\beta} \;=\; \frac{-y(1-t\beta)}{\beta} \;=\; -y \Big/ \frac{\beta}{(1-t\beta)} \;=\; \frac{-y}{\tilde\beta}\,.$$

Thus,

$$\begin{aligned} m_Y(t) \;&= E(e^{tY}) \;=\; \int_0^\infty e^{ty}\left[\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right] dy\\ &= \frac{1}{\beta^\alpha}\int_0^\infty \frac{y^{\alpha-1}}{\Gamma(\alpha)}\exp\left[ty \;-\; \frac{y}{\beta}\right] dy \overset{\textbf{(B)}}{=} \frac{1}{\beta^\alpha}\int_0^\infty \frac{y^{\alpha-1}\,e^{-y/\tilde\beta}}{\Gamma(\alpha)}\,dy\end{aligned}$$

Part of **(B)** is $\dfrac{-y(1-t\beta)}{\beta} \;=\; \dfrac{-y}{\tilde\beta}$. Thus, $(1-t\beta)\,\tilde\beta \;=\; \beta$; thus, $\beta^\alpha \;=\; (1-t\beta)^\alpha\cdot\tilde\beta^\alpha$; thus,

$$m_Y(t) \;=\; \frac{1}{(1-t\beta)^\alpha}\cdot\int_0^\infty \frac{y^{\alpha-1}\,e^{-y/\tilde\beta}}{\tilde\beta^\alpha\Gamma(\alpha)}\,dy \;=\; \frac{1}{(1-t\beta)^\alpha}\cdot\int_0^\infty \varphi(y)\,dy\,.$$

Here, the function $\varphi(y)$ is the PDF of a gamma$(\alpha,\tilde\beta)$ random variable. It follows that $\int_0^\infty \varphi(y)\,dy = 1$ and we conclude that $m_Y(t) = 1/(1-t\beta)^\alpha$. ■

Theorem 10.14 (WMS Ch.04.6, Theorem 4.8)**.**

Let the random variable Y be gamma(α, β) with $\alpha, \beta > 0$. Then

$$E[Y] \;=\; \alpha\beta \quad \textit{and} \quad Var[Y] \;=\; \alpha\beta^2\,.$$

PROOF: We obtain those results by differentiating the MGF of Y.

$$\begin{aligned} m_Y(t) \;=\; (1-\beta t)^{-\alpha} \;&\Rightarrow m_Y'(t) \;=\; (-\alpha)(1-\beta t)^{-\alpha-1}(-\beta) \\ &\Rightarrow m_Y''(t) \;=\; (-\alpha)(-\beta)(-\beta)(-\alpha-1)(1-\beta t)^{-\alpha-2}\,. \end{aligned}$$

Thus,

$$\begin{aligned} m_Y'(0) \;&=(-\alpha)(1-0)^{-\alpha-1}(-\beta) \;=\; \alpha\beta\,, \\ m_Y''(0) \;&=(-\alpha)\beta^2(-\alpha-1)(1-0)^{-\alpha-2} \;=\; (-\alpha)^2\beta^2 \;-\; (-\alpha)\beta^2 \;=\; \alpha^2\beta^2 \;+\; \alpha\beta^2\,. \end{aligned}$$

In other words, $E[Y] = \alpha\beta$ and $E[Y^2] = \alpha\beta^2$ From this,

$$Var[Y] \;=\; E[Y^2] \;-\; \big(E[Y]\big)^2 \;=\; (\alpha^2\beta^2 \;+\; \alpha\,\beta^2) \;-\; \alpha^2\beta^2 \;=\; \alpha\,\beta^2\,. \;\blacksquare$$

Definition 10.11 (Chi–square distribution)**.**

Let $\nu \in \mathbb{N}$. We say that a random variable Y has a **chi–square distribution** with ν **degrees of freedom**, in short, Y is **chi–square with ν df** or Y is **chi–square(ν)**, or Y is $\chi^2(\nu)$, if Y is gamma$(\nu/2, 2)$. In other words, Y must have a gamma distribution with shape parameter $\nu/2$ and scale parameter 2. $\square$

Theorem 10.15 (WMS Ch.04.6, Theorem 4.9)**.**

A chi–square random variable Y with ν degrees of freedom has expectation and variance

$$E[Y] \;=\; \nu \qquad \textit{and} \qquad Var[Y] \;=\; 2\nu.$$

PROOF: This follows from Theorem 10.14 with $\alpha = \nu/2$ and $\beta = 2$. $\blacksquare$

Definition 10.12 (Exponential distribution)**.**

We say that a random variable Y has an **exponential distribution** with parameter $\beta > 0$, in short, Y is **expon(β)**, if it has density function

$$f_Y(y) \;=\; \begin{cases} \dfrac{1}{\beta}\, e^{-y/\beta}\,, & \text{for } 0 \le y < \infty\,, \\ 0\,, & \text{else}\,. \end{cases} \qquad (10.46)$$

$\square$

Remark 10.11. In many textbooks exponential random variables are expressed in terms of $\lambda = 1/\beta$. Then its PDF is

$$f_Y(y) \;=\; \begin{cases} \lambda\, e^{-\lambda y}\,, & \text{for } 0 \le y < \infty\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{10.47}$$ □

Theorem 10.16.

An exponential random variable Y with parameter β has expectation and variance

$$E[Y] \;=\; \beta \qquad \textit{and} \qquad Var[Y] \;=\; \beta^2.$$

PROOF: This follows from Theorem 10.14 with $\alpha = 1$. ■

Proposition 10.9 (Memorylessness of the exponential distribution)**.** *Let Y be an exponential random variable. Let $t > 0$ and $h > 0$. Then*

$$P\{Y \;>\; t+h \mid Y > t\} \;=\; P\{Y \;>\; h\}\,. \tag{10.48}$$

PROOF: From the definition of conditional probability and

$$\{Y \;>\; t+h\} \cap \{Y > t\} \;=\; \{Y \;>\; t+h\}\,,$$

it follows that

$$P\{Y \;>\; t+h \mid Y > t\} \;=\; \frac{P\{Y \;>\; t+h\}}{P\{Y \;>\; t\}}\,.$$

We obtain

$$P\{Y \;>\; t+h\} \;=\; \int_{t+h}^{\infty} \frac{1}{\beta}\, e^{-y/\beta} dy \;=\; -\frac{1}{1/\beta} \cdot \frac{1}{\beta} \cdot e^{-y/\beta}\Big|_{t+h}^{\infty} \;=\; -e^{-y/\beta}\Big|_{t+h}^{\infty} \;=\; e^{-(t+h)/\beta}$$

and

$$P\{Y \;>\; t\} \;=\; \int_{t}^{\infty} \frac{1}{\beta}\, e^{-y/\beta} dy \;=\; -e^{-y/\beta}\Big|_{t}^{\infty} \;=\; e^{-t/\beta}\,.$$

Thus,

$$P\{Y \;>\; t+h \mid Y > t\} \;=\; \frac{e^{-(t+h)/\beta}}{e^{-t/\beta}} \;=\; e^{-h/\beta} \;=\; P\{Y \;>\; h\}\,.$$ ■

Remark 10.12. The property (10.48) of an exponential distribution is referred to as the **memoryless property** of the exponential distribution. It also occurs in the geometric distribution. □

10.7 The Beta Distribution

This chapter is merely a summary of the most impportant material of WMS Chapter 4.7 (The Beta Probability Distribution).

Like the gamma PDF, the beta density function is a two–parameter PDF defined over the closed interval $0 \le y \le 1$. y often plays the role of a proportion, such as the proportion of impurities in a chemical product or the proportion of time that a machine is under repair.

Definition 10.13 (Beta distribution). ⋆

A random variable Y has a **beta probability distribution** with parameters $\alpha > 0$ and $\beta > 0$ if it has density function

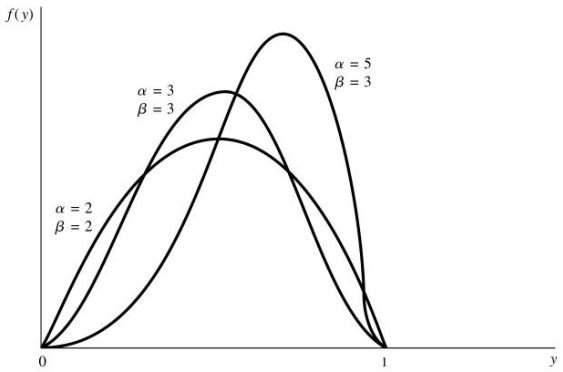
$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, & \text{if } 0 \le y \le 1, \\ 0, & \text{else}, \end{cases} \tag{10.49}$$

where

$$B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{10.50}$$

We also express that by saying that Y is beta(α,β). □

Beta density functions come in a large variety of shapes for different values of α and β. Some of these are shown in the figure to the right.
Note that $0 \le y \le 1$ does not restrict the use of the beta distribution. It can be applied to a random variable defined on the interval $c \le y \le d$ by means of the transformation $\tilde{y} = (y-c)/(d-c)$ which defines a new variable $0 \le \tilde{y} \le 1$ which has the correct domain for the beta density.



Beta density functions. Source: WMS

Theorem 10.17. ⋆

If Y is a beta–distributed random variable with parameters $\alpha > 0$ and $\beta > 0$, then

$$E[Y] = \frac{\alpha}{\alpha+\beta} \qquad \textit{and} \qquad Var[Y] = \frac{\alpha\,\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

PROOF: See the WMS text ■

10.8 Inequalities for Probabililities

This chapter lists some very useful estimates for probabilities which involve the moments of a random variable. Among them is the Tchebysheff inequality.

Theorem 10.18. ⋆

Let Y, Z be continuous or discrete random variables and $a > 0$. Assume further that $Y \geq 0$. Then

$$P\{Y \geq a\}) \leq \frac{E[Y]}{a}, \tag{10.51}$$

$$P\{|Z| \geq a\}) \leq \frac{E[\,|Z|^n]}{a^n}. \tag{10.52}$$

(10.51) *is known as the* ***Markov inequality***

PROOF of (10.51): [100] We give the proof for continuous random variables. The discrete case is even simpler since it involves summation instead of integration.

Let $f_Y(y)$ be the PDF of Y. We observe the following:

(a) $Y \geq 0$ implies $y\, f_Y(y) = 0$ for $-\infty < y < 0$.
(b) $y\, f_Y(y) \geq 0$ for $0 \leq y < \infty$.
(c) $y\, f_Y(y) \geq a\, f_Y(y)$ for $a \leq y < \infty$.

Thus,

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} y\, f_Y(y)dy \overset{\text{(a)}}{=} \int_0^{\infty} y\, f_Y(y)dy = \int_0^a y\, f_Y(y)dy + \int_a^{\infty} y\, f_Y(y)dy \\ &\overset{\text{(b)}}{\geq} \int_a^{\infty} y\, f_Y(y)dy \overset{\text{(c)}}{\geq} \int_a^{\infty} a\, f_Y(y)dy = a\int_a^{\infty} f_Y(y)dy = a\, P\{Y \geq a\}. \end{aligned}$$

We divide by $a > 0$ and obtain (10.51).

PROOF of (10.52): Since $|Z|^n \geq 0$ and $a^n > 0$, we can apply (10.51) with $|Z|^n$ in place of Y and a^n in place of a:

$$P\{|Z|^n \geq a^n\} \leq \frac{E[|Z|^n]}{a^n}. \tag{A}$$

Since the function $x \mapsto x^n$ is (strictly) increasing, $|Z(\omega)|^n \geq a^n \Leftrightarrow |Z(\omega)| \geq a$.

Thus, **(A)** yields $P\{|Z| \geq a\} \leq E[|Z|^n]/a^n$ and this proves (10.52). ■

The work we have done here allows us to quickly prove the Tchebysheff inequalities in the form listed in WMS Chapter 4.10 (Tchebysheff's Theorem).

Theorem 10.19 (Tchebysheff Inequalities)**.**

Let Y be a random variable with mean $\mu = E[Y]$ and standard deviation $\sigma = \sqrt{Var[Y]}$. Let $k > 0$. Then

$$P\{|Y - \mu| \geq k\sigma\} \leq \frac{1}{k^2}, \tag{10.53}$$

$$P\{|Y - \mu| < k\sigma\} \geq 1 - \frac{1}{k^2}. \tag{10.54}$$

Both (10.53) *and* (10.54) *are known as the* ***Tchebysheff inequalities***

[100] Source: https://en.wikipedia.org/wiki/Markov%27s_inequality

PROOF: We apply (10.52) with $n = 2$, $Y - \mu$ in place of Z, and $k\sigma$ in place of a. We obtain

$$P\{|Y - \mu| \geq k\sigma\}) \;\leq\; \frac{E[\,|Y-\mu|^2]}{(k\sigma)^2} \;=\; \frac{E[\,(Y-\mu)^2]}{(k\sigma)^2} \;=\; \frac{\sigma^2}{k^2\sigma^2} \;=\; \frac{1}{k^2}\,.$$

This proves (10.53). Since the event $\{|Y-\mu| < k\sigma\}$ is the complement of the event $\{|Y-\mu| \geq k\sigma\}$, (10.54) follows. ■

Remark 10.13. Some comments about the Tchebysheff inequalities:

(a) Both inequalities state the same, since the events $\{|Y - \mu| < c\sigma\}$ and $\{|Y - \mu| \geq c\sigma\}$ are complements of each other. We had noted this in the proof of Theorem 10.19.

(b) The inequalities are not particularly powerful, but consider that they are universally valid, regardless of any particulars concerning Y!

(c) If we write $a := k\sigma$ and thus, $k = a/\sigma$, we obtain

$$P\{|Y-\mu| < a\} \;\geq\; 1 - \frac{Var[Y]}{a^2} \qquad \text{and} \qquad P\{|Y-\mu| \geq a\} \;\leq\; \frac{Var[Y]}{a^2}\,. \;\square$$

Example 10.3. ACME Co. produces screws. Their lengths follow a distribution with a mean of $\mu = 18.40$ mm and a variance of $\sigma^2 = 0.64$ mm^2. In other words, the length Y of a randomly picked screw (a sample of size 1) has $E[Y] = 18.40$ and $Var[Y] = 0.64$.

A screw can only be sold if its length is within 17.20 and 19.60 mm. How likely is it that a screw is produced that cannot be sold?

Solution: We observe that $E[Y] = 18.40$ is the midpoint of the interval $[17.20, 19.60]$ and that

- a screw cannot be sold $\Leftrightarrow\; Y(\omega) \notin [17.20, 19.60] \;\Leftrightarrow\; |Y(\omega) - E[Y]| > (17.20, 19.60)/2 = 1.2.$

We solve

$$k\sigma \;=\; |Y - E[Y]| \;=\; 1.2\,, \quad \text{i.e.,} \quad \sqrt{0.64}k \;=\; 0.8k \;=\; 1.2\,,$$

for k and obtain $k = 1.2/0.8 = 3/2$. Thus, $k^2 = 9/4$.

Tchebysheff's inequality (10.54) then yields the following upper bound for the probability of obtaining a sample with a difference $\bar{Y}(\omega) - \mu$ as large as or even larger than the one we have sampled:

$$P\{|Y - \mu| > k\sigma\} \;\leq\; \frac{1}{k^2} \;=\; 4/9\,.$$

This example demonstrates the low quality of the bounds that we obtain from Tchebysheff's inequalities. For example, let us assume we know that Y follows a normal distribution, i.e.,

$$Y \;\sim\; \mathcal{N}(\mu = 18.40, \sigma^2 = 0.64)\,,$$

then we can deduce from the empirical rule (the 68%–95%–99.7% rule) [101] that

$$\begin{aligned} 0.32 \;=\; 1 - 0.68 \;&\approx P\{|Y - \mu| > 1 \cdot \sigma\} \\ &\geq P\{|Y - \mu| > 1.5\sigma\} \\ &\geq P\{|Y - \mu| > 2\sigma\} \;\approx\; 1 - 0.95 \;=\; 0.05\,. \end{aligned}$$

[101] see the introcduction to subch.10.5: The Normal Probability Distribution

Thus, higher precision calculations show that the more likely event of $Y(\omega)$ not being within one standard deviation of 18.40 mm only has a probability of 0.32, substantially less than our overly generous estimate of $4/9 = 0.44\bar{4}$ for the less likely event of being within 1.5 standard deviations. By the way, the exact figure (in the case of $Y \sim \mathcal{N}(18.40, 0.64)$) is $P\{|Y - \mu| > 1.5\sigma\} \approx 0.1336$. This is less than one third of the Tchebysheff estimate. □

Example 10.4. It has been established some time ago that the data in the population of interest follow a distribution with a mean of $\mu = 18.40$. In other words, a random pick Y (a sample of size 1) from that population has $E[Y] = 18.40$. There have been concerns that the composition of the population has changed significantly and μ with it. An SRS (simple random sample) is drawn from that population and mean and variance are estimated from the realization of this sample as

$$\bar{Y}(\omega) = 17.60 \qquad \text{and} \qquad S^2(\omega) = 6.25\,.^{102}$$

Is the deviation of $\bar{Y}(\omega)$ from μ big enough to discard $\mu = 18.40$ and go through the process of establishing a new population mean?

Solution: We use $S^2 = 6.25$ for $\sigma^2 = Var[Y]$. Then $\sigma = \sqrt{6.25} = 2.5$. We solve

$$k\sigma = |\bar{Y} - E[Y]|\,, \qquad \text{i.e.,} \quad 0.25k = |17.60 - 18.40| = 0.8\,,$$

for k and obtain $k = 3.2$. Thus, $k^2 = 10.24$. Since $E[\bar{Y}] = E[Y]$ it follows from Tchebysheff's inequality (10.54) that the probability of obtaining a sample with a difference $\bar{Y}(\omega) - E[Y]$ as large as or even larger than the one of the sample we have drawn, is

$$P\{|Y - \mu| < k\sigma\} \geq 1 - \frac{1}{k^2} = 1 - \frac{1}{10.24} = 0.902344\,.$$

This probability is very large and shows that our sample mean $\bar{Y} = 17.60$ does not contradict the assumption that the population mean 18.40 □

[102] $\bar{Y} = 17.60$ is the so called sample mean (see Example 11.5: Variance of the sample mean on p.262) and $S^2 = S^2(\omega) = \frac{1}{n-1}\left(\sum_{j=1}^{n} \left(Y_j(\omega) - \bar{Y}(\omega)\right)^2\right)$ is the so called sample variance which will be introduced in subchapter 13.3 (Sampling Distributions) of Chapter 13(Limit Theorems). See Definition 13.4: Sample variance on p.325.

11 Multivariate Probability Distributions

Like the previous chapter, this one is extremely skeletal in nature. It contains very few examples. You are reminded again that you must work through the corresponding chapters in the WMS text. In this case, that would be WMS Chapter 5 (Multivariate Probability Distributions).

11.1 Multivariate CDFs, PMFs and PDFs

Assumption 11.1 (Comma separation denotes intersection)**.** We will follow the following convention for the notation of events that are generated by random variables or random elements $X, Y, Z \ldots$

Separating commas are to be interpreted as "**and**" and not as "**or**". Thus, for example,

$$\begin{aligned}\{X \in B,\, Y = \alpha,\, 5 \le Z < 8\} &= \{X \in B \text{ **and** } Y = \alpha \text{ **and** } 5 \le Z < 8\} \\ &= \{X \in B\} \cap \{Y = \alpha\} \cap \{5 \le Z < 8\}\,. \ \square\end{aligned}$$

Definition 11.1 (Joint cumulative distribution function)**.**

Given are two random variables Y_1 and Y_2. No assumption is made whether they are discrete or continuous. We call

$$F(y_1, y_2) \ := F_{Y_1,Y_2}(y_1, y_2) \ := \ P(Y_1 \le y_1, Y_2 \le y_2)\,, \qquad \text{where } y_1, y_2 \in \mathbb{R}\,, \tag{11.1}$$

the **joint cumulative distribution function** or **bivariate cumulative distribution function** or **joint CDF** or **joint distribution function** of Y_1 and Y_2. $\square$

Theorem 11.1.

Let Y_1 and Y_2 be random variables with joint CDF $F_{Y_1,Y_2}(y_1, y_2)$. Further, assume that $\vec{a} := (a_1, a_2) \in \mathbb{R}^2$ and $\vec{b} := (b_1, b_2) \in \mathbb{R}^2$ satisfy $a_1 < b_1$ and $a_2 < b_2$. Then,

$$F_{Y_1,Y_2}(-\infty, -\infty) \ = \ F_{Y_1,Y_2}(-\infty, y_2) \ = \ F_{Y_1,Y_2}(y_1, -\infty) \ = \ 0\,. \tag{11.2}$$

$$F_{Y_1,Y_2}(\infty, \infty) \ = 1\,, \tag{11.3}$$

$$\begin{aligned}P\{a_1 < Y_1 \le b_1,\, a_2 < Y_2 \le b_2\} \ &= F_{Y_1,Y_2}(b_1, b_2) \ - \ F_{Y_1,Y_2}(a_1, b_2) \\ &\quad - \ F_{Y_1,Y_2}(b_1, a_2) \ + \ F_{Y_1,Y_2}(a_1, a_2)\,,\end{aligned} \tag{11.4}$$

$$F_{Y_1,Y_2}(b_1, b_2) \ - \ F_{Y_1,Y_2}(a_1, b_2) \ - \ F_{Y_1,Y_2}(b_1, a_2) \ + \ F_{Y_1,Y_2}(a_1, a_2) \ \ge 0\ , \tag{11.5}$$

PROOF:

(11.2) follows from

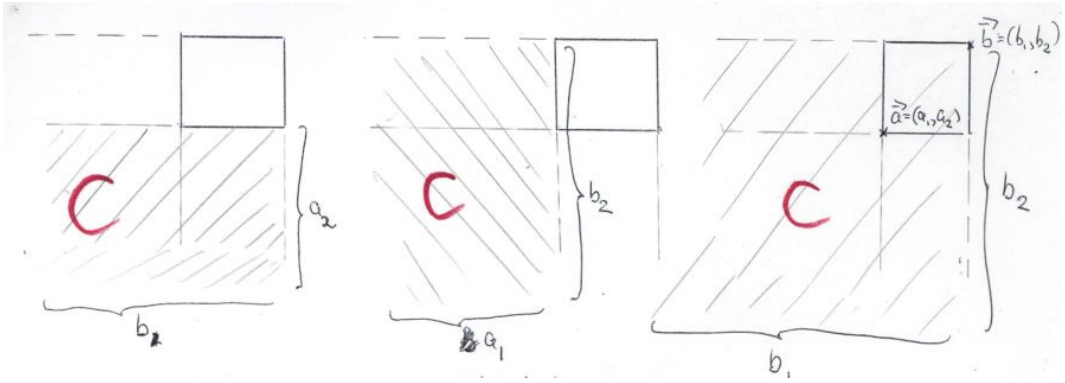
$$P\{Y_1 < -\infty\} \ = \ P\{Y_2 < -\infty\} \ = \ 0\,.$$

(11.3) follows from

$$P\{Y_1 < \infty,\, Y_2 < \infty\} \;=\; P(\Omega) \;=\; 1\,.$$

(11.5) is immediate from (11.4).

Finally, for the proof of (11.4), we see from the three pictures below the following:



- $P\{a_1 < Y_1 \le b_1,\, a_2 < Y_2 \le b_2\} \mathrel{\widehat{=}}$ black rectangle in the upper right corner
- $F_{Y_1,Y_2}(b_1,b_2) \mathrel{\widehat{=}}$ shaded area in the right drawing
- $F_{Y_1,Y_2}(b_1,a_2) \mathrel{\widehat{=}}$ shaded area (below black rectangle) in the left drawing
- $F_{Y_1,Y_2}(b_1,a_2) \mathrel{\widehat{=}}$ shaded area (to left of black rectangle) in the middle drawing
- $F_{Y_1,Y_2}(a_1,a_2) \mathrel{\widehat{=}}$ area marked with a red **C**

The expression $F_{Y_1,Y_2}(b_1,b_2) - F_{Y_1,Y_2}(a_1,b_2) - F_{Y_1,Y_2}(b_1,a_2)$ would correspond to the black rectangle, except that we subtracted the red **C** area twice. We add $F_{Y_1,Y_2}(a_1,a_2)$ to compensate. ■

Definition 11.2 (Joint probability mass function)**.**

Let Y_1 and Y_2 be discrete random variables. We call

(11.6) $$p(y_1,y_2) \;:= p_{Y_1,Y_2}(y_1,y_2) \;:=\; P\{Y_1 = y_1, Y_2 = y_2\}\,, \qquad \text{where } y_1, y_2 \in \mathbb{R}\,,$$

the **joint probability mass function** or **bivariate probability mass function** or **joint PMF** of Y_1 and Y_2. □

Just as in the univariate case, $p_{Y_1,Y_2}(y_1,y_2)$ assigns nonzero probabilities to only finitely or countably many pairs of values (y_1,y_2). As in the univariate case, by definition,

$$\sum_{(y_1,y_2)\in B} p_{Y_1,Y_2}(y_1,y_2) \;=\; \sum_{\substack{(y_1,y_2)\in B,\\ p_{Y_1,Y_2}(y_1,y_2)>0}} p_{Y_1,Y_2}(y_1,y_2)\,.$$

Proposition 11.1 (WMS Ch.05.2, Theorem 5.1)**.**

If Y_1 and Y_2 are discrete random variables with joint PMF $p_{Y_1,Y_2}(y_1,y_2)$, then

(1) $p_{Y_1,Y_2}(y_1,y_2) \ge 0$ *for all* $y_1, y_2 \in \mathbb{R}$,

(2) $\sum\limits_{y_1,y_2} p_{Y_1,Y_2}(y_1,y_2) \;=\; 1.$

(3) $$F_{Y_1,Y_2}(y_1,y_2) \;=\; \sum_{u_1\le y_1,\, u_2\le y_2} p_{Y_1,Y_2}(u_1,u_2) \;=\; \sum_{u_1\le y_1} \sum_{u_2\le y_2} p_{Y_1,Y_2}(u_1,u_2)\,.$$

PROOF: Obvious. ■

Definition 11.3 (Jointly continuous random variables)**.**

Let Y_1 and Y_2 be random variables with joint CDF $F(y_1, y_2)$. We call Y_1 and Y_2 **jointly continuous** if $F(y_1, y_2)$ is a continuous function of both arguments. □

Assumption 11.2 (Jointly continuous random variables have PDFs)**.** We will follow the following convention for the notation of events that are generated by random variables or random elements $X, Y, Z \ldots$

We assume for all jointly continuous random variables Y_1 and Y_2 that $\dfrac{\partial^2 F_{Y_1,Y_2}}{\partial y_1 \partial y_2}$ exists and is continuous except for $(y_1, y_2) \in B^*$, where the set $B^* \subseteq \mathbb{R}^2$ satisfies that $B^* \cap B$ is finite for any bounded subset $B \in \mathbb{R}^2$ (bounded sets are those contained in a circle with sufficiently large radius).

This assumption guarantees for all $y_1, y_2 \in \mathbb{R}$, when we write f_{Y_1,Y_2} for $\dfrac{\partial^2 F_{Y_1,Y_2}}{\partial y_1 \partial y_2}$, that

$$
\begin{aligned}
F_{Y_1,Y_2}(y_1, y_2) &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{Y_1,Y_2}(u_1, u_2)\, du_2\, du_1 \\
&= \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} f_{Y_1,Y_2}(u_1, u_2)\, du_1\, du_2\,. \qquad (11.7) \\
&= \iint\limits_{]-\infty, y_1 \times]-\infty, y_2]} f_{Y_1,Y_2}(u_1, u_2)\, du_1\, du_2\,. \quad \square
\end{aligned}
$$

Definition 11.4 (WMS Ch.05.2, Definition 5.3)**.**

Let Y_1 and Y_2 be continuous random variables with joint distribution function $F(y_1, y_2)$ and second derivative $f_{Y_1,Y_2}(y_1, y_2) = \dfrac{\partial^2 F_{Y_1,Y_2}}{\partial y_1 \partial y_2}(y_1, y_2)$. We call $f_{Y_1,Y_2}(y_1, y_2)$ the **joint probability density function** or **joint PDF** of Y_1 and Y_2. □

Theorem 11.2.

Let Y_1 and Y_2 be jointly continuous random variables with joint PDF $f_{Y_1,Y_2}(y_1, y_2)$, then

(1) $f_{Y_1,Y_2}(y_1, y_2) \geq 0$ *for all* y_1, y_2.

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 = 1$.

PROOF: An easy consequence of Theorem 11.1 on p.241. ■

11.2 Marginal and Conditional Probability Distributions

Definition 11.5 (Marginal distribution of two random variables)**.**

Let $\vec{Y} = (Y_1, Y_2)$ be a vector of two random variables with joint distribution

$$(B_1, B_2) \mapsto P_{Y_1,Y_2}(B_1, B_2) = P\{Y_1 \in B_1, Y_2 \in B_2\}, \quad \text{where } B_1, B_2 \subseteq \mathbb{R}.$$

We call the probability measures

$$Q_1 : B_1 \mapsto P_{Y_1,Y_2}(B_1, \mathbb{R}) \quad \text{and} \quad Q_2 : B_2 \mapsto P_{Y_1,Y_2}(\mathbb{R}, B_2) \tag{11.8}$$

the **marginal distributions** of $\vec{Y} = (Y_1, Y_2)$. □

Proposition 11.2.

The marginal distributions of $\vec{Y} = (Y_1, Y_2)$ are the distributions P_{Y_1} and P_{Y_2} of the coordinates Y_1 and Y_2. In other words, $Q_1 = P_{Y_1}$ and $Q_2 = P_{Y_2}$

PROOF: Since, $Y_1(\omega) \in B \Leftrightarrow Y_1(\omega) \in B$ and $Y_2(\omega) \in \mathbb{R}$ holds for all $B \subseteq \mathbb{R}$, we obtain

$$Q_1(B) = P_{Y_1,Y_2}(B, \mathbb{R}) = P\{Y_1 \in B, Y_2 \in \mathbb{R}\} = P\{Y_1 \in B\} = P_{Y_1}(B), \quad \text{whenever } B \subseteq \mathbb{R}.$$

Thus, $Q_1 = P_{Y_1}$. We obtain in a similar fashion from $Y_2(\omega) \in B \Leftrightarrow Y_1(\omega) \in \mathbb{R}$ and $Y_2(\omega) \in B$, that

$$Q_2(B) = P_{Y_2}(B), \quad \text{for all } B \subseteq \mathbb{R}. \blacksquare$$

Henceforth, we will retire the symbols Q_1, Q_2 and denote the marginal distributions of $\vec{Y} = (Y_1, Y_2)$ by P_{Y_1} and P_{Y_2}.

Definition 11.5 translates for discrete random variables, whose distribution is determined by their joint PMF and for continuous random variables, whose distribution is determined by their joint PDF, to the following.

Definition 11.6 (Marginal PMF and PDF)**.**

(a) Let Y_1 and Y_2 be discrete random variables with joint PMF $p_{Y_1,Y_2}(y_1, y_2)$. We call

$$p_{Y_1}(y_1) = \sum_{\text{all } y_2} p_{Y_1,Y_2}(y_1, y_2) \quad \text{and} \quad p_{Y_2}(y_2) = \sum_{\text{all } y_1} p_{Y_1,Y_2}(y_1, y_2) \tag{11.9}$$

the **marginal probability mass functions** or **marginal PMFs** of Y_1 and Y_2.

(b) Let Y_1 and Y_2 be continuous random variables with joint PDF $f_{Y_1,Y_2}(y_1, y_2)$. We call

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \quad \text{and} \quad f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1\,. \tag{11.10}$$

the **marginal density functions** or **marginal PDFs** of Y_1 and Y_2. □

Remark 11.1. We recall Definition 5.7 of $P(A \mid B)$, the probability of the event A conditioned on the event B, which is defined for $P(B) > 0$ as

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

We also recall that, if $P(B) > 0$, the set function $A \mapsto P(A \mid B)$ is a probability measure on Ω. See Theorem 5.8 on p.118. We replace the general events A and B with events $\{Y_1 = y_1\}$ and $\{Y_2 = y_2\}$ and obtain, if $P\{Y_2 = y_2\} > 0$,

$$P\{Y_1 = y_1 \mid Y_2 = y_2\} = \frac{P\{Y_1 = y_1,\, Y_2 = y_2\}}{P\{Y_2 = y_2\}}\,. \tag{11.11}$$

As we always do for conditional probabilities, we interpret (11.11) as the probability that the random variable Y_1 equals y_1, given that Y_2 equals y_2.

Not much can be done with formula (11.11) for continuous random variables Y_1 and Y_2, because $P\{Y_2 = y_2\} = 0$ for all $y_2 \in \mathbb{R}$; but it shows us how to define conditional PMFs for discrete random variables. □

Definition 11.7 (Conditional probability mass function)**.**

Let Y_1 and Y_2 be discrete random variables with joint PMF $p_{Y_1,Y_2}(y_1, y_2)$ and marginal PMFs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$. Then we call

$$p_{Y_1|Y_2}(y_1 \mid y_2) := \begin{cases} P\{Y_1 = y_1 \mid Y_2 = y_2\}\,, & \text{if } P\{Y_2 = y_2\} > 0\,, \\ \textbf{undefined}\,, & \text{if } P\{Y_2 = y_2\} = 0\,, \end{cases} \tag{11.12}$$

the **conditional probability mass function** or the **conditional PMF** of Y_1 given Y_2.

Likewise, we call

$$p_{Y_2|Y_1}(y_2 \mid y_1) := \begin{cases} P\{Y_2 = y_2 \mid Y_1 = y_1\}\,, & \text{if } P\{Y_1 = y_1\} > 0\,, \\ \textbf{undefined}\,, & \text{if } P\{Y_1 = y_1\} = 0\,, \end{cases} \tag{11.13}$$

the **conditional PMF** of Y_2 given Y_1. □

Remark 11.2. Note that conditional PMFs can be expressed in terms of joint PMF and marginal PMFs:

$$p_{Y_1|Y_2}(y_1 \mid y_2) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)} \quad \text{if } p_{Y_2}(y_2) > 0\,, \tag{11.14}$$

$$p_{Y_2|Y_1}(y_2 \mid y_1) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)} \quad \text{if } p_{Y_1}(y_1) > 0\,. \tag{11.15}$$ □

We had mentioned in Remark 11.1 on p.244 that we must find an alternative to the formula

$$P\{Y_1 = y_1 \mid Y_2 = y_2\} = \frac{P\{Y_1 = y_1,\, Y_2 = y_2\}}{P\{Y_2 = y_2\}}\,,$$

used for discrete random variables conditioning, when conditioning a continous random variable on another continuous random variable. And yet, the discrete case formulas. (11.14) and (11.14) will guide us in creating the appropriate definitions.

From a modeling perspective, when one is concerned with expressing reality in mathematical terms, the next two definition have proven very useful.

Definition 11.8 (Conditional probability density function)**.**

Let Y_1 and Y_2 be continuous random variables with joint PDF $f_{Y_1|Y_2}(y_1, y_2)$ and marginal densities $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$. Then we call

$$f_{Y_1|Y_2}(y_1 \mid y_2) := \begin{cases} \dfrac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)}, & \text{if } f_{Y_2}(y_2) > 0, \\ \textbf{undefined}, & \text{if } f_{Y_2}(y_2) = 0, \end{cases} \tag{11.16}$$

the **conditional probability density function** or the **conditional PDF** of Y_1 given Y_2.

Likewise, we call

$$f_{Y_2|Y_1}(y_2 \mid y_1) := \begin{cases} \dfrac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)}, & \text{if } f_{Y_1}(y_1) > 0, \\ \textbf{undefined}, & \text{if } f_{Y_1}(y_1) = 0, \end{cases} \tag{11.17}$$

the **conditional PDF** of Y_2 given Y_1. □

Definition 11.9. ⋆

Let Y_1 and Y_2 be two jointly continuous random variables. Then,

$$F_{Y_1|Y_2}(y_1 \mid y_2) := \int_{-\infty}^{y_1} \frac{f_{Y_1,Y_2}(u_1,y_2)}{f_{Y_2}(y_2)}\, du_1 \tag{11.18}$$

defines the **conditional distribution function** or **conditional CDF** of Y_1 given $Y_2 = y_2$. □

Remark 11.3. We interpret $f_{Y_1|Y_2}(y_1 \mid y_2)$ and $F_{Y_1|Y_2}(y_1 \mid y_2)$ as follows:

$$f_{Y_1|Y_2}(y_1 \mid y_2) \approx \frac{P\{y_1 < Y_1 \le y_1 + \delta,\, y_2 < Y_2 \le y_2 + \delta\}}{P\{y_2 < Y_2 \le y_2 + \delta\}},$$
$$F_{Y_1|Y_2}(y_1 \mid y_2) \approx P\{Y_1 \le y_1 \mid y_2 < Y_2 \le y_2 + \delta\},$$

for very small $\delta > 0$. As $\delta \to 0$, the error of approximation becomes smaller and smaller. □

11.3 Independence of Random Variables and Discrete Random Elements

Introduction 11.1. Let $X_1, X_2 : (\Omega, P) \to \Omega'$ be two random elements (recall that they are called random variables only if $\Omega' \subseteq \mathbb{R}$). Not all events $A \subseteq \Omega$ are meaningful for X_1 and X_2. Rather, only **events generated by** X_1 and by X_2, i.e., events of the form $\{X_1 \in B_1\}$ and $\{X_2 \in B_2\}$ for suitable $B_1, B_2 \subseteq \Omega'$ will matter.

Since independence of two events A_1 and A_2 is defined by $P(A_1 \cap A_2) = P(A_1)P(A_2)$, the proper way to define independence of X_1 and X_2 seems to be

$$P\{X_1 \in B_1,\, X_2 \in B_2,\} = P\{X_1 \in B_1\} \cdot P\{\, X_2 \in B_2,\} \quad \text{for all relevant } B_1, B_2 \subseteq \Omega'. \tag{11.19}$$

What are the relevant sets B_j? We answer that question for discrete random elements (hence, also for discrete random variables) and for continuous random variables.

(a) Assume that $X : (\Omega, P) \to \Omega'$ is a discrete random element with PMF $p_X(x)$. In other words, there is a countable $\Omega^* \subseteq \Omega'$ such that, for any $B \subseteq \Omega'$,

$$P\{X \in B\} \;=\; P_X(B) \;=\; \sum_{x \in \Omega^* \cap B} p_X(x) \;=\; \sum_{x \in B} p_X(x) \;=\; \sum_{x \in B} P\{X = x\}\,.$$

These equations show that the distribution of X is determined by the events $\{X = x\}$. Thus, the relevant sets for X are of the form $B = \{x\}$.

(b) Assume that Y is a continuous random variable on (Ω, P) with PDF $f_Y(y)$. Then the probabilities for the events that matter, the events $\{a < Y \le b\}$ where $a < b$, are

$$P\{a < Y \le b\} \;=\; \int_a^b f_Y(y)dy\,.$$

(See (10.4) in heorem 10.2 on p.214.) This equation shows that the distribution of Y is determined by the probability density function $f_Y(y)$. Thus, the relevant sets for Y are the intervals $B =]a, b]$. [103]

In summary, we could define independence of discrete random elements X_1 and X_2 as

$$P\{X_1 = x_1,\, X_2 = x_2,\} \;=\; P\{X_1 = x_1\} \cdot P\{\, X_2 = x_2,\} \quad \text{for all } x_1, x_2 \in \Omega'.$$

Equivalently, this can be expressed as

$$p_{X_1,X_2}(x_1, x_2) \;=\; p_{X_1}(x_1) \cdot p_{X_2}(x_2) \qquad \text{for all } x_1, x_2 \in \Omega'. \tag{11.20}$$

Moreover, independence of continuous random variables Y_1 and Y_2 could be defined as

$$P\{a < X_1 \le b,\, c < X_2 \le d\} \;=\; P\{a < X_1 \le b\} \cdot P\{c < X_2 \le d\} \quad \text{for all } a < b \text{ and } c < d.$$

Equivalently, this can be expressed as

$$\int_a^b \int_c^d f_{Y_1,Y_2}(y_1, y_2)dy_2\, dy_1 \;=\; \int_a^b f_{Y_1}(y_1)dy_1 \cdot \int_c^d f_{Y_2}(y_2)dy_2 \quad \text{for all } a < b \text{ and } c < d. \tag{11.21}$$

The CDF (cumulative distribution function) $F_Y(y)$ gives us for both discrete and continuous random variables (but we must exclude discrete random elements) a unified way to express what was stated in **(a)** and **(b)** as follows.

In the discrete case **(a)** we have

$$P\{Y = y\} \;=\; P\{Y \le y\} \;-\; P\{Y < y\} \;=\; F_Y(y) - F_Y(y-)\,.$$

Here $F_Y(y-) = \lim_{a<y, a\to y} F_Y(a)$ is the left–hand limit of the (monotone) function $F_Y(\cdot)$ at y.

In the continuous case **(b)** we have

$$P\{a < Y \le b\} \;=\; P\{Y \le b\} \;-\; P\{Y \le a\} \;=\; F_Y(b) - F_Y(a)\,.$$

In both cases, independence of Y_1 and Y_2 can now be defined as

$$F_{Y_1,Y_2}(y_1, y_2) \;=\; F_{Y_1}(y_1) \cdot F_{Y_2}(y_2) \quad \text{for all } y_1, y_2 \in \mathbb{R}. \;\square \tag{11.22}$$

[103] Since $P\{X = a\} = 0$ for all $a \in \mathbb{R}$, it does not matter whether we do or do not include the end points. See Proposition 10.1 on p.213.

We make (11.22) the basis for the definition of independence of random variables.

Definition 11.10 (Independent random variables)**.**

Let Y_1 and Y_2 be random variables with CDFs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and with joint CDF $F_{Y_1,Y_2}(y_1, y_2)$. We call Y_1 and Y_2 **independent** if

$$F_{Y_1,Y_2}(y_1, y_2) \;=\; F_{Y_1}(y_1) \cdot F_{Y_2}(y_2) \quad \text{for all } y_1, y_2 \in \mathbb{R}. \tag{11.23}$$

If Y_1 and Y_2 are not independent, we call them **dependent**.

We must treat discrete random elements separately since there are no CDFs.

Let X_1 and X_2 be discrete random elements with PMFs $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ and with joint PMF $p_{X_1,X_2}(x_1, x_2)$. We call X_1 and X_2 **independent** if

$$p_{X_1,X_2}(x_1, x_2) \;=\; p_{X_1}(x_1) \cdot p_{X_2}(x_2) \quad \text{for all } x_1, x_2 \in \mathbb{R}. \tag{11.24}$$

If X_1 and X_2 are not independent, we call them **dependent**. □

Theorem 11.3 (Functions of independent random variables are independent)**.**

Let $\vec{Y} = (Y_1, \ldots, Y_k) : (\Omega, P) \to \mathbb{R}$ be a vector of k independent random variables and $h_j : \mathbb{R} \to \mathbb{R}$.

- *Then the random variables $h_1 \circ Y_1, \ldots, h_k \circ Y_k$ also are independent.*

PROOF: We recall (2.50) of Proposition 2.8 (Preimages of function composition) on p.45: Let $f : X \to Y$ and $g : Y \to Z$ and $W \subseteq Z$. Then

$$\textbf{(A)} \qquad (g \circ f)^{-1} \;= f^{-1} \circ g^{-1}, \text{ i.e., } \quad (g \circ f)^{-1}(W) \;=\; f^{-1}\big(g^{-1}(W)\big)\,. \tag{A}$$

We use this twice in the following calculations.

$$\begin{aligned} P\{h_j \circ Y_j \in B_j, (j = 1, \ldots, n)\} \;&= P\{(h_j \circ Y_j)^{-1}(B_j), (j = 1, \ldots, n)\} \\ &\overset{\textbf{(A)}}{=} P\{Y_j^{-1} \circ h_j^{-1}(B_j), (j = 1, \ldots, n)\} \;=\; P\{Y_j \in h_j^{-1}(B_j), (j = 1, \ldots, n)\}\,. \end{aligned}$$

Since the Y_j are independent, the product rule holds. We obtain

$$\begin{aligned} P\{h_j \circ Y_j \in B_j, (j = 1, \ldots, n)\} \;&= \prod_j P\{Y_j \in h_j^{-1}(B_j)\} \;=\; \prod_j P\{Y_j^{-1} \circ h_j^{-1}(B_j)\} \\ &\overset{\textbf{(A)}}{=} \prod_j P\{\prod_j P\{(h_j \circ Y_j)^{-1}(B_j)\} \;=\; \prod_j P\{h_j \circ Y_j \in B_j\}\,. \quad \blacksquare \end{aligned}$$

Theorem 11.4 (WMS Ch.05.4, Theorem 5.4)**.**

If Y_1 and Y_2 are discrete random variables with joint PMF $p_{Y_1,Y_2}(y_1, y_2)$ and marginal PMFs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$, then

$$Y_1, Y_2 \text{ are independent} \quad \Leftrightarrow \quad p_{Y_1,Y_2}(y_1, y_2) \;=\; p_{Y_1}(y_1) \cdot p_{Y_2}(y_2) \text{ for all } y_1, y_2 \in \mathbb{R}. \tag{11.25}$$

If Y_1 and Y_2 are continuous random variables with joint PDF $f_{Y_1,Y_2}(y_1,y_2)$ and marginal PDFs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, then

(11.26) *Y_1, Y_2 are independent* $\Leftrightarrow$ $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)\cdot f_{Y_2}(y_2)$ *for all* $y_1, y_2 \in \mathbb{R}$.

PROOF: We only prove here the $\Rightarrow$ directions of (11.25) and (11.26). The proof of the opposite direction is left as an excercise to the reader.

We apply (11.4) of Theorem 11.1 on p.241 and (11.23) of Definition 11.10 (Independent random variables) on p.248 as follows.

$$\begin{aligned} &P\{a_1 < Y_1 \le y_1,\ a_2 < Y_2 \le y_2\} \\ &\overset{(12.18)}{=} F_{Y_1,Y_2}(y_1,y_2) - F_{Y_1,Y_2}(a_1,y_2) - F_{Y_1,Y_2}(y_1,a_2) + F_{Y_1,Y_2}(a_1,a_2) \\ &\overset{(11.23)}{=} F_{Y_1}(y_1)F_{Y_2}(y_2) - F_{Y_1}(a_1)F_{Y_2}(y_2) - F_{Y_1}(y_1)F_{Y_2}(a_2) + F_{Y_1}(a_1)F_{Y_2}(a_2) \\ \textbf{(A)}\qquad &= \big(F_{Y_1}(y_1) - F_{Y_1}(a_1)\big)\big(F_{Y_2}(y_2) - F_{Y_2}(a_2)\big) = P\{a_1 < Y_1 \le y_1\}\cdot P\{a_2 < Y_2 \le y_2\} \end{aligned}$$

For discrete Y_1 and Y_2, we obtain with $a_1 = y_1-$ and $a_2 = y_2-$,

$$\begin{aligned} p_{Y_1,Y_2}(y_1,y_2) &= P\{y_1- < Y_1 \le y_1,\ y_2- < Y_2 \le y_2\} \\ &\overset{\textbf{(A)}}{=} P\{y_1- < Y_1 \le y_1\}\cdot P\{y_2- < Y_2 \le y_2\} = p_{Y_1}(y_1)\cdot p_{Y_2}(y_2)\,. \end{aligned}$$

For continuous Y_1 and Y_2, we obtain,

$$\begin{aligned} \int_{a_1}^{y_1}\int_{a_2}^{y_2} f_{Y_1,Y_2}(u_1,u_2)\,du_1du_2 &= P\{a_1 < Y_1 \le y_1,\ a_2 < Y_2 \le y_2\} \\ &\overset{\textbf{(A)}}{=} P\{a_1 < Y_1 \le y_1\}\cdot P\{a_2 < Y_2 \le y_2\} = \int_{a_1}^{y_1} f_{Y_1}(u_1)\,du_1\cdot\int_{a_2}^{y_2} f_{Y_2}(u_2)\,du_2 \end{aligned}$$

We differentiate with respect to y_1 and y_2 and obtain $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)\,f_{Y_2}(y_2)$. ■

The next theorem will be generalized in Theorem 11.10 on p.256. There Y_1 and Y_2 will be replaced with functions $g(Y_1)$ and (Y_2).

Theorem 11.5.

If Y_1 and Y_2 are independent random variables, then

(11.27) $$E[Y_1\cdot Y_2] = E[Y_1]\cdot E[Y_2]\,.$$

PROOF: We show the proof for continuous Y_1 and Y_2. Since $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)\cdot f_{Y_2}(y_2)$,

$$\begin{aligned} E[Y_1Y_2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2 f_{Y_1,Y_2}(y_1,y_2)\,dy_1\,dy_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2 f_{Y_1}(y_1)\,f_{Y_2}(y_2)\,dy_1\,dy_2 \\ &= \int_{-\infty}^{\infty} y_2\left[\int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1)\,dy_1\right] f_{Y_2}(y_2)\,dy_2 = \int_{-\infty}^{\infty} y_2E[Y_1]f_{Y_2}(y_2)\,dy_2 \\ &= E[Y_1]\int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2)\,dy_2 = E[Y_1]\,E[Y_2]\,. \end{aligned}$$

The proof for discrete random variables is obtained by employing $p_{Y_1,Y_2}(y_1,y_2) = p_{Y_1}(y_1)\cdot p_{Y_2}(y_2)$ and replacing integration with summation. ■

Theorem 11.6 (WMS Ch.05.4, Theorem 5.5)**.**

Let the continuous random variables Y_1 and Y_2 have a joint PDF $f_{Y_1,Y_2}(y_1,y_2)$ that is strictly positive if and only if there are constants $a<b$ and $c<d$ such that

$$f_{Y_1,Y_2}(y_1,y_2) > 0 \quad \Leftrightarrow \quad a\le y_1\le b \text{ and } c\le y_2\le d\,.$$

(11.28) *Then Y_1, Y_2 are independent* $\quad\Leftrightarrow\quad f_{Y_1,Y_2}(y_1,y_2) = g_1(y_1)\cdot g_2(y_2)$

for suitable nonnegative functions $g_1, g_2 : \mathbb{R}\to\mathbb{R}$ such that the only argument of g_1 is y_1 and the only argument of g_2 is y_2.

PROOF:

The $\Rightarrow$ direction is trivially true: Choose the marginal densities f_{Y_1} and f_{Y_1} for g_1 and g_2.

PROOF of $\Leftarrow$: From $f(y_1,y_2) = g_1(y_1)\,g_2(y_2)$, we obtain for the marginal densities,

(**A**)
$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f(y_1,y_2)\,dy_2 = \int_{-\infty}^{\infty} g_1(y_1)\,g_2(y_2)\,dy_2 = g_1(y_1)\int_{-\infty}^{\infty} g_2(y_2)\,dy_2 = \alpha g_1(y_1)\,,\\ f_{Y_2}(y_2) &= \int_{-\infty}^{\infty} f(y_1,y_2)\,dy_1 = \int_{-\infty}^{\infty} g_1(y_1)\,g_2(y_2)\,dy_1 = g_2(y_2)\int_{-\infty}^{\infty} g_1(y_1)\,dy_1 = \beta g_2(y_2)\,, \end{aligned}$$

Here, the constants $\alpha = \int_{-\infty}^{\infty} g_2(y_2)\,dy_2$ and $\beta = \int_{-\infty}^{\infty} g_1(y_1)\,dy_1$ satisfy

(**B**)
$$\begin{aligned} \alpha\,\beta &= \int_{-\infty}^{\infty} g_2(y_2)\,dy_2\cdot\int_{-\infty}^{\infty} g_1(y_1)\,dy_1\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(y_1)g_2(y_2)\,dy_1dy_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1,y_2)\,dy_1dy_2 = 1\,. \end{aligned}$$

We conclude that

$$f_{Y_1,Y_2}(y_1,y_2) \overset{\text{(B)}}{=} \alpha\beta\, f_{Y_1,Y_2}(y_1,y_2) = \alpha\beta\, g_1(y_1)g_2(y_2) = \big(\alpha g_1(y_1)\big)\big(\beta g_2(y_2)\big) \overset{\text{(A)}}{=} f_{Y_1}(y_1)f_{Y_2}(y_2)\,.$$

We have proved independence. ■

Example 11.1 (Buffon's needle)**.** The plane is segmented by paralled lines into strips of width $d>0$. A needle of length $\lambda<d$ is dropped at random onto the plane. What is the probability that the line will intersect one of those parallel lines?

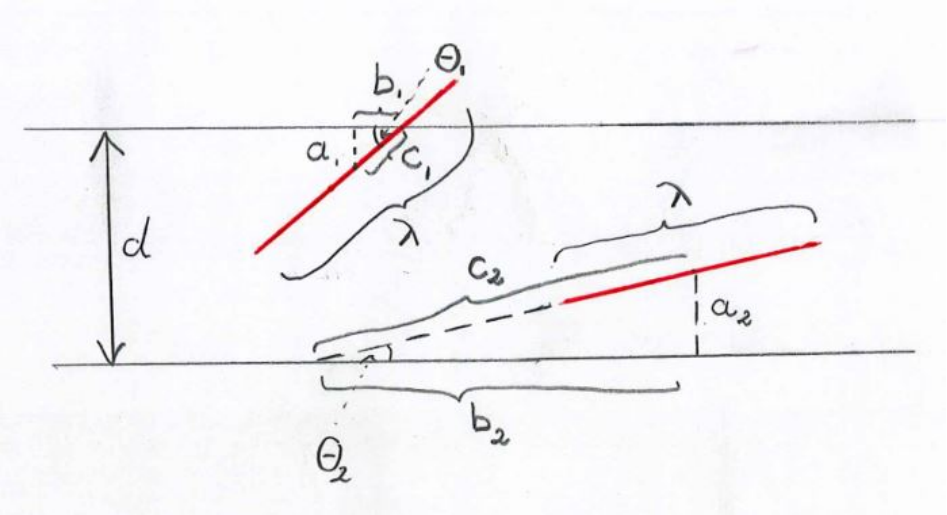
Solution: A needle that is dropped on the plane uniquely determines a right–angled triangle as follows:

- Leg #1 is perpendicular to the parallels. It extends from the midpoint of the needle to the nearest parallel line. Its length is denoted a.
- Its hypothenuse of length c is on the same line as the needle. Thus, it extends from the midpoint of the needle to the point of intersection with that parallel line.
- Leg #2 is located on that parallel line. Its length is denoted b.

We denote the angle formed by the hypothenuse and leg #2 by θ. Thus,

(A) $$\sin(\theta) \;=\; \frac{a}{c}\,, \;\text{ thus, }\; c \;=\; \frac{a}{\sin(\theta)}\,.$$

(B) The needle intersects the (nearest) parallel $\;\Leftrightarrow\; c \,<\, \lambda/2 \;\overset{\text{(A)}}{\Longleftrightarrow}\; \dfrac{a}{\sin(\theta)} \,<\, \lambda/2\,.$



11.1 (Figure). **Buffon's needle**

In Figure 11.1, the triangle on the left satisfies **(B)**:

- $c_1 < \lambda/2$ means that theNE part of the needle extends past the nearest parallel.

On the other hand, the one on the right does not satisfy **(B)**:

- $c_2 > \lambda/2$ means that the SW end of the needle does not reach the nearest parallel.

Note that the triangle created by the random position of the needle is uniquely determined by the two random variables

$$\begin{aligned} \omega \;\mapsto\; A(\omega) \;&:=\; \text{length of leg \#2}\,, \\ \omega \;\mapsto\; \Theta(\omega) \;&:=\; \text{angle between leg \#1 and the hypothenuse.} \end{aligned}$$

Let $\Gamma \subseteq \Omega$ be the event that the needle intersects with a parallel line. We have seen that

$$\omega \in \Gamma \;\overset{\text{(B)}}{\Longleftrightarrow}\; \frac{A(\omega)}{\sin\big(\Theta(\omega)\big)} \,<\, \frac{\lambda}{2}\,. \;\Leftrightarrow\; \big(A(\omega),\Theta(\omega)\big) \in B\,,$$

where

$$B \;=\; \left\{\, (a,\theta) \in]0,d/2[\,\times\,]0,\pi[\,:\; \frac{a}{\sin(\theta)} \,<\, \frac{\lambda}{2} \,\right\} \;=\; \left\{\, (a,\theta) \in]0,d/2[\,\times\,]0,\pi[\,:\; a \,<\, \frac{\lambda}{2}\cdot\sin(\theta) \,\right\}.$$

Here, the constraint $0 < a < d/2$ results from the fact that the midpoint of the needle has a distance of at most $d/2$ from the nearest parallel. Thus, the length $A(\omega)$ of leg #2 cannot exceed $d/2$.

The randomness of the needle toss ensures that

- $A \sim \text{uniform}(0,\lambda/2)$ • $\Theta \sim \text{uniform}(0,\pi)$ • A and Θ are independent.

It follows that the joint PDF of (A,Θ) is

$$f_{A,\Theta}(a,\theta) \;=\; f_A(a)\cdot f_\Theta(\theta) \;=\; \begin{cases} \dfrac{2}{d\pi}\,, & \text{if } 0 \,<\, a \,<\, \dfrac{d}{2}\,,\; 0 \le \theta \le \pi\,, \\ 0\,, & \text{elsewhere}\,. \end{cases}$$

We obtain the probability that a randomly tossed needle intersects one of the parallel lines as

$$P(\Gamma) \;= P\{(A,\Theta)\in B\} \;=\; \iint_B f_{A,\Theta}(a,\theta)\,da\,d\theta$$
$$= \int_0^{\pi}\int_0^{(\lambda/2)\sin(\theta)} \frac{2}{d\pi}\,da\,d\theta \;=\; \frac{\lambda}{d\pi}\int_0^{\pi}\sin(\theta)\,d\theta \;=\; \frac{\lambda}{d\pi}(-\cos\theta)\Big|_0^{\pi} \;=\; \frac{2\lambda}{d\pi}\,.$$

Note that $\int \ldots d\theta$ must go from 0 to π and not just from 0 to $\pi/2$, because a needle with an angle of $30°$ (sloping up) is different from one with an angle of $150°$ (sloping down). □

11.4 The Mulitivariate Uniform Distribution

In this section we extend uniform distribution of Chapter 10.4 (The Uniform Probability Distribution) to regions in two– and threedimensional space.

Definition 11.11 (Multivariate continuous, uniform random variable)**.**

(A) Let $\vec{Y} = (Y_1, Y_2)$ be a twodimensional random vector of continuous random variables with a joint PDF $f_{\vec{Y}}(y_1, y_2)$ that satisfies the following:

- There is a constant $c > 0$ such that either $f_{\vec{Y}}(y_1, y_2) = c$ or $f_{\vec{Y}}(y_1, y_2) = 0$.

Let $C := \{(y_1, y_2) \in \mathbb{R}^2 : f_{\vec{Y}}(y_1, y_2) > 0\}$. Then we say that $\vec{Y}$ has a **continuous uniform probability distribution** on C. □

(B) Let $\vec{Y} = (Y_1, Y_2, Y_3)$ be a threedimensional random vector of continuous random variables with a joint PDF $f_{\vec{Y}}(y_1, y_2, y_3)$ that satisfies the following:

- There is a constant $d > 0$ such that either $f_{\vec{Y}}(y_1, y_2, y_3) = d$ or $f_{\vec{Y}}(y_1, y_2, y_3) = 0$.

Let $D := \{(y_1, y_2, y_2, y_3) \in \mathbb{R}^3 : f_{\vec{Y}}(y_1, y_2, y_3) > 0\}$. Then we say that $\vec{Y}$ has a **continuous uniform probability distribution** on D. □

Remark 11.4. The constants c and d of the previous definition are uniquely determined as follows:
(A) In the twodimensional case,

$$\iint_{\mathbb{R}^2} f_{\vec{Y}}(y_1, y_2)\,dy_1\,dy_2 \;=\; 1 \quad\Rightarrow\quad c \;=\; 1\Big/\iint_C dy_1\,dy_2\,.$$

In other words, c is the reciprocal of the area of C.

(B) In the threedimensional case,

$$\iiint_{\mathbb{R}^3} f_{\vec{Y}}(y_1, y_2, y_3)\,dy_1\,dy_2\,dy_3 \;=\; 1 \quad\Rightarrow\quad d \;=\; 1\Big/\iiint_D dy_1\,dy_2\,dy_3\,.$$

Thus, d is the reciprocal of the volume of D.

(C) It should be obvious how to generalize uniform distribution to n–dimensional random vectors: Let $\vec{Y} = (Y_1, \ldots, Y_n)$ be an n–dimensional random vector of continuous random variables with a joint PDF $f_{\vec{Y}}(\vec{y})$ that satisfies the following:

- There is a constant $e > 0$ such that either $f_{\vec{Y}}(\vec{y}) = e$ or $f_{\vec{Y}}(\vec{y}) = 0$.

Let $E := \{\vec{y} \in \mathbb{R}^n : f_{\vec{Y}}(\vec{y}) > 0\}$. Then we say that $\vec{Y}$ has a **continuous uniform probability distribution** on E.

Similarly to the cases $n = 2$ and $n = 3$, we obtain that e is the reciprocal of the (n–dimensional) volume of E: $e = 1/e'$, where

$$e' := \int_{\vec{y} \in E} \cdots \int d\vec{y} \quad \square$$

Example 11.2. **(a)** What is the uniform density on $C := C_1 \uplus C_2$, where

$$C_1 := \{\vec{y} \in \mathbb{R}^2 : y_1 < 0,\, 0 \le y_2 \le e^{y_1}\}\,, \quad C_2 := \{\vec{y} \in \mathbb{R}^2 : 0 \le y_1 \le 2,\, 0 \le y_2 \le 1\}\,?$$

Note that C_1 has area $\int_{-\infty}^{0} e^{y_1}\, dy_1 = 1$ and C_2, a rectangle of with 2 and height 1, has area 2. Thus, C has area 3 and thus, $c = 1/3$. It follows that

$$f_{\vec{Y}}(\vec{y}) = \begin{cases} \frac{1}{3}\,, & \text{if } y_1 < 0,\, 0 \le y_2 \le e^{y_1},\ \text{or } 0 \le y_1 \le 2,\, 0 \le y_2 \le 1\,, \\ 0\,, & \text{else}\,. \end{cases}$$

(b) Determine the uniform density on

$$D := \{\vec{y} \in \mathbb{R}^3 : y_1 > 0, y_2 > 0, y_3 > 0,\, y_1^2 + y_2^2 + y_3^2 \le 1\}\,.$$

Since Vol(D), the volume of D, is one eighth of $(4/3)\pi$, the volume of the unit sphere, we obtain

$$d = \frac{1}{\text{Vol}(D)} = \frac{8}{(4/3)\pi} = \frac{6}{\pi}\,.$$

Thus,

$$f_{\vec{Y}}(\vec{y}) = \begin{cases} \frac{6}{\pi}\,, & \text{if } y_1 > 0, y_2 > 0, y_3 > 0,\, y_1^2 + y_2^2 + y_3^2 \le 1\,, \\ 0\,, & \text{else}\,. \end{cases} \quad \square$$

11.5 The Expected Value of a Function of Several Random Variables

In this section we must work with vectors $(x_1, x_2, \ldots, x_k)$ of fixed,but arbitrary dimension k, where each component x_j is a real number and thus, $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$. Since it is extremely space consuming to repeatedly write such lengthy objects, we remind you of the "arrow notation" that was introduced in Example 2.21 on p.51.

Notation 11.1 (Arrow notation for vectors)**.**

- We write $\vec{x}$ as an abbreviation for a vector $\big(x_1, x_2, \ldots, x_n\big)$. The dimension n is either explicitly stated or known from the context.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is a function of n real numbers and $U = [a_1, b_1] \times \cdot \times [a_n, b_n]$ is an n–dimensional rectangle, we write

$$\int_A f(\vec{x})\, d\vec{x} \;=\; \int_{a_1}^{b_1} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n)\, dy_1 dy_2 \cdots dy_n$$

  Note that all integrands that occur in this course are so well behaved that the order in which those n integrations take place can be switched around, just as you remember it in the cases $n = 2$ and $n = 3$ from multidimensional calculus.
- Let $a_1 < b_1$, $a_2 < b_2, \ldots, a_n < b_n$ for some $n \in \mathbb{N}$. Then $\vec{y} \in]a_1, b_1] \times \cdot \times]a_d, b_d]$ denotes the following: $\vec{y} = \big(y_1, y_2, \ldots, y_d\big)$ and $a_i < y_i \leq b_i$ for $i = 1, \ldots, d$.

Here are some examples.

(a) $\vec{z} \in \mathbb{R}^m$ means: $\vec{z} = \big(z_1, z_2, \ldots, z_m\big)$ and $z_j \in \mathbb{R}$ for all j.
(b) If $f : \mathbb{R}^k \to \mathbb{R}$, then $g(\vec{y})$ means: $f\big(y_1, \ldots, y_k\big)$.
(c) If $g : \mathbb{R}^d \to \mathbb{R}$, then $g(\vec{Y})$ means: $g\big(Y_1, \ldots, Y_d\big)$; $g\big(\vec{Y}(\omega)\big)$ means: $g\big(Y_1(\omega), \ldots, Y_d(\omega)\big)$.
(d) If $\psi : \mathbb{R}^n \to \mathbb{R}$, then $E\left[\psi(\vec{Y})\right]$ means: $E\left[\psi\big(Y_1, \ldots, Y_n\big)\right]$.

Definition 11.12 (Expected value of $g(\vec{Y})$)**.**

(a) Let $k \in \mathbb{N}$ and let $\vec{Y} = \big(Y_1, Y_2, \ldots, Y_k\big)$ be a vector of discrete random variables on a probability space (Ω, P) with PMF $p_{\vec{Y}}(\vec{y})$. Further, let $g : \mathbb{R}^k \to \mathbb{R}$ be a function of k real numbers $y_1, y_2, \ldots, y_k$. Then

$$E\big[g(\vec{Y})\big] \;=\; E\big[g(Y_1, Y_2, \ldots, Y_k)\big] \;:=\; \sum_{y_1,\, y_2,\, \ldots,\, y_k} \cdots \sum g(\vec{y})\, p_{\vec{Y}}(\vec{y}) \tag{11.29}$$

is called the **expected value** or **mean** of the random variable $g(\vec{Y})$. As usual, the sum on the right is countable summation over those $\vec{y} = \big(y_1, y_2, \ldots, y_k\big)$ for which $p_{\vec{Y}}(\vec{y}) \neq 0$.

(b) Let $k \in \mathbb{N}$ and let $\vec{Y} = \big(Y_1, Y_2, \ldots, Y_k\big)$ be a vector of continuous random variables on a probability space (Ω, P) with PDF $f_{\vec{Y}}(\vec{y})$. Further, let $h : \mathbb{R}^k \to \mathbb{R}$ be a function of k real numbers $y_1, y_2, \ldots, y_k$. Then

$$E\big[h(\vec{Y})\big] \;=\; E\big[h(Y_1, Y_2, \ldots, Y_k)\big] \;:=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\vec{y})\, f_{\vec{Y}}(\vec{y}) d\vec{y} \tag{11.30}$$

is called the **expected value** or **mean** of the random variable $g(\vec{Y})$.

See Notations 11.1 (Arrow notation for vectors) for an explanation of $\int \cdots d\vec{y}$.

As in the onedimensional case, we only are allowed to say that $E\big[g(\vec{Y})\big]$ exists if $\sum\cdots\sum|g(y_1,\ldots,y_k)|\,p(y_1,\ldots,y_k)$ is finite and that $E\big[h(\vec{Y})\big]$ exists if $\int\cdots\int|g(y_1,\ldots,y_k)|\,f(y_1,\ldots,y_k)\,dy_1\ldots dy_k$ is finite. The functions g and h we deal with in this course will always satisfy that assumption. □

Example 11.3. As an example of the power of that definition, we give here the proof that

$$E[Y_1 + \cdots + Y_n] \;=\; E[Y_1] + \cdots + E[Y_n]\,.$$

Let $h(\vec{y}) := y_1 + \cdots + y_n$. Then, by definition 11.12,

$$E[h(\vec{Y})] \;=\; \int_{\mathbb{R}^n}(y_1+\cdots+y_n)f_{\vec{Y}}(\vec{y})d\vec{y} \;=\; \sum_{j=1}^{n}\int_{\mathbb{R}^n} y_j f_{\vec{Y}}(\vec{y})d\vec{y}.$$

Let $\vec{\tilde{y}} := \big(y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_n\big)$. Then $\int(\cdots)d\vec{y} = \int(\cdots)d\vec{\tilde{y}}dy_j)$ because the order of integration can be switched. Since y_j is constant with respect to $\vec{\tilde{y}}$,

$$\int_{\mathbb{R}^n} y_j f_{\vec{Y}}(\vec{y})\,d\vec{y} \;=\; \int_{\mathbb{R}}\left(\int_{\mathbb{R}^{n-1}} y_j f_{\vec{Y}}(\vec{y})\,d\vec{\tilde{y}}\right)dy_j \;=\; \int_{-\infty}^{\infty} y_j\left(\int_{\mathbb{R}^{n-1}} f_{\vec{Y}}(\vec{y})\,d\vec{\tilde{y}}\right)dy_j\,.$$

The inner integral "integrates out" all variables except y_j from the PDF of $\vec{Y}$. Thus, it is the marginal PDF f_{Y_j} of Y_j. It follows from $E[Y_j] = \int_{-\infty}^{\infty} y_j f_{Y_j} dy_j$ that

$$E[h(\vec{Y})] \;=\; \sum_{j=1}^{n}\int_{\mathbb{R}^n} y_j f_{\vec{Y}}(\vec{y})d\vec{y}. \;=\; \sum_{j=1}^{n}\int_{-\infty}^{\infty} y_j f_{Y_j}\,dy_j\,. \;=\; \sum_{j=1}^{n} E[Y_j]\,. \;\square$$

We list here the theorems of WMS Chapter 5.6 (Special Theorems) that detail the rules for evaluating expectations. For the remainder of this section we assume that $Y_1, Y_2, \ldots$ are random variables on a common probability space (Ω, P)

Theorem 11.7 (WMS Ch.05.6, Theorem 5.6)**.**

(11.31) $$c \in \mathbb{R} \;\Rightarrow\; E[c] \;=\; c\,.$$

PROOF: Trivial. ■

Theorem 11.8 (WMS Ch.05.6, Theorem 5.7)**.**

Let $c \in \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ Then the random variable $g(Y_1, Y_2)$ satisfies

(11.32) $$E[cg(Y_1,Y_2)] \;=\; cE[g(Y_1,Y_2)]\,.$$

PROOF: Trivial. ■

Theorem 11.9 (WMS Ch.05.6, Theorem 5.8)**.**

Let $g_1, g_2, \ldots, g_k : \mathbb{R}^n \to \mathbb{R}$ and $\vec{Y} := (Y_1, \ldots, Y_n)$. Then the random variables $g_j(\vec{Y})$ $(j = 1, \ldots, k)$ satisfy

$$E[g_1(\vec{Y}) + g_2(\vec{Y}) + \cdots + g_k(\vec{Y})] = E[g_1(\vec{Y})] + E[g_2(\vec{Y})] + \cdots + E[g_k(\vec{Y})]\,. \tag{11.33}$$

PROOF: We proved in Example 11.3 on p.255 that $E\left[\sum_j U_j\right] = \sum_j E[U_j]$ for discrete or continuous random variables $U_1, \ldots, U_k$. We apply this formula to $U_j := g_j(\vec{Y})$ and the theorem follows. ■

The next theorem generalizes Theorem 11.5 on p.249. That one stated that, for independent random variables, the expectation of the product is the product of the expectations.

Theorem 11.10.

Let $g, h : \mathbb{R} \to \mathbb{R}$ be functions of a single variable and assume that the random variables Y_1 and Y_2 are independent. Then the random variables $g(Y_1)$ and $h(Y_2)$ also are independent and they satisfy

$$E[g(Y_1)\,h(Y_2)] \;=\; E[g(Y_1)]\,E[h(Y_2)]\,. \tag{11.34}$$

PROOF: We give the proof for the continuous case only. It is the WMS proof without any alterations. The proof for the discrete case is similar.

Let $f_{Y_1,Y_2}(y_1, y_2)$ denote the joint PDF of Y_1 and Y_2. Independence of Y_1 and Y_2 yields

$$f_{Y_1,Y_2}(y_1, y_2) \;=\; f_{Y_1}(y_1)\,f_{Y_2}(y_2)\,.$$

The product $g(Y_1)h(Y_2)$ is a function $\varphi(Y_1, Y_2)$ of Y_1 and Y_2. Hence, by Definition 11.12 (Expected value of $g(\vec{Y})$) on p.254,

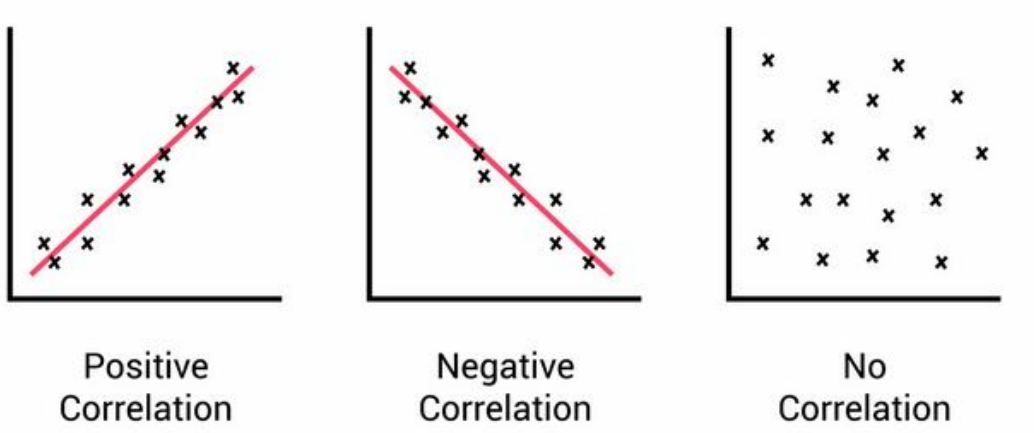
$$\begin{aligned}
E[g(Y_1)h(Y_2)] \;&=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f_{Y_1,Y_2}(y_1,y_2)\,dy_2\,dy_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f_{Y_1}(y_1)\,f_{Y_2}(y_2)\,dy_2\,dy_1 \\
&= \int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\left[\int_{-\infty}^{\infty} h(y_2)f_{Y_2}(y_2)\,dy_2\right] dy_1 \\
&= \int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\,E[h(Y_2)]\,dy_1 \\
&= E[h(Y_2)]\int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\,dy_1 \;=\; E[g(y_1)]\,E[h(Y_2)]\,.
\end{aligned}$$

The proof of the independence of $g \circ Y_1$ and $h \circ Y_2$ is based on a characterization of the independence if random elements X_i which involves $\sigma\{X_i\}$, the sigma algebras generated by each X_i. it is omitted here. ■

11.6 The Covariance of Two Random Variables

Introduction 11.2. If we examine how two random variables Y_1 and Y_2 relate to each other, we can consider among other issues the following:

- **(a)** If the values of Y_1 increase, will the values of Y_2, on average, also tend to increase? One says in this case that Y_1 and Y_2 have **positive correlation**.
- **(b)** Or will the values of Y_2, on average, tend to decrease as the values of Y_1 increase? One says in this case that Y_1 and Y_2 have **negative correlation**.
- **(c)** Or will the values of Y_2, on average, have neither increasing nor falling tendency as the values of Y_1 increase? One says in this case that Y_1 and Y_2 have **zero correlation** or that they are **uncorrelated**.
- **(d)** What if Y_1 and Y_2 are independent? We should expect in that case that Y_1 and Y_2 are uncorrelated.



One can associate with Y_1 and Y_2 a number ρ, their which measures the strength of their correlation. More precsisely, it measures the strength of the linear association between Y_1 and Y_2 and whether that association is of an increasing or decreasing nature. ρ is defined in terms of the covariance of Y_1 and Y_2 and this will be the topic of the current section. □

In this entire section, we consider two random variables Y_1 and Y_2 on a probability space (Ω, P). As usual, we denote mean and standard deviation

$$\mu_j := E[Y_j], \quad \sigma_j := \sqrt{Var[Y_j]}, \quad \text{for } j = 1, 2.$$

Definition 11.13 (Covariance)**.**

The **covariance** of Y_1 and Y_2 is

$$Cov[Y_1, Y_2] = E\big[(Y_1 - E[Y_1])\,(Y_2 - E[Y_2])\big] = E\big[(Y_1 - \mu_1)\,(Y_2 - \mu_2)\big]. \quad \square \tag{11.35}$$

Remark 11.5. $Cov[Y_1, Y_2]$ has the following properties:

(a) The larger the absolute value of the covariance of Y_1 and Y_2, the greater the linear dependence between Y_1 and Y_2.
(b) $Cov[Y_1, Y_2] > 0$ indicates that, on average, Y_1 increases as Y_2 increases.
(c) $Cov[Y_1, Y_2] < 0$ indicates that, on average, Y_1 decreases as Y_2 increases.
(d) $Cov[Y_1, Y_2] = 0$ indicates that, on average, Y_1 remains constant as Y_2 increases. It is a peculiarity of the statistician's lingo that this kind of linear relationship, even if it is very strong, is defined to be as **NO linear relationship** between Y_1 and Y_2.
(e) If we consider $10Y_1$ instead of Y_1 and $10Y_2$ instead of Y_2 the correlation changes by a factor of $10^2 = 100$: $Cov[10Y_1, 10Y_2] = 100Cov[Y_1, Y_2]$. This is not convenient in many situations and one defines a standardized correlation by relating Y_1 and Y_2 to their variances. This will be done in the next definition. □

Definition 11.14 (Correlation coefficient)**.**

The **correlation coefficient**, of Y_1 and Y_2 is

$$\rho = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2} \quad \square \tag{11.36}$$

We say that Y_1 and Y_2 have **positive correlation** if $\rho > 0$, (i.e., if $Cov(Y_1, Y_2) > 0$), they have **negative correlation** if $\rho < 0$, (i.e., if $Cov(Y_1, Y_2) < 0$), and that they have **zero correlation** or that they are **uncorrelated** if $\rho = 0$, (i.e., if $Cov(Y_1, Y_2) = 0$).

Proposition 11.3. *The correlation coefficient satisfies the inequality*

$$-1 \leq \rho \leq 1 \quad \square \tag{11.37}$$

PROOF: Omitted ■

The next formula often makes it easier to compute the covariance.

Theorem 11.11.

$$Cov[Y_1, Y_2] = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = E[Y_1 Y_2] - E[Y_1]\, E[Y_2]\,. \tag{11.38}$$

PROOF: Since $E[U + V] = E[U] + E[V]$ and $E[cU] = cE[U]$ and $E[c] = c$ for all random variables U, V and numbers c,

$$\begin{aligned} Cov[Y_1, Y_2] &= E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= E(Y_1 Y_2 - \mu_1 Y_2 - \mu_2 Y_1 + \mu_1 \mu_2) \\ &= E[Y_1 Y_2] - \mu_1 E[Y_2] - \mu_2 E[Y_1] + \mu_1 \mu_2 \\ &= E[Y_1 Y_2] - \mu_1 \mu_2 - \mu_2 \mu_1 + \mu_1 \mu_2 = E[Y_1 Y_2] - \mu_1 \mu_2\,. \ \blacksquare \end{aligned}$$

Theorem 11.12.

Independent random variables are uncorrelated.

PROOF: By Theorem 11.5 on p.249, independent random variables Y_1 and Y_2 satisfy $E[Y_1Y_2] = E[Y_1]E[Y_2]$. Together with (11.38), we obtain

$$Cov[Y_1, Y_2] \;= E[Y_1Y_2] \;-\; E[Y_1]\,E[Y_2] \;=\; 0\,. \;\blacksquare$$

Example 11.4 (Uncorrelated,but not independent)**.** The following simple examle shows two discrete random variables Y_1 and Y_2 which are uncorrelated, but they are not independent.

We obtain from the joint PMF $p(y_1, y_2)$ of Y_1 and Y_2, shown at the right, that

$E[Y_1] = (-1)\frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 0,$
$E[Y_2] = (-1)\frac{1}{2} + 1 \cdot \frac{1}{2} = 0,$
$E[Y_1Y_2] = (-1)(-1)0 + 0(-1)\frac{1}{2} + (1)(-1)0$
$\quad +(-1)(1)\frac{1}{4} + 0 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot \frac{1}{4} = 0.$

	$\mathbf{Y_2}$	
$\mathbf{Y_1}$	-1	1
-1	0	$1/4$
0	$1/2$	0
1	0	$1/4$

Thus, $E[Y_1Y_2] = E[Y_1]E[Y_2] = 0$ and Y_1 and Y_2 are uncorrelated. On the other hand, $p(-1,-1) = 0$, whereas $p_{Y_1}(-1) \cdot p_{Y_2}(-1) = \frac{1}{4} \cdot \frac{1}{2} \neq 0$. Thus, Y_1 and Y_2 are not independent. $\square$

Definition 11.15 (Linear function)**.** ⋆

Let $n \in \mathbb{N}$. We call a function $\varphi : \mathbb{R}^n \to \mathbb{R};\ \vec{x} = (x_1, \ldots, x_n) \mapsto \varphi(\vec{x})$, a **linear function**, of $x_1, \ldots, x_n$, if there are constants $a_1, \ldots, a_n \in \mathbb{R}$ such that

$$\varphi(\vec{x}) \;=\; a_1x_1 \;+\; a_2x_2 \;+ \cdots + a_nx_n \;=\; \sum_{j=1}^{n} a_jx_j\,. \;\square \tag{11.39}$$

Remark 11.6. Note that if $\vec{Y} = (Y_1, \ldots, Y_n)$ is a vector of random variables, then the function φ of (11.39) defines a random variable $V \;=\; \varphi(\vec{Y}) \;=\; \sum\limits_{j=1}^{n} a_jY_j$. $\square$

Theorem 11.13 (WMS Ch.05.8, Theorem 5.12)**.** *Let $\vec{X} = X_1, \ldots, X_m$ and $\vec{Y} = Y_1, \ldots, Y_n$ be random variables on a probability space (Ω, P). For $i = 1, \ldots, m$ and $j = 1, \ldots, n$, let $\xi_i := E(X_i)$ and $\eta_j := E(Y_j)$. Further, let*

$$U \;:=\; \sum_{i=1}^{m} a_iX_i \qquad \textit{and} \qquad V \;:=\; \sum_{j=1}^{n} b_jY_j\,,$$

where $\vec{a} = (a_1, a_2, \ldots, a_m)$ and $\vec{b} = (b_1, b_2, \ldots, b_n)$ are two constant vectors. Then

$$E[U] = \sum_{i=1}^{m} a_i \xi_i \,, \tag{11.40}$$

$$Var[U] = \sum_{i=1}^{m} a_i^2 Var[X_i] + 2 \sum_{1 \le i < j \le m} \sum a_i a_j Cov[X_i, X_j] \,. \tag{11.41}$$

$$Cov[U, V] = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j Cov[X_i, Y_j] \,. \tag{11.42}$$

In (11.41), $\sum\limits_{1 \le i < j \le m} \sum \cdots$ *refers to summation over all pairs* (i, j) *satisfying* $i < j$.

PROOF: The theorem consists of three parts, of which (11.40) follows directly from Theorems 11.8 and 11.9.

Proof of (11.41): From the definition of variance we obtain

$$\begin{aligned} Var[U] &= E[U - E[U]]^2 = E\left[\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \xi_i\right]^2 = E\left[\sum_{i=1}^{n} a_i (X_i - \xi_i)\right]^2 \\ &= E\left[\sum_{i=1}^{n} a_i^2 (X_i - \xi_i)^2 + \sum_{\substack{i=1 \\ i \ne j}}^{n} \sum_{j=1}^{n} a_i a_j (X_i - \xi_i)(X_j - \xi_j)\right] \\ &= \sum_{i=1}^{n} a_i^2 E[X_i - \xi_i]^2 + \sum_{\substack{i=1 \\ i \ne j}}^{n} \sum_{j=1}^{n} a_i a_j E[(X_i - \xi_i)(X_j - \xi_j)] \,. \end{aligned}$$

By the definitions of variance and covariance, we have

$$E[(X_i - \xi_i)^2] = Var[X_i] \qquad \text{and} \qquad E[(X_i - \xi_i)(X_j - \xi_j)] = Cov[X_i, X_j] \,.$$

Thus,

$$Var[U] = \sum_{i=1}^{n} a_i^2 Var[X_i] + \sum_{\substack{i=1 \\ i \ne j}}^{n} \sum_{j=1}^{n} a_i a_j Cov[X_i, X_j] \,.$$

We apply symmetry $Cov[X_i, X_j] = Cov[X_j, X_i]$ to the double summation and obtain

$$Var[U] = \sum_{i=1}^{n} a_i^2 Var[X_i] + 2 \sum_{1 \le i < j \le n} \sum a_i a_j Cov[X_i, X_j] \,.$$

We have shown (11.41). To prove (11.42), we proceed in a similar fashion: We have

$$\begin{aligned} Cov[U,V] &= E\big[\,(U-E[U])(V-E[V])\,\big] \\ &= E\left[\left(\sum_{i=1}^{m} a_i X_i - \sum_{i=1}^{m} a_i \xi_i\right)\left(\sum_{j=1}^{n} b_j X_j - \sum_{j=1}^{n} b_j \eta_j\right)\right] \\ &= E\left[\left(\sum_{i=1}^{m} a_i (X_i - \xi_i)\right)\left(\sum_{j=1}^{n} b_j (Y_j - \eta_j)\right)\right] \end{aligned}$$

$$\begin{aligned} \text{Thus,}\quad Cov[U,V] &= E\left[\sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j (X_i - \xi_i)(Y_j - \eta_j)\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j E[(X_i - \xi_i)(Y_j - \eta_j)] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov[X_i, Y_j]\,. \quad \blacksquare \end{aligned}$$

Remark 11.7. Note the following about Theorem 11.13:

(a) Neither CDFs, PMFs or PDFs were needed to prove the theorem. Thus, the proof applies to both discrete and continuous random variables.

(b) Since $Cov[Y_i, Y_i] = Var[Y_i]$, (11.41) is a particular version of (11.42). □

We are now in a position to prove (10.28) of Theorem 10.8 on p.224 Those formulas state that, for independent random variables, the variance of the sum equals the sum of the variances. Even better, independence can be replaced with the weaker assumption of correlation zero. (See Theorem 11.12.)

Corollary 11.1 (Bienaymé formula for uncorrelated variables). ⋆

Let $Y_1, Y_2, \ldots, Y_n : \Omega \to \mathbb{R}$ be uncorrelated random variables (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$. Then

(11.43) $$Var\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var[Y_j]\,.$$

PROOF: Since $Y_1, \ldots, Y_n$ are uncorrelated, $Cov[Y_i, Y_j] = 0$ for $1 \le i, j \le n$ and $i \ne j$. We employ (11.41) on p.260 with $a_1 = a_2 = \cdots = a_n = 1$ and obtain

$$Var\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} Var[Y_i] + 2\sum_{1\le i<j\le n}\sum Cov[Y_i, Y_j] = \sum_{i=1}^{n} Var[Y_i] + 0\,. \quad \blacksquare$$

Example 11.5 (Variance of the sample mean [104])**.** This example belongs thematically to Section 8.2 (Sampling and Urn Models With and Without Replacement). We model SRS sampling from a population to infer statistical knowledge about it as follows.

- The population is represented by a probability space (Ω, P) and the statistical knowledge we are interested in is part of the distribution of a random variable Y on (Ω, P).
- Picking at random an item from the population is modeled as the outcome $Y(\omega)$ of an invocation of Y.
- Picking an SRS sample of size n from the population is modeled as the n outcomes $\vec{Y}(\omega) = \big(Y_1(\omega), \ldots, Y_n(\omega)$ of n independent random variables $Y_1, \ldots Y_n$ which have the same distribution as Y. In other words, the Y_j are a (finite) iid sequence in the sense of Definition 5.18 on p.136.
- Of course, that last point is an idealization, since independent sample picks correspond to sampling with replacement, whereas SRS models to sampling without replacement. See Definitions 8.3 on p.186 and 8.5 about SRS and urn models. On the other hand, the computational differences between results based on sampling with and without replacement are of practical insignificance if the sample size is small when compared to the population size. [105]

In this example we specifically consider the mean of the population data.

- It seems natural to model this mean it by the mean of Y, i.e., the expectation $\mu = E[Y]$ of Y.
- So that's it then. $E[Y]$ is the answer we are looking for. Well, it would be if we only knew the distribution of Y or, at least, $E[Y]$.
- But we don't! We "defined" Y as the action of taking a single random pick from the population, and that is the extent of our knowledge of Y.
- This is why we introduced the vector $\vec{Y}$ of n iid sample picks. The randomness and independence of $Y_1, \ldots Y_n$ should make the specific sample $\vec{y}$ that consists of the outcomes $y_j = Y_j(\omega)$ representative of the population. Thus, its **sample mean** $\bar{y} = \bar{Y}(\omega)$ which is obtained by averaging the sample data, i.e.,

$$\bar{Y}(\omega) \;=\; \frac{Y_1(\omega) + Y_2(\omega) + \cdots + Y_n(\omega)}{n}\,,$$

  should result in a good estimate of the population mean.

All of the above serves as motivation for the following setup. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with common expectation $E[Y_j] = \mu$ and variance $Var[Y_j] = \sigma^2$ $(j = 1, \ldots, n)$. Let

$$\bar{Y} \;:=\; \frac{1}{n}\sum_{j=1}^{n} Y_j\,. \tag{11.44}$$

[104] This is a modified version of WMS, Example 5.27.

[105] See parts **(c)** and **(d)** of Remark 8.2 on p.185.

It follows from (11.40) on p.260 and Corollary 11.1 on p.261 that

$$E[\bar{Y}] = E\left[\frac{1}{n}\sum_{j=1}^{n} Y_j\right] = \frac{1}{n}E\left[\sum_{j=1}^{n} Y_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[Y_j] = \frac{1}{n}(n\,\mu) = \mu\,,$$

$$Var[\bar{Y}] = Var\left[\frac{1}{n}\sum_{j=1}^{n} Y_j\right] = \frac{1}{n^2}Var\left[\sum_{j=1}^{n} Y_j\right] = \frac{1}{n^2}\sum_{j=1}^{n} Var[Y_j] = \frac{1}{n^2}(n\,\sigma^2) = \frac{\sigma^2}{n}\,.$$

We infer from those two formulas the following.

Recall that the purpose of $\bar{Y}$ is to serve as an **estimator** for the following population parameter: The population mean, which is the mean of anyone of the sample picks $\mu = E[Y_j]$.

The significance of the formula $E[\bar{Y}] = \mu$ is as follows

- The expected value of this estimator equals the parameter it is meant to estimate.

An estimator with that property is referred to as an **unbiased estimator**.

Now to the formula $Var[\bar{Y}] = \sigma^2/n$. We use it to compare the standard deviations

$$\sigma_{Y_j} = \sqrt{Var[Y_j]} \qquad \text{and} \qquad \sigma_{\overline{Y}} = \sqrt{Var[\bar{Y}]}$$

of a single pick Y_j and the average $\bar{Y}$ of n such independent picks. Note that the standard deviation of a random variable U is a measure for its concentration about its expected value. (And the same is true for its variance.) A small σ_U signifies that most outcomes $U(\omega)$ are in close vicinity of $E[U]$.

Thus, $\sigma_{\overline{Y}}$ is a measure for the lack of precision with which $\bar{Y}$ estimates $E[\bar{Y}] = \mu$.

- In the extreme case of a sample of size 1, i.e., $n = 1$, that lack of precision is σ.
- For $n = 100$, that lack of precision goes down to $\dfrac{\sigma}{10}$. Thus, precision has improved by a factor of 10.
- Generally speaking, increasing the sample size by the factor K (and spending all that time and money doing so) does not reward us with a proportionate improvement of the precision of the estimate $\bar{Y}$. It only increases by the factor $\sqrt{K}$. □

11.7 Conditional Expectations and Conditional Variance

11.7.1 The Conditional Expectation With Respect to an Event ⋆

We will start with a definition of the conditional expectation $E[Y \mid B]$ of a random variable Y where conditioning happens with respect to an event $B \subseteq \Omega$. This definition is usually not taught in an undergraduate level course on probability theory for the following reason: It cannot be extended, in the case of continuous random variables Y and $\tilde{Y}$, to $E[Y \mid \tilde{Y} = \tilde{y}]$, i.e., conditioning on $\tilde{Y}$ having a fixed outcome $\tilde{y}$.

All that follows in this subsection is based on Theorem 5.8 on p.118 which states the following: If (Ω, P) is a probability space and $B \subseteq \Omega$ is an event that satisfies $P(B) > 0$, then the function $Q(\cdot)$, defined as $Q(A) := P(A \mid B)$ for $A \subseteq \Omega$, is a probability measure on Ω. [106]

[106] To be exact, there also was a σ–algebra $\mathscr{F}$ and we had to assume that $B \in \mathscr{F}$ and that $P(A)$ is defined only for $A \in \mathscr{F}$. This in turn implies that $Q(A) = P(A \mid B)$ only is defined for arguments $A \in \mathscr{F}$. We do not mention $\mathscr{F}$ since we decided to avoid dealing with σ–algebras whenever possible.

Assumption 11.3.

In all of this subsection we deal with a fixed probability space (Ω, P) and a fixed event $B \subseteq \Omega$ that satisfies $P(B) > 0$. We further assume that $Q(\cdot)$ is the probability measure

$$A \mapsto Q(A) := P(A \mid B), \quad \text{where } A \subseteq \Omega. \tag{11.45}$$

The symbols $X, X_1, X_2, \ldots$ denote random elements and $X, X_1, X_2, \ldots$ denote random variables on Ω. We need not be specific about whether we mean (Ω, P) or (Ω, Q), because the definition of random element and random variable does not involve the probability measure, only the carrier space Ω. □

Remark 11.8. The following mathematical triviality allows us to translate much that we have done with random variables in connection with P to their analogues with respect to $Q = P(\cdot \mid B)$.

- All definitions, propositions and theorems in which an unspecified probability measure P is involved can be reformulated by replacing P with Q.

Here is a list (certainly not complete) of many such concepts.

• cumulative distribution function, • probability mass function

• probability density function • joint CDF • joint PMF • joint PDF

• expectation • variance • moments • moment generating function

BEWARE: The above does not apply to cases where a specific probability measure is considered. An example for this would be, e.g., Proposition 10.9 on p.236 (memorylessness of the exponential distribution). Here the probability measure is an exponential distribution P_Y.

We will elaborate on some of the items in that bulleted list in the next remark. □

Remark 11.9. In the following, the phrase "Q–....." serves as an abbreviation for the lengthier "..... with respect to Q".

(a) The Q–CDF of a random variable Y is $F_Y^Q(y) = Q\{Y \le y\} = P\{Y \le y \mid B\}$.

(b) The Q–PMF of a discrete random element [107] X is $p_X^Q(x) = Q\{X = x\} = P\{X = x \mid B\}$.

(c) Assume that the derivative $f_Y^Q(y) = \frac{dF_Y^Q(y)}{dy}$ of the Q–CDF of a random variable Y exists and is continuous except for at most finitely many y in any finite interval. Then Y is a Q–continuous random variable with Q–PDF $f_Y^Q(y)$. [108]

(d) We skip joint Q–CDFs and joint Q–PDFs and only elaborate on the joint Q–PMF. of two random elements X_1 and X_2. It is, as one should expect, defined as $p_{X_1,X_2}^Q(x_1, x_2) = Q\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1, X_2 = x_2 \mid B\}$.

(e) The Q–expected value of a discrete random variable Y is $E^Q[Y] = \sum_y y \cdot p_Y^Q(y) = \sum_y y \cdot P\{Y = y \mid B\}$. ($\sum_y$ is over all y where $p_Y^Q(y) > 0$.)

(f) The Q–expectation of a continuous random variable Y is $E^Q[Y] = \int_{-\infty}^{\infty} y \cdot f_Y^Q(y)dy$.

[107] Since $P\{X = x\} \cap B \le P\{X = x\}$, $P\{X = x\} = 0$ implies $Q\{X = x\} = 0$. Thus, any P–discrete random element also is Q–discrete.

[108] There may be some reasonably general and simple conditions that guarantee Y being Q–continuous from being P–continuous, but this author is not aware of them.

(g) The Q–variance of a random variable Y is $Var^Q[Y] = E^Q\left[(Y - E^Q[Y])^2\right]$.
(h) The Q–MGF of a random variable Y is $m_Y^Q(t) = E^Q\left[e^{tY}\right]$.

For expectations of functions of random variables we skip the case of one or two random variables and proceed directly to the case of a vector $\vec{Y} = \left(Y_1, Y_2, \ldots, Y_k\right)$ of random variables. (See Definition 11.12 on p.254.)

(i) If the Y_j are discrete and $g : \mathbb{R}^k \to \mathbb{R}$, then $E^Q\left[g(\vec{Y})\right] = \sum_{y_1,\, y_2,\, \ldots,\, y_k} \cdots \sum g(\vec{y})\, p_{\vec{Y}}^Q(\vec{y})$.
(j) If the Y_j are continuous and $h : \mathbb{R}^k \to \mathbb{R}$, then $E^Q\left[h(\vec{Y})\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\vec{y})\, f_{\vec{Y}}^Q(\vec{y}) d\vec{y}$. □

Here are some of the theorems we get for free because we have shown them for any probability measure. Again, BEWARE: We made the assumption $P(B) > 0$!

Theorem 11.14.

If $\vec{Y} = \left(Y_1, Y_2, \ldots, Y_k\right)$ is a vector of k discrete or Q–continuous random variables, then

$$E^Q\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E^Q[Y_j]\,. \tag{11.46}$$

PROOF: This follows from Theorem 10.7 on p.224. ■

Theorem 11.15. *If Y is a discrete or Q–continuous random variable and $\vec{Y} = \left(Y_1, Y_2, \ldots, Y_k\right)$ is a vector of k Q–independent discrete or Q–continuous random variables, then*

$$Var^Q[Y] = E^Q\left[Y^2\right] - \left(E^Q[Y]\right)^2, \tag{11.47}$$

$$Var^Q\left[aY + b\right] = a^2 Var^Q[Y]\,, \tag{11.48}$$

$$Var^Q\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var^Q[Y_j]\,. \tag{11.49}$$

PROOF: This follows from Theorem 10.8 on p.224. ■

There is an issue with that last theorem. Not just with the proof, but with the assumptions that were made. How is Q–independence defined for random variables, or even for events A_1, A_2, A_k? The answer is, of course, that we apply all previously made definitions of independence of two or more events or random variables, replacing the original probability measure P with Q.

The following theorem about the Q–independence of two events is worthwhile mentioning.

Theorem 11.16.

Let $A_1, A_2, B \subseteq \Omega$ be three events such that $P(A_1 > 0$, $P(A_2 > 0$ and $P(B > 0$. Then

$$
\begin{array}{lll}
 & \textbf{(a)} & P(A_1 \cap A_2 \mid B) \;=\; P(A_1 \mid B) \cdot P(A_2 \mid B) \\
(11.50) \quad \Leftrightarrow & \textbf{(b)} & P(A_1 \mid A_2 \cap B) \;=\; P(A_1 \mid B) \\
\Leftrightarrow & \textbf{(c)} & P(A_2 \mid A_1 \cap B) \;=\; P(A_2 \mid B)\,.
\end{array}
$$

In other words, if A_i and A_j are independent with respect to "just" conditioning on B, then "further" conditioning of A_i on both A_j and B has no effect. Here, either $i = 1, j = 2$ or $i = 2, j = 1$.

PROOF: Since **(a)** is aymmetrical in A_1 and A_2 and **(c)** is obtained from **(b)** by switching the roles of A_1 and A_2, it suffices to prove **(a)** $\Leftrightarrow$ **(b)**.

PROOF that **(a)** $\Rightarrow$ **(b)**:

$$
\begin{aligned}
P(A_1 \mid A_2 \cap B) \;&=\; \frac{P(A_1 \cap A_2 \cap B)}{P(A_2 \cap B)} \;=\; \frac{P(A_1 \cap A_2 \cap B)}{P(B)} \cdot \frac{P(B)}{P(A_2 \cap B)} \\
&= P(A_1 \cap A_2 \mid B) \cdot \frac{1}{P(A_2 \mid B)} \overset{\textbf{(a)}}{=} P(A_1 \mid B) \cdot P(A_2 \mid B) \cdot \frac{1}{P(A_2 \mid B)} \\
&= P(A_1 \mid B)\,.
\end{aligned}
$$

PROOF that **(b)** $\Rightarrow$ **(a)**:

$$
\begin{aligned}
P(A_1 \cap A_2 \mid B) \;&=\; \frac{P(A_1 \cap A_2 \cap B)}{P(B)} \;=\; \frac{P(A_1 \cap A_2 \cap B)}{P(A_2 \cap B)} \cdot \frac{P(A_2 \cap B)}{P(B)} \\
&= P(A_1 \mid A_2 \cap B) \cdot P(A_2 \mid B) \overset{\textbf{(b)}}{=} P(A_1 \mid B) \cdot P(A_2 \mid B)\,. \quad \blacksquare
\end{aligned}
$$

11.7.2 The Conditional Expectation w.r.t a Random Variable or Random Element

Remark 11.10. $\boxed{\star}$ We mentioned at the beginning of the previous subsection 11.7.1 (The Conditional Expectation With Respect to an Event), that conditioning with respect to an event B constitutes a dead end street. This is the reason why the material has been marked as $\boxed{\star}$ (optional). Now let us give the reason.

As far as modeling reality by means of probability theoretical concepts is concerned, the primary interest of conditioning is being able to assume during certain calculations of the probability involving a random element X_1, that another random element X_2 has as its outcome a fixed value x_2. Thus, we typically are interested in

- $P\{X_1 \in B_1 \mid X_2 = x_2\}$, where x_2 is some fixed outcome that can be attained by X_2.

Having stated the issue in the most general terms, we will restrict ourselves for the remainder of this remark to random variables Y_1 and Y_2 rather than working with random elements. This will allow us to contrast discrete and continuous random variables.

The method of subsection 11.7.1 (using the probability measure $Q(A) = P\{A \mid Y_2 = y_2\}$ will actually work if we condition on specific values of a discrete random variable Y_2. This is so because we only are interested in those outcomes y_2 for which

$$
p_{Y_2}(y_2) \;=\; P\{Y_2 = y_2\} \;>\; 0
$$

and the conditional probability $P\{A \mid Y_2 = y_2\}$ exist for such outcomes y_2.

On the other hand, we have nothing at all to work with if Y_2 is continuous, since $P\{Y_2 = y_2\} = 0$ for all numbers y_2 (see Proposition 10.1 on p.213), since this results in $P\{Y_1 \in B_1 \mid Y_2 = y_2$ being **UNDEFINED** for all numbers y_2!

To overcome this hurdle we will work with the conditional PMFs and PDFs

- $p_{Y_1|Y_2}(y_1 \mid y_2) = \dfrac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)}$, if Y_1 and Y_2 are discrete random variables,
- $f_{Y_1|Y_2}(y_1 \mid y_2) = \dfrac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}$, if Y_1 and Y_2 are continuous random variables.

We close this remark by noticing that, in the case of discrete random variables, working with $Q\{Y_1 \in B_1\} = P\{Y_1 \in B \mid Y_2 = y_2\}$ or with $p_{Y_1|Y_2}(y_1 \mid y_2)$ amounts to the same, because Q and $p_{Y_1|Y_2}$ satisfy

$$Q\{Y_1 \in B_1\} = \sum_{y_1 \in B_1} P\{Y_1 = y_1 \mid Y_2 = y_2\} = \sum_{y_1 \in B_1} p_{Y_1|Y_2}(y_1 \mid y_2)\,. \quad \square$$

Compare the following remark to Remark 11.8 on p.264 for discrete random variables.

Remark 11.11. The following allows us to translate much that we have done with continuous random variables in connection with P to their analogues where we replace the (marginal) PDF $f_{Y_1}(y_1)$ with the conditional PDF $f_{Y_1|Y_2}(y_1 \mid y_2)$:

- Assume that $y_2 \in \mathbb{R}$ satisfies $f_{Y_2}(y_2) > 0$. Then the integrable function

$$f_{Y_1|Y_2}(\cdot \mid y_2) \,:\, y_1 \mapsto f_{Y_1|Y_2}(y_1 \mid y_2) \qquad \text{satisfies}$$

$$\boxdot\ f_{Y_1|Y_2}(y_1 \mid y_2) \geq 0 \quad \text{for } -\infty < y_1 < \infty \qquad \boxdot\ \int_{-\infty}^{\infty} f_{Y_1|Y_2}(y_1 \mid y_2) dy_1 = 1\,.$$

- According to Theorem 10.3 on p.215, $f_{Y_1|Y_2}(\cdot \mid y_2)$ is the PDF of the probability measure P_{y_2} on Ω, defined by

$$P_{y_2}\{a < Y_1 \leq b\} = \int_a^b f_{Y_1|Y_2}(y_1 \mid y_2)\, dy_1\,.$$

- Thus, all definitions, propositions and theorems in which an unspecified probability measure P is involved can be reformulated by replacing P with P_{y_2}.

This applies, among others, to the following concepts which were listed in Remark 11.8 on p.264 for discrete random variables:

• cumulative distribution function, • probability mass function
• probability density function • joint CDF • joint PMF • joint PDF
• expectation • variance • moments • moment generating function

- All that was said above extends to a random vector $\vec{U} = (U_1, \ldots, U_k)$ in place of Y_1. We only must replace $f_{Y_1,Y_2}(y_1, y_2)$ with $f_{\vec{U},Y_2}(u_1, \ldots, u_k, y_2)$, etc. $\square$

Definition 11.16 (Conditional expectation)**.**

Let Y_1 and Y_2 be two random variables which are either jointly discrete or jointly continuous and $g : \mathbb{R} \to \mathbb{R}$. Let

$$E[g(Y_1) \mid Y_2 = y_2] := \sum_{y_1} g(y_1)\, p(y_1 \mid y_2) \quad \text{(discrete case)}, \tag{11.51}$$

$$E[g(Y_1) \mid Y_2 = y_2] := \int_{-\infty}^{\infty} g(y_1)\, f(y_1 \mid y_2)\, dy_1 \quad \text{(continuous case)}. \tag{11.52}$$

We call $E[g(Y_1) \mid Y_2 = y_2]$ the **conditional expectation** of $g(Y_1)$, given that $Y_2 = y_2$. □

Remark 11.12. Note for the following that the function

$$\omega \;\mapsto\; E[g(Y_1) \mid Y_2 = Y_2(\omega)] \;=\; E[g(Y_1) \mid Y_2 = y_2]\Big|_{y_2 = Y_2(\omega)}$$

defines a random variable on (Ω, P). It is customary in many situations to suppress the argument ω and write

$$E[g(Y_1) \mid Y_2] \tag{11.53}$$

for this random variable. Clearly, if we write $Z(\omega)$ for $E[g(Y_1) \mid Y_2 = Y_2(\omega)]$, we can take its (unconditional) expectation

$$E[Z] = E\big[E[g(Y_1) \mid Y_2]\big]\,. \tag{11.54}$$

In particular, if $g(y) = y$, we can take the expectation $E\big[E[Y_1 \mid Y_2]\big]$ of $E[Y_1 \mid Y_2]$. We will do so in the next theorem. □

Theorem 11.17 (WMS Ch.05.11, Theorem 5.14)**.**

Let Y_1 and Y_2 be either jointly continuous or jointly discrete random variables. Then

$$E[Y_1] \;=\; E\big[E[Y_1 \mid Y_2]\big]\,. \tag{11.55}$$

See Remark 11.12 concerning the interpretation of the right–hand side.

PROOF: We give the proof for jointly continuous Y_1 and Y_2. With the usual notation for joint PDF, marginal densities and conditional PDF we obtain

$$\begin{aligned}
E[Y_1] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1\, f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1\, f_{Y_1|Y_2}(y_1 \mid y_2) f_{Y_2}(y_2)\, dy_1\, dy_2 \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y_1\, f_{Y_1|Y_2}(y_1 \mid y_2)\, dy_1\right) f_2(y_2)\, dy_2 \\
&= \int_{-\infty}^{\infty} E[Y_1 \mid Y_2 = y_2]\, f_{Y_2}(y_2)\, dy_2 \;=\; E\big[E[Y_1 \mid Y_2]\big]\,.
\end{aligned}$$

The proof for the discrete case is done by doing summation instead of integration and replacing joint, marginal and conditional PDFs with the corresponding PMFs. ■

We define the conditional variance of Y_1 given $Y_2 = y_2$ by applying Definition 11.16 to the functions $g(y_1) = y_1$ and $g(y_1) = y_1^2$.

Definition 11.17 (Conditional variance)**.**

Let Y_1 and Y_2 be two random variables which are either jointly discrete or jointly continuous. Let

$$Var[Y_1 \mid Y_2 = y_2] := E[Y_1^2 \mid Y_2 = y_2] - \big(E[Y_1 \mid Y_2 = y_2]\big)^2 . \tag{11.56}$$

We call $Var[Y_1 \mid Y_2 = y_2]$ the **conditional variance** of $g(Y_1)$, given that $Y_2 = y_2$. □

Theorem 11.18.

Let Y_1 and Y_2 be jointly discrete or jointly continuous random variables. Then

$$Var[Y_1 \mid Y_2] = E\big[(Y_1 - E[Y_1 \mid Y_2])^2 \mid Y_2\big], \tag{11.57}$$

$$Var[Y_1] = E\big[Var[Y_1 \mid Y_2]\big] + Var\big[E[Y_1 \mid Y_2]\big]. \tag{11.58}$$

PROOF: We only give the proof of (11.58). Note that

$$\textbf{(A)} \qquad Var[Y_1 \mid Y_2] = E[Y_1^2 \mid Y_2] - \big(E[Y_1 \mid Y_2]\big)^2,$$

$$\textbf{(B)} \qquad E\big[Var[Y_1 \mid Y_2]\big] = E\big[E[Y_1^2 \mid Y_2]\big] - E\big[\big(E[Y_1 \mid Y_2]\big)^2\big].$$

By the definition of (unconditional) variance,

$$\textbf{(C)} \qquad Var\big[E[Y_1 \mid Y_2]\big] = E\big[\big(E[Y_1 \mid Y_2]\big)^2\big] - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2.$$

Further,

$$\begin{aligned}
Var[Y_1] &= E\big[Y_1^2\big] - \big(E[Y_1]\big)^2 \\
&= E\big[E[Y_1^2 \mid Y_2]\big] - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2 \\
&= E\big[E[Y_1^2 \mid Y_2]\big] - E\big[\big(E[Y_1 \mid Y_2]\big)^2\big] + E\big[\big(E[Y_1 \mid Y_2]\big)^2\big] - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2 \\
&= E\big[E[Y_1^2 \mid Y_2] - \big(E[Y_1 \mid Y_2]\big)^2\big] + \big\{E\big[\big(E[Y_1 \mid Y_2]\big)^2 - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2\big\} \\
&= E\big[Var[Y_1 \mid Y_2]\big] + Var\big[E[Y_1 \mid Y_2]\big]. \ ■
\end{aligned}$$

11.7.3 Conditional Expectations as Optimal Mean Squared Distance Approximations

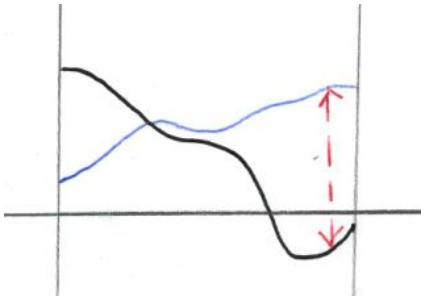
The presentation of the material presented here follows [1] Bickel and Doksum: Mathematical Statistics.

Introduction 11.3. One can measure the distance between two real–valued functions in several ways.

For example, one can define for $\varphi, \psi : A \to \mathbb{R}$,

$$\text{dist}_1(\varphi, g) := \max\{|\varphi(a) - \psi(a)| : a \in A\}\,.$$

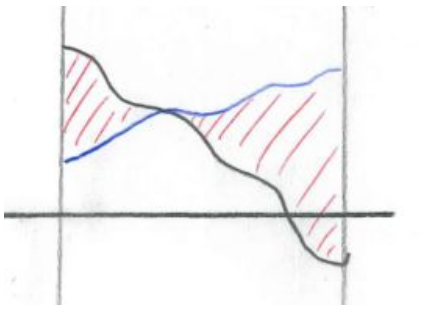
In other words, one takes the maximum displacement over all arguments of φ and ψ. This "worst case scenario" as the advantage that it works for any kind of domain A, since all that is needed is that the function values are numeric.



However, it often makes more sense to consider the area between the curves defined by φ and ψ.

$$\text{dist}_2(\varphi, \psi) := \int_a^b |\varphi(x) - \psi(x)|\, dx\,.$$

Doing so averages out all individual displacements $|\varphi(x) - \psi(x)|$ over all arguments and one obtains a measure of distance which is not distorted just one potential outlier.



There are mathematical reasons why one would rather work with the squared difference and consider

$$\text{dist}_3(\varphi, \psi) \;:=\; \int_a^b |\varphi(x) - \psi(x)|^2\, dx \;=\; \int_a^b \big(\varphi(x) - \psi(x)\big)^2\, dx\,.$$

Moreover, one can replace the ordinary integral $\int \cdots dx$ with a weighted integral $\int \cdots w(x)dx$ where $w(x) \geq 0$ for all x and define

$$\text{dist}_4(\varphi, \psi) := \int_a^b \big(\varphi(x) - \psi(x)\big)^2 w(x)\, dx\,.$$

Here, bigger values $w(x)$ of the weight function w lead to a stronger contribution of $\varphi(x) - \psi(x)$ to the distance between φ and ψ

That last example shows us how the expectation of the difference of two functions of two continuous random variables can be viewed as a distance

$$\text{dist}\big(\varphi(Y_1), \psi(Y_2)\big) \;= E[\,\big(\varphi(Y_1) - \psi(Y_2)\big)^2\,] \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(\varphi(y_1) - \psi(y_2)\big)^2 f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2\,.$$

Since $E[\,\big(\varphi(Y_1) - \psi(Y_2)\big)^2\,]$ also is defined for discrete random variables, we obtain for those a corresponding definition by replacing the joint PDF with the joint PMF and integration with summation:

$$\text{dist}\big(\varphi(Y_1), \psi(Y_2)\big) \;= E[\,\big(\varphi(Y_1) - \psi(Y_2)\big)^2\,] \;=\; \sum_{y_1, y_2} \big(\varphi(y_1) - \psi(y_2)\big)^2 p_{Y_1,Y_2}(y_1, y_2)\,.$$

In either discrete or continuous case, we are particularly interested in the case $\varphi(y_1) = y_1$ and examine the distance

$$\text{dist}\big(Y_1, \psi(Y_2)\big) \;= E[\big(Y_1 - \psi(Y_2)\big)^2]$$

for all possible functions $y_2 \mapsto \psi(y_2)$. It turns out that the minimum

$$\min\{\,\text{dist}\big(Y_1, \psi(Y_2)\big) : \text{ all suitable functions } \psi\}$$

is attained by selecting $\psi : y_2 \mapsto E[Y_1 \mid Y_2 = y_2]$. $\square$

Lemma 11.1. ⋆

Let Y be a random variable on (Ω, P) that satisfies $E[Y^2] < \infty$. Then, $E[|Y|] < \infty$.

PROOF:

Let $A := |Y| < 1$ and $Z := \mathbf{1}_A + |Y^2|$, i.e., $Z(\omega) = \begin{cases} 1 + |Y^2|\,, & \text{if } |Y\omega| < 1\,, \\ |Y^2|\,, & \text{if } |Y\omega| \geq 1\,. \end{cases}$

Since $|Y(\omega)| < 1$ for $\omega \in A$ and $Y(\omega)^2 \geq 1$ for $\omega \in A^{\complement}$, we obtain $|Y(\omega)| \leq Z(\omega)$ for all ω. Thus,

$$E[|Y|] \;\leq\; E[|Z|] \;\leq\; E[1] + E[Y^2]\,.$$

The assertion follows. ■

Lemma 11.2. ⋆

Let Y be a random variable on (Ω, P) and $h : \mathbb{R} \to [0, \infty]$ defined by $a \mapsto E[(Y-a)^2]$. Then,
either ***(a)*** *$h(a) = \infty$ for all $a \in \mathbb{R}$,*
or ***(b)*** *h attains a unique minimum at $a = E[Y]$.*

PROOF:

Step I: We show that either $h(y) \equiv \infty$ for all y or $h(y) \in \mathbb{R}$ for all y.

For fixed $a \in \mathbb{R}$,we define $F : \mathbb{R} \to \mathbb{R}$ by $F(y) := (y-a)^2 - \big((1/2)y^2 - a^2\big)$. Then,

$$F'(y) \;= 2(y-a) - y \;=\; y - 2a \qquad \text{and} \qquad F''(y) \;=\; 1.$$

It follows that F attains a (unique) minimum at $y = 2a$. From $F(2a) \;=\; a^2 - (2a^2 - a^2) \;=\; 0$, we obtain that $F(y) \geq 0$ for all y. Thus, $(y-a)^2 \;\geq\; (1/2)y^2 - a^2$. This yields

(**A**) $$\frac{1}{2}\,y^2 - a^2 \;\leq (y-a)^2 \;=\; y^2 - 2ay + a^2\,.$$

Next, we obtain from $(y-a)^2 \;\leq\; (y-a)^2 + (y+a)^2$ that

(**B**) $$y^2 \;-\; 2ay \;+\; a^2 \;\leq\; (y^2 - 2ay + 2a^2) \;+\; (y^2 = 2ay + 2a^2) \;=\; 2y^2 \;+\; 2a^2\,.$$

Let $\omega \in \Omega$ and $y := Y(\omega)$. We combine **(A)** and **(B)** and obtain $\frac{1}{2}\,y^2 - a^2 \;\leq\; (y-a)^2 \;\leq\; 2y^2 \;+\; 2a^2\,.$ Since ω is an arbitrary elment of Ω, we have the following inequality of random variables:

(**C**) $$\frac{1}{2}\,Y^2 \;-\; a^2 \;\leq (Y-a)^2 \;\leq\; 2Y^2 \;+\; 2a^2\,.$$

Taking expectations maintains inequalities. Since $E[(Y-a)^2] = h(a)$ and $E[Y]^2 = h(0)$,

(**D**) $$\frac{1}{2}\,h(0) \;-\; a^2 \;\leq h(a) \;\leq\; 2h(0) \;+\; 2a^2\,.$$

From this we see that either $\big[\,h(0) = \infty$ and in this case, $h(a) = \infty$ for all $a\,\big]$,
or $\big[\,h(0) < \infty$ and in this case, $h(a) < \infty$ for all $a\,\big]$.

Step II: We show that **(b)** holds if $h(y) \not\equiv \infty$ According to **Step I**, we may assume that $h(0) < \infty$, i.e., $E[(Y^2] < \infty$. We obtain from Lemma 11.1 that $|E[Y]| < \infty$. Thus,

$$\begin{aligned} h(a) &= E[(Y-a)^2] \;=\; E[(Y^2] \,-\, 2aE[Y] \,+\, a^2 \\ &= E[(Y^2] \,-\, (E[Y])^2 \,+\, \big(a^2 \,-\, 2aE[Y] \,+\, (E[Y])^2\big) \\ &= Var[Y] \,+\, (a - E[Y])^2 \end{aligned} \tag{E}$$

It follows that h attains a unique minimum in height of $Var[Y]$ at $a = E[Y]$. This concludes the proof of the lemma. ■

Theorem 11.19.

Assume that Y is a random variable and $\vec{X} = (X_1, \ldots, X_k)$ is a random vector on (Ω, P). Then, either $E\big[(Y - g \circ \vec{X})\big] = \infty$ for all real–valued functions $g : \mathbb{R}^k \to \mathbb{R}$ of k real arguments, or

$$E\left[\big(Y - E[Y \mid \vec{X}]\big)^2\right] \;\le\; E\left[(Y - g \circ \vec{X})^2\right],$$

for all such functions g. Further, this is a strict inequality if $E[Y \mid \vec{X}] \neq g \circ \vec{X}$.

Note that, as always, we consider equations and inequalities involving random variables to be true as long as they are satisfied on a set of probability 1.

PROOF: Let us fix $\vec{x} \in \mathbb{R}^k$ for which $E[Y \mid \vec{X} = \vec{x}]$ is defined.

- **(a)** In the case of discrete Y and $\vec{X}$ this means that $p_{\vec{X}}(\vec{x}) > 0$ and then $B \mapsto \sum_{y \in B} p_{y|\vec{X}}(y \mid \vec{x})$ is a probability measure $P_{\vec{x}}$ on Ω for which we denote expectations by $E_{\vec{x}}[\ldots]$. Further, for $\psi : \mathbb{R} \to \mathbb{R}$, $E[\psi(Y) \mid \vec{X} = \vec{x}] \;=\; E_{\vec{x}}[\psi(Y)]$
- **(b)** For continuous Y and $\vec{X}$ this means that $f_{\vec{X}}(\vec{x}) > 0$. We have seen in Remark 11.11 on p.267 that $B \mapsto \int_B f_{y|\vec{X}}(y \mid \vec{x})dy$ is a probability measure $P_{\vec{x}}$ on Ω for which we denote expectations by $E_{\vec{x}}[\ldots]$. Further, for $\psi : \mathbb{R} \to \mathbb{R}$, $E[\psi(Y) \mid \vec{X} = \vec{x}] \;=\; E_{\vec{x}}[\psi(Y)]$
- **(c)** Thus, in both cases, all we have learned about ordinary expectations can be applied, for fixed $\vec{x}$, to the conditional expectations $E[\cdots \mid \vec{X} = \vec{x}]$.
- **(d)** When we condition an expression on $\vec{X} = \vec{x}$, we can replace in that expression all occurrences of $\vec{X}$ with $\vec{x}$.

It follows from **(d)** that

$$E\left[\big(Y \,-\, g(\vec{X})\big)^2 \mid \vec{X} = \vec{x}\right] \;= E\left[\big(Y \,-\, g(\vec{x})\big)^2 \mid \vec{X} = \vec{x}\right]. \tag{A}$$

We can apply Lemma 11.2 with $E_{\vec{x}}(\ldots)$ instead of $E(\ldots)$ and the constant $g(\vec{x})$ instead of a and conclude that

$$E\left[\big(Y \,-\, g(\vec{x})\big)^2 \mid \vec{X} = \vec{x}\right] \;\ge E\left[\big(Y \,-\, E[Y \mid \vec{X}]\big)^2 \mid \vec{X} = \vec{x}\right]. \tag{B}$$

We apply both **(A)** and **(B)** and evaluate both sides of the resulting inequality for $\vec{x} = \vec{X}(\omega)$:

$$E\left[\big(Y \,-\, g(\vec{X})\big)^2 \mid \vec{X} = \vec{X}(\omega)\right] \;\ge E\left[\big(Y \,-\, E[Y \mid \vec{X}]\big)^2 \mid \vec{X} = \vec{X}(\omega)\right].$$

As we have done before, we streamline this expression by replacing $\vec{X} = \vec{X}(\omega)$ with X:

$$E\left[\left(Y - g(\vec{X})\right)^2 \mid \vec{X}\right] \geq E\left[\left(Y - E[Y \mid \vec{X}]\right)^2 \mid \vec{X}\right].$$

Taking expectations on both sides, we obtain

$$E\left[\left(Y - g(\vec{X})\right)^2\right] \geq E\left[\left(Y - E[Y \mid \vec{X}]\right)^2\right].$$

We have shown the inequality that was asserted in the theorem.

We still must prove that this inequality is strict if $E[Y \mid \vec{X}] \neq g \circ \vec{X}$. To do so we apply the reasoning above to formula **(E)** of Lemma 11.2 and obtain

$$E\left[\left(Y - g(\vec{X})\right)^2\right] = Var[Y] + E\left[\left(g(\vec{X}) - E[Y \mid \vec{X}]\right)^2\right].$$

Since $E\left[\left(g(\vec{X}) - E[Y \mid \vec{X}]\right)^2\right] > 0$ unless $P\{g(\vec{X}) \neq E[Y \mid \vec{X}]\} = 0$, the assertion at the end of the theorem follows. ■

The last theorem can be phrased as follows:

We interpret random variables of the form $g(\vec{X})$, where $\vec{x} \mapsto g(\vec{x})$ is a (deterministic) function of $\vec{x}$, as those random variables that only use the information available to $\vec{X}$.
If we measure the quality of the approximation of a random variable Y by $g(\vec{X})$ as their mean squared distance, $E\left[\left(Y - g(\vec{X})\right)^2\right]$, then

- $E[Y \mid \vec{X}]$ is the best approximation of Y which is based only on information provided by $\vec{X}$.

11.8 The Multinomial Probability Distribution

Introduction 11.4. In Definition 7.3 (p.174) of Chapter 7 (Combinatorial Analysis) we discussed multinomial coefficients

$$\binom{n}{n_1\, n_2 \cdots n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

when counting the ways of classifying n items into k classes in such a way that n_1 items belong to class 1, n_2 items belong to class 2, ... n_k items belong to class k ($n_1 + \cdots + n_k = n$). The multinomial probability distribution is based on those coefficients and generalizes the binomial distribution of Section 9.2 (Bernoulli Variables and the Binomial Distribution).

The binomial distribution is that of a random variable Y which counts the number of successes in n Bernoulli trials. (See Definition 9.4 on p.197 about Bernoulli trials.) To say this differently, Y counts the number of those Bernoulli trials which result in an outcome that falls into the "success class".

The multinomial distribution will not be about a single random variable Y, but about a random vector $\vec{Y} = (Y_1, \ldots, Y_k)$ of k random variables Y_j, which count the number of the n trials resulting in an outcome that falls into class j. What kind of trials are we talking about? We should expect those n random elements, let us call them $X_1, \ldots, X_n$, to show some similarities to Bernoulli trials. Of course, there must be some significant differences. For example, each X_i will not have two outcomes (success or failure), but k outcomes corresponding to the k classes. □

Definition 11.18 (Multinomial Sequence)**.**

Let $X_1, X_2, \dots$ be a finite or infinite sequence of random elements on a probability space (Ω, P) which take values in a set Ω'. We call this sequence a **multinomial sequence**, if the following are satisfied:

(1) The sequence is iid.

(2) There is some $k \in \mathbb{N}$ such that the outcome of each X_j is one of k distinct values $\omega'_1, \omega'_2, \dots, \omega'_k \in \Omega'$.

Since the X_j have identical distribution, there are probabilities $p_1, p_2, \dots, p_k$ such that

(3) $p_i := P\{X_j = \omega'_i\}$ is the same for all j and $p_1 + \cdots + p_k = 1$.

If we consider a finite multinomial sequence $X_1, X_2, \dots, X_n$, we adopt the WMS notation and speak of a **multinomial experiment** of size n wich consists of the **trials** X_j □

Definition 11.19 (Multinomial distribution)**.**

Assume that $\vec{Y} = \big(Y_1, Y_2, \dots, Y_k\big)$ is a vector of random variables which possesses the joint probability mass function

$$p_{\vec{Y}}(y_1, y_2, \dots, y_k) = \binom{n}{y_1, \cdots, y_k} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}, \tag{11.59}$$

subject to the following conditions:

- $p_j \geq 0$ for $j = 1, 2, \dots, k$ and $\sum_{j=1}^{k} p_j = 1$.
- $y_i = 0, 1, 2, \dots, n$ for $i = 1, 2, \dots, k$ and $\sum_{i=1}^{k} y_i = n$.

Then we say that the random variables Y_i have a **multinomial distribution** with parameters n and $\vec{p} = (p_1, p_2, \dots, p_k)$. □

Theorem 11.20.

Let $n \in \mathbb{N}$ and $X_1, \dots, X_n$ be a multinomial sequence of size n. Let $p_j := P\{X_i = \omega'_j\}$. (That probability is the same for all i, since the X_i have identical distribution.)
Let $\vec{Y} = \big(Y_1, \dots, Y_k\big)$ be a vector of k random variables, such that each Y_j equals the number of the n trials resulting in an outcome that falls into class j. In other words,

(A) *$Y_i(\omega) = y_i \Leftrightarrow X_j(\omega) = \omega'_i$ for exactly y_i of the multinomial items X_j.*

Then $\vec{Y}$ has a multinomial distribution with parameters n and $p_{\vec{Y}}(y_1, y_2, \dots, y_k)$.

PROOF: For fixed $\vec{y} = \big(y_1, \dots, y_k\big)$, the event $A := \{\vec{Y} = \vec{y}\}$ corresponds to all different ways that $\{1, 2, \dots, n\}$ can be partitioned into k subsets

(A) $$\{1, 2, \dots, n\} = J_1 \uplus J_2 \uplus \cdots \uplus J_k$$

such that each J_i contains y_i of those n indices. It follows from Theorem 7.6 on p.175 that

(**B**) there are $\binom{n}{y_1, y_2, \dots y_k}$ different ways of creating such a partition.

Thus, if we write

$$A(J_1, \dots, J_k) \ := \{X_{i_{m,1}} = \cdots = X_{i_{m,y_m}} = \omega'_m \text{ for all } 1 \le m \le k\}\,,$$

it follows that

(**C**) $$P(A) \ = P\{\vec{Y} = \vec{y}\} \ = \ P\left(\biguplus A(J_1, \dots, J_k)\right)\,,$$

where this union is taken over all $\binom{n}{y_1, \dots y_k}$ partitions $J_1, \dots, J_k$ of $[1, n]_{\mathbb{Z}}$.
For a fixed $1 \le m \le k$, we write $J_m = \{i_{m,1} < i_{m,2} < \cdots < i_{m,y_m}\}$. Since the X_j are independent,

$$P\{X_{i_{m,1}} = X_{i_{m,2}} = \cdot = X_{i_{m,y_m}} = \omega'_m\} \ = P\big(\{X_{i_{m,1}} = \omega'_m\} \cap \cdots \cap \{X_{i_{m,y_m}} = \omega'_m\} \ = \ (p_m)^{y_m}$$

Since the X_j are independent not only for indices j belonging to J_m, but also across all J_m, it follows from the definition of $A(J_1, \dots, J_k)$ that

(**D**) $$P\big(A(J_1, \dots, J_k)\big) \ = (p_1)^{y_1}\,(p_2)^{y_2} \cdots (p_k)^{y_k}\,.$$

The right–hand side is independent of the particular partition $J_1, \dots, J_k$. We obtain from **(B)**, **(C)** and **(D)** that

$$P\{\vec{Y} = \vec{y}\} \ = \binom{n}{y_1, \cdots, y_k} (p_1)^{y_1}\,(p_2)^{y_2} \cdots (p_k)^{y_k}\,.$$

Thus, $\vec{Y}$ has the joint PMF that was specified in (11.59). We conclude that $\vec{Y}$ has a multinomial distribution with parameters n and $p_{\vec{Y}}(y_1, y_2, \dots, y_k)$. ■

Example 11.6. Research by the marketing division of GreatWidgets Corp. has established that their customers' age is distributed as shown in the table to the right. A random sample of eight customers is taken. Assume that the proportions shown accurately reflect those of GreatWidgets Corp.

Age	Proportion
Group 1: $15 - 20$	0.2
Group 2: $21 - 30$	0.2
Group 3: $31 - 40$	0.1
Group 4: $41 - 50$	0.2
Group 5: > 50	0.3

what is the probability that the sample is composed as follows:
• Group 1: 1 person • Group 2: 3 persons • Group 4: 2 persons • Group 5: 2 persons?

Solution:
We interpret the sample picks as the members $X_1, \dots, X_8$ of a multinomial sequence each of which has age group k as an outcome with probability p_k as indicated in the table.
Then the probability we are looking for is given by (11.59) on p.274

$$p_{\vec{y}}(y_1, y_2, y_3, y_4, y_5) \ = \ \frac{n!}{y1!\,y2!\,y3!\,y4!\,y5!}\,p_1^{y_1}\,p_2^{y_2}\,p_3^{y_3}\,p_4^{y_4}\,p_5^{y_5}\,, \ \square$$

In the context of this example we obtain

$$p(1, 3, 2, 0, 2) \ = \frac{8!}{1!\,3!\,2!\,0!\,2!}\,0.2^1\,0.2^2\,0.1^0\,0.2^2\,0.3^0\,, \ = \ 0.009768. \ \square$$

Theorem 11.21 (WMS Ch.05.9, Theorem 5.13)**.**

Assume that the random vector $\vec{Y} = (Y_1, Y_2, \ldots, Y_k)$ follows a multinomial distribution with parameters n and $\vec{p} = (p_1, p_2, \ldots, p_k)$. Then, for $i, i' \in [1,k]_{\mathbb{Z}}$ and $q_i = 1 - p_i$,

(a) $E[Y_i] = np_i$ ***(b)*** $Var[Y_i] = np_iq_i$ ***(c)*** *If $i \neq i'$, then* $Cov[Y_i, Y_{i'}] = -np_ip_{i'}$

PROOF: We may assume that there is an underlying multinomial sequence $(X_j)_j$ and distinct items $\omega'_1, \ldots, \omega'_k$ such that

(1) $P\{X_j = \omega'_i\} = p_i$ for each j.

(2) $Y_i(\omega)$ counts the number of indices j such that $X_j(\omega) = \omega'_i$.

Clearly, $Y_i \sim$ binom(n, p_i). Thus, $E[Y_i] = np_i$ and $Var[Y_i] = np_iq_i$. We have shown **(a)** and **(b)**.

For the proof of **(c)**, we associate with the (fixed and different!) indices i and i' two 0–1 encoded binomial sequences $(Z_j)_j$ and $(Z'_j)_j$ as follows:

- $Z_j(\omega) = 1$ if $X_j(\omega) = \omega'_i$ and 0 else,
- $Z'_j(\omega) = 1$ if $X_j(\omega) = \omega'_{i'}$ and 0 else.

Then,

- $\sum_{j=1}^{n} Z_j(\omega) =$ count of indices j such that $X_j(\omega) = \omega'_i = Y_i(\omega)$. See **(2)**.
- $\sum_{j=1}^{n} Z'_j(\omega) =$ count of indices j such that $X_j(\omega) = \omega'_{i'} = Y_{i'}(\omega)$. See **(2)**.

It follows that

(3) $Cov[Y_i, Y_{i'}] = Cov\left[\sum_{j=1}^{n} Z_j, \sum_{m=1}^{n} Z'_m\right]$.

This is easily computed by employing the formulas of Theorem 11.13 (WMS Ch.05.8, Theorem 5.12) on p.259 and the following:

Since $Z_j \sim$ binom$(1, p_i)$ and $Z'_m \sim$ binom$(1, p'_i)$ for each j and m,

(4) $E[Z_j] = p_i$ and $E[Z'_m] = p'_i$ for each j and each m.

Let $j \in [1,n]_{\mathbb{Z}}$ and $\omega \in \Omega$. Since $i \neq i'$ implies $\omega'_i \neq \omega'_{i'}$, at least one of $Z_j(\omega), Z'_j(\omega)$ is zero.

(5) Thus, $E[Z_jZ'_j] = 0$ for all j.

It follows from **(4)** and **(5)** that

(6) $Cov[Z_j, Z'_j] = E[Z_jZ'_j] - E[Z_j] \cdot E[Z'_j] = -p_jp'_j$.

Let $j, m \in [1,n]_{\mathbb{Z}}$ and $j \neq m$. Since $X_1, \ldots, X_n$ are independent and Z_j only depends on X_j and Z'_m only depends on X_m, the random variables Z_j and Z'_m also are independent.

(7) Thus, $Cov[Z_j, Z'_m] = 0$ for all $j, m \in [1,n]_{\mathbb{Z}}$ such that $j \neq m$.

We apply (11.42) on p.260 and obtain

$$Cov[Y_i, Y_{i'}] \overset{(3)}{=} Cov\left[\sum_{j=1}^{n} Z_j, \sum_{m=1}^{n} Z'_m\right] = \sum_{j=1}^{n}\sum_{m=1}^{n} Cov[Z_j, Z'_m]$$
$$= \sum_{j=1}^{n} Cov[Z_j, Z'_j] + \sum\sum_{j\neq m} Cov[Z_j, Z'_m]$$
$$\overset{(6),(7)}{=} \sum_{j=1}^{n}(-p_i p'_i) + \sum\sum_{j\neq m} 0 = -np_i p'_i\,. \quad \blacksquare$$

Note that it makes perfect sense for $Cov[Y_i, Y_{i'}]$ to be negative if $i \neq i'$: If a large proportion of the X_j have the outcome ω'_i, then fewer trials remain to take one of the other values.

11.9 Order Statistics


@@Author

The presentation of the material in this section is largely based on the 2015 Math 447 lecture notes of Prof. Xingye Qiao, Binghamton University

Given are n random variables $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$. One can sort them, for any fixed $\omega \in \Omega$, in nondecreasing order. One obtains in this fashion a sequence, of size n, of numbers

$$Y_{(1)}(\omega) \leq Y_{(2)}(\omega) \leq Y_{(3)}(\omega) \leq \cdots \leq Y_{(n)}(\omega)\,.$$

Since these numbers depend on randomness ω, each $Y_{(j)}(\omega)$ represents an outcome of a random variable $Y_{(j)}$.

Example 11.7. Here are some examples.

(a) 70 students are randomly selected when exiting lecture hall and their age is rounded to the closest year. Those 70 ages, $A_1(\omega), \ldots, A_{70}(\omega)$, are sorted in increasing order:

- $A_{(1)}(\omega)$ = height of the smallest person in the sample
- $A_{(2)}(\omega)$ = height of the second smallest person in the sample
- -
- $A_{(j)}(\omega)$ = height of the jth smallest person in the sample
- -
- $A_{(n)}(\omega)$ = height of the tallest person in the sample

Clearly, $A_{(1)}(\omega) \leq A_{(2)}(\omega) \leq A_{(3)}(\omega) \leq \cdots \leq A_{(n)}(\omega)\,.$

Almost all of those ages will be one of $18, 19, .., 25$. Accordingly, it is not only possible that we encounter an index j that results in equality, $A_{(j)}(\omega) = A_{(j+1)}(\omega)$, but this will be the rule rather than the exception.

(b) Rather than considering the age of those 70 students, we now look at their height, measured in millimeters. Those 70 heights, $H_1(\omega), \ldots, H_{70}(\omega)$, are sorted in increasing order.

Height can be considered a continuous random variable. Since the probability of two students having precisely the same height is zero, we may consider the outcomes $H_{(j)}$ distinct. Accordingly, we can replace "less or equal" with strict inequality and obtain

$$H_{(1)}(\omega) < H_{(2)}(\omega) < H_{(3)}(\omega) < \cdots < H_{(n)}(\omega)\,. \quad \square$$

- We will deal in this section exclusively with continuous random variables.
- When considering a finite or infinite sequence $Y_1, Y_2, Y_3, \dots$ of such random variables, we assume that they are iid (independent and identically distributed).

Definition 11.20 (Order statistics)**.**

Given n iid continuous random variables $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$, we sort them in inreasing order. The resulting sequence of random variables, which we denote as $Y_{(j)}, j = 1, \dots, n$, then satisfies, for each $(omega \in Omega$,

$$Y_{(1)}(\omega) \le Y_{(2)}(\omega) \le Y_{(3)}(\omega) \le \cdots \le Y_{(n)}(\omega)\,. \tag{11.60}$$

We call $Y_{(j)}$ the **jth order statistic** of $\vec{Y}$.

See Example 11.7**(b)** why we may consider strictly increasing rather than nondecreasing. □

Assumption 11.4.

Unless explicitly stated otherwise,
- $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ denotes a list of n iid continuous random variables $(n \in \mathbb{N})$.
- $Y_1 \sim Y_2 \sim \cdots \sim Y_n$ implies $F_{Y_1} = F_{Y_2} = \cdots = F_{Y_n}$ and $f_{Y_1} = f_{Y_2} = \cdots = f_{Y_n}$
  We write $F(y) := F_{Y_j}(y)$ and $f(y) := f_{Y_j}(y)$ for the common CDF and PDF. □

Remark 11.13. Note that
- The **first order statistic** or **smallest order statistic** is $Y_{(1)} = \min\{Y_1, \dots, Y_n\}$.
- The nth order statistic or **largest order statistic** is $Y_{(n)} = \max\{Y_1, \dots, Y_n\}$.
- A simple consequence of the definition of min and max are the following formulas:

$$Y_{(1)}(\omega) > y \Leftrightarrow \min\big(Y_1(\omega), \dots, Y_n(\omega)\big) > y \Leftrightarrow Y_j(\omega) > y \text{ for all } j \in [1,n]_{\mathbb{Z}}\,, \tag{11.61}$$

$$Y_{(n)}(\omega) \le y \Leftrightarrow \max\big(Y_1(\omega), \dots, Y_n(\omega)\big) \le y \Leftrightarrow Y_j(\omega) \le y \text{ for all } j \in [1,n]_{\mathbb{Z}}\,. \tag{11.62}$$

□

Theorem 11.22 (CDF and PDF of the jth order statistic)**.**

For $y \in \mathbb{R}$, the CDF of the kth order statistic ($k = 1, \dots, n$) satisfies the following:

$$F_{Y_{(1)}(y)} = 1 - [1 - F(y)]^n\,, \tag{11.63}$$

$$F_{Y_{(n)}(y)} = [F(y)]^n\,, \tag{11.64}$$

$$F_{Y_{(k)}(y)} = 1 - \sum_{j=0}^{k-1} \binom{n}{j} [F(y)]^j\,[1 - F(y)]^{n-j} = \sum_{j=k}^{n} \binom{n}{j} [F(y)]^j\,[1 - F(y)]^{n-j}\,. \tag{11.65}$$

For $y \in \mathbb{R}$, the PDF of the kth order statistic ($k = 1, \ldots, n$) satisfies the following:

$$f_{Y_{(1)}(y)} = n\,[1 - F(y)]^{n-1}\,f(y)\,, \tag{11.66}$$

$$f_{Y_{(n)}(y)} = n\,[F(y)]^{n-1}\,f(y)\,, \tag{11.67}$$

$$f_{Y_{(k)}(y)} = \sum_{j=0}^{k-1} \binom{n}{j} f(y) \left(n\,[F(y)]^{n-1} - j[F(y)]^{j-1} \right). \tag{11.68}$$

$$f_{Y_{(k)}(y)} = n \binom{n-1}{k-1} f(y) \cdot \big[\, F(y) \,\big]^{k-1} \cdot \big[\, 1 - F(y) \,\big]^{n-k}. \tag{11.69}$$

Note that the proofs are not given in the order of the seven formulas of the theorem.

PROOF of (11.64):

$$F_{Y_{(n)}(y)} \overset{(11.62)}{=} P\big(\{Y_1 \le y\} \cap \{Y_2 \le y\} \cap \cdots \cap \{Y_n \le y\}\big)$$
$$\overset{\textbf{indep}}{=} P\{Y_1 \le y\} \cdot P\{Y_2 \le y\} \cdots P\{Y_n \le y\} \;=\; [F(y)]^n\,.$$

PROOF of (11.63):

$$P\{Y_{(1)} > y\} \overset{(11.61)}{=} P\big(\{Y_1 > y\} \cap \{Y_2 > y\} \cap \cdots \cap \{Y_n > y\}\big)$$
$$\overset{\textbf{indep}}{=} P\{Y_1 > y\} \cdot P\{Y_2 > y\} \cdots P\{Y_n > y\} \;=\; [1 - F(y)]^n\,.$$

Thus, $F_{Y_{(1)}(y)} = 1 - P\{Y_{(1)} > y\} = 1 - [1 - F(y)]^n$.

PROOF of (11.66) and (11.67):

This follows from $\dfrac{d}{dy}\left(1 - [1 - F(y)]^n\right) = -n[1 - F(y)]^{n-1}\big(-f(y)\big)$

and $\dfrac{d}{dy}\left([F(y)]^n\right) = n[F(y)]^{n-1}\,f(y)$.

PROOF of (11.65):

This proof requires a lot more work than the proofs we have done so far. It will be done by constructing a binomial random variable.

- Since y is fixed, so is $p := F(y) = P\{Y_j \le y\}$. (Identical for all j, since the Y_j are iid.)
- For $j = 1, \ldots, n$, let $X_j(\omega) := \begin{cases} 1 & \text{if } Y_j(\omega) \le y\,, \\ 0 & \text{else}\,. \end{cases}$ Let $U(\omega) := \sum_{j=1}^{n} X_j(\omega)$.
- We interpret $Y_j(\omega) \le y$ as a success and $Y_j(\omega) > y$ as a failure. Then $X_1, \ldots, X_n$ form a 0–1 encoded Bernoulli sequence [109] and $U \sim \text{binom}(n, p)$, since U counts the number of successes.
- Observe that $Y_{(k)}(\omega) \le y \;\Leftrightarrow\; Y_j(\omega) \le y$ at least k times $\;\Leftrightarrow\;$ there are at least k successes $\;\Leftrightarrow\; U(\omega) \ge k$. It does not matter whether or not there are more than k successes.
- Thus, $F_{Y_{(k)}}(y) = P\{Y_{(k)} \le y\} = P\{U \ge k\} = \sum_{j=k}^{n} P\{U = j\} = 1 - \sum_{j=0}^{k-1} P\{U = j\}$.
- Since $U \sim \text{binom}(n, p)$ and $p = F(y)$, $F_{Y_{(k)}}(y) = 1 - \sum_{j=0}^{k-1} \binom{n}{j} [F(y)]^j\,[1 - F(y)]^{n-j}$.

[109] See Definition 9.4 (Bernoulli trials and variables) on p.197.

PROOF of (11.68):

This is done by differentiation. For each $j = 0, \ldots, k-1$,

$$\textbf{(A)} \qquad \begin{aligned} \frac{d}{dy}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j} &= \binom{n}{j}\frac{d}{dy}\left([F(y)]^j \;-\; F(y)]^n\right) \\ &= \binom{n}{j}\;\left(\,j\,[F(y)]^{j-1}\,f(y) \;-\; n\,F(y)]^{n-1}\,f(y)\,\right) \end{aligned}$$

$$\begin{aligned} \text{Thus,} \qquad f_{Y_{(k)}} &= \frac{d}{dy}\left[\,1 \;-\; \sum_{j=0}^{k-1}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j}\right] \\ &= -\sum_{j=0}^{k-1}\frac{d}{dy}\binom{n}{j}\left(\,[F(y)]^j\,[1-F(y)]^{n-j}\,\right) \\ &\overset{\textbf{(A)}}{=} \sum_{j=0}^{k-1}\binom{n}{j}\,f(y)\;\left(\,n\,[F(y)]^{n-1} \;-\; j\,F(y)]^{j-1}\,\right)\,. \end{aligned}$$

This finishes the proof of (11.68).

The proof of (11.69) is based on an entirely different approach. Before we do that proof, we first illustrate that approach by redoing those of (11.66) and (11.67). Those proofs are much simpler and are a good preparation for that of (11.69).

ALTERNATE PROOF of (11.66):

First, we note the following for a continuous random variable U with density $f_U(u)$ Assume that $\delta > 0$ is very close to zero. Since we assumed for all our continuous random variables that they have continuous density, $f_U(\cdot) \approx \text{const} = f_U(u)$ on $]u-\delta, u+\delta[$.

(a) Thus, $P\{u < U \le u+\delta\} \;=\; \int_u^{u+\delta} f_U(t)\,dt \;\approx\; f_U(u)\cdot\delta$.

(b) For the fixed y and some "really small" δy, we create three events:
$\boxdot$ L (for "left–hand side") $\boxdot$ I (for "inside") $\boxdot$ R (for "right–hand side"),
and a sequence of random elements $X_1, \ldots, X_n$ as follows.
$\boxdot$ $X_j(\omega) = L \;\Leftrightarrow\; Y_j(\omega) \le y$. Then $P\{X_j = L\} \;=\; P\{Y_j \le y\} \;=\; F(y)$.
$\boxdot$ $X_j(\omega) = I \;\Leftrightarrow\; y < Y_j(\omega) \le y+\delta$, Then $P\{X_j = I\} \;=\; P\{y < Y_j \le y+\delta\} \overset{\textbf{(a)}}{\approx} f_U(u)\cdot\delta$.
$\boxdot$ $X_j(\omega) = R \;\Leftrightarrow\; Y_j(\omega) > y+\delta$. Then $P\{X_j = R\} \;=\; P\{Y_j > y+\delta\} \;=\; 1 - F(y+\delta)$.

(c) By construction, the X_j form a multinomial sequence. Let $\vec{U} := \big(U_1, U_2, U_3\big)$, where
$\boxdot$ $U_1 :=$ # of indices j such that $X_j = L$,
$\boxdot$ $U_2 :=$ # of indices j such that $X_j = I$,
$\boxdot$ $U_3 :=$ # of indices j such that $X_j = R$.

(d) Then $\vec{U}$ is multinomial with parameters $n,\;\; p_1 = F(y),\;\; p_2 = f(y)\delta,\;\; p_3 = 1 - F(y+\delta)$.

(e) Since we assume that $Y_{(j)}(\omega)$ is strictly increasing with j for all ω, it seems reasonable that, for "really small" δ, the following is true:

- If $Y_{(1)}(\omega) > y$, then $Y_{(j)}(\omega) > y+\delta$ for all $j > 1$.

(f) Thus, $f_{Y_{(1)}}(y)\cdot\delta \overset{\text{(a)}}{\approx} P\{\, y < Y_{(1)} \le y+\delta \,\}$
$= P\{$ exactly one of $Y_1,\ldots,Y_n \in]y, y+\delta]$ and $Y_j > y+\delta$ for all other $j\,\}$.
$= P\{$ none of the X_j are L and exactly one is I and $n-1$ are $R\,\}$.
$= P\{U_1=0,\, U_2=1,\, U_3=n-1,\,\} \overset{\text{(d)}}{=} \binom{n}{0,1,n-1} [F(y)]^0 \left[f(y)\delta\right]^1 [1-F(y+\delta)]^{n-1}$.

(g) Since $\binom{n}{0,1,n-1} = \dfrac{n!}{0!\cdot 1!\cdot (n-1)!} = n$,
we obtain $f_{Y_{(1)}}(y)\cdot\delta \approx n\,[1-F(y+\delta)]^{n-1} f(y)\delta$.

We divide both expressions by δ, then let $\delta\to 0$. Since $t\mapsto F(t)$ is continuous, $\lim\limits_{\delta\to 0} F(y+\delta) = F(y)$. We conclude that the density of $Y_{(1)}$ is

$$f_{Y_{(1)}}(y) \;=\; n\,[1-F(y)]^{n-1}\, f(y)\,.$$

ALTERNATE PROOF of (11.67):
We can adapt the alternate proof for the density of $Y_{(1)}$ to obtain that of $Y_{(n)}$ as follows.
We keep all items through **(e)** and modify **(f)** and **(g)** as follows.

(f′) $f_{Y_{(n)}}(y)\cdot\delta \overset{\text{(a)}}{\approx} P\{\, y < Y_{(n)} \le y+\delta \,\}$
$= P\{$ exactly one of $Y_1,\ldots,Y_n \in]y, y+\delta]$ and $Y_j \le y$ for all other $j\,\}$.
$= P\{$ none of the X_j are R and exactly one is I and $n-1$ are $L\,\}$.
$= P\{U_1=n-1,\, U_2=1,\, U_3=0,\,\} \overset{\text{(d)}}{=} \binom{n}{n-1,1,0} [F(y)]^{n-1} \left[f(y)\delta\right]^1 [1-F(y+\delta)]^0$.

(g′) Since $\binom{n}{n-1,1,0} = \dfrac{n!}{(n-1)!\cdot 1!\cdot 0!} = n$,
we obtain $f_{Y_{(n)}}(y)\cdot\delta \approx n\,[F(y)]^{n-1} f(y)\delta$.

We divide both expressions by δ, then let $\delta\to 0$. We obtain the density of $Y_{(n)}$ as

$$f_{Y_{(n)}}(y) \;=\; n\,[F(y)]^{n-1}\, f(y)\,.$$

PROOF of (11.69):
This time we adapt the alternate proof for the density of $Y_{(1)}$ to obtain that of $Y_{(k)}$ as follows.
We keep all items through **(e)** and modify **(f)** and **(g)** as follows.

(f″) $f_{Y_{(k)}}(y)\cdot\delta \overset{\text{(a)}}{\approx} P\{\, y < Y_{(k)} \le y+\delta \,\}$
$= P\{$ exactly one of $Y_1,\ldots,Y_n \in]y, y+\delta]$ and $Y_j \le y$ for $k-1$ indices j
and $Y_j > y$ for $n-k$ indices $j\,\}$
$= P\{k-1$ of the X_j are $L, n-k$ of the X_j are R and exactly one is $I\,\}$
$= P\{U_1=k-1,\, U_2=1,\, U_3=n-k\}$
$\overset{\text{(d)}}{=} \binom{n}{k-1,1,n-k} [F(y)]^{k-1} \left[f(y)\delta\right]^1 [1-F(y+\delta))]^{n-k}$.

(g") Since $\binom{n}{k-1,1,n-k} = \frac{n\cdot(n-1)!}{(k-1)!\cdot 1!\cdot(n-k)!} = n\cdot\binom{n-1}{k-1}$,
we obtain $f_{Y_{(k}}(y)\cdot\delta \approx n\cdot\binom{n-1}{k-1}\,[F(y)]^{k-1}\,f(y)\delta\,[1-F(y+\delta)]^{n-k}$.

We divide both expressions by δ, then let $\delta\to 0$. Since $t\mapsto F(t)$ is continuous, $\lim_{\delta\to 0} F(y+\delta) = F(y)$. We conclude that the density of $Y_{(1)}$ is

$$f_{Y_{(k)}}(y) \;=\; n\binom{n-1}{k-1}\,[F(y)]^{k-1}\,f(y)\,[1-F(y)]^{n-k}\,. \;\blacksquare$$

Remark 11.14. (11.65) yields (11.63) for $k=1$ and (11.64) for $k=n$. This can be seen as follows: Recall that

(A)
$$\begin{aligned} 1 \;&=\big(F(y)\;+\;[1-F(y)]\big)^n \;=\; \sum_{j=0}^{n}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j} \\ &= \sum_{j=0}^{n-1}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j} \;+\; \binom{n}{0}[F(y)]^0\,[1-F(y)]^n\,. \end{aligned}$$

If we evaluate (11.65) for $k=1$ and $k=n$, we obtain

$$F_{Y_{(1)}(y)} \;=1\;-\;\binom{n}{0}[F(y)]^0\,[1-F(y)]^n \;=\; 1\;-\;1\cdot 1\cdot[1-F(y)]^n \;=\; [1-F(y)]^n\,,$$

$$F_{Y_{(n)}(y)} \;=1\;-\;\sum_{j=0}^{n-1}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j} \overset{\text{(A)}}{=} \binom{n}{0}[F(y)]^0\,[1-F(y)]^n \;=\; [1-F(y)]^n\,. \;\square$$

Remark 11.15. You may have noticed that there are two formulas for $f_{Y_{(k)}(y)}$.
(11.69) was shown by means of the "density approach" that utilized a limiting process $\delta\to 0$ in conjunction with the multinomial distribution. The proof was harder than that of (11.68). In return, (11.69) has computational advantages, since no more summation $\sum_{j=0}^{k-1}$ is required. $\square$

The next remark belongs thematically into Section 7.2 (Permutations) of Chapter 7. However, it has been placed here, since every order statistic

$$\vec{Y}_{(\bullet)} \;=\; \big(Y_{(1)},\ldots,Y_{(n)}\big).$$

is a (specific) permutation of $\vec{Y}=\big(Y_1,\ldots,Y_n\big)$, and every other permutation

$$\big(Y_{i_1},Y_{i_2},\ldots,Y_{i_n}\big)$$

of $\vec{Y}=\big(Y_1,\ldots,Y_n\big)$, possesses the same order statistic.

Remark 11.16. If we deal with a list $\vec{a}=(a_1,a_2,\ldots,a_n)$ of distinct numbers, e.g.,

(A) $$\vec{a} \;=\; (13.2,\,-3,\,6.6,\,2,\,-1.5)\,,$$

then there is a uniquely determined permutation, $\vec{a}_{(\bullet)} = (a_{(1)}, a_{(2)}, \ldots, a_{(n)})$ of $\vec{a}$, which has those a_j in increasing order. In other words,

$$a_{(1)} \;<\; a_{(2)} \;<\cdots<\; a_{(n)}\,.$$

In the specific example **(A)**, we obtain

$$\vec{a}_{(\bullet)} \;=\; (-3,\, -1.5,\, 2,\, 6.6,\, 13.2)\,.$$

If $\vec{b} = (b_1, b_2, \ldots, b_n)$ is another list of distinct numbers, then

$$\vec{b}_{(\bullet)} \;=\; \vec{a}_{(\bullet)} \qquad \Leftrightarrow \qquad \vec{b} \text{ is a permutation of } \vec{a}\,.$$

Going back to our example, if

$$\begin{aligned}\vec{b} &= (13.2,\, 6.6,\, -1.5,\, -3,\, 2)\,,\\ \vec{c} &= (13.2,\, -3,\, 6.6,\, 2,\, -1.51)\,,\end{aligned}$$

then $\vec{b}_{(\bullet)} = \vec{a}_{(\bullet)}$, but $\vec{c}_{(\bullet)} \neq \vec{a}_{(\bullet)}$, since $\vec{a}_{(\bullet)}$ does not include the number -1.51. □

Theorem 11.23 (Joint PDF of the order statistic)**.**

A: *Let $\vec{y} \in \mathbb{R}^n$ satisfy*

(11.70) $$y_1 \;<\; y_2 \;<\cdots<\; y_n\,.$$

For the vector $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$, let $\vec{Y}_{(\bullet)}$ be the vector of its associated order statistics, i.e.,

(11.71) $$\vec{Y}_{(\bullet)} \;=\; \big(Y_{(1)}, \ldots, Y_{(n)}\big).$$

Then its density function at $\vec{y}$ is given by

(11.72) $$f_{\vec{Y}_{(\bullet)}}(\vec{y}) \;=\; n! \cdot \prod_{j=1}^{n} f(y_j) \;=\; n!\, f(y_1)\cdots f(y_n)\,.$$

B: *If $\vec{y}$ does not satisfy (11.70), then $f_{\vec{Y}_{(\bullet)}}(\vec{y}) \;=\; 0$.*

FIRST PROOF:

Let Δ be a "small" n–dimensional cube with volume $Vol(\Delta)$ that is centered at $\vec{y}$. Study the proof of (11.65) of Theorem 11.22 on p.278. It explains (in the onedimensional case), why one can approximate

$$\begin{aligned}P\{\vec{Y} \in \Delta\} \;&\approx\; f_{\vec{Y}}(\vec{y}) \cdot Vol(\Delta)\,,\\ P\{\vec{Y}_{(\bullet)} \in \Delta\} \;&\approx\; f_{\vec{Y}_{(\bullet)}}(\vec{y}) \cdot Vol(\Delta)\,.\end{aligned}$$

A cube of sidelength 2ε has volume $Vol(\Delta) = (2\varepsilon)^n$. If we solve that equation for ε, we obtain

$$\varepsilon \;=\; \frac{Vol(\Delta)^{1/n}}{2}\,.$$

Since $y_1 < y_2 < \cdots < y_n$, one can choose Δ and hence, $\varepsilon = Vol(\Delta)^{1/n}/2$, so small, that any two intervals $[\,y_i - \varepsilon, y_i + \varepsilon\,]$ and $[\,y_j - \varepsilon, y_j + \varepsilon\,]$ have empty intersection for $i \neq j$.

For the following, see Remark 11.16 on p.282. Note that

(A)
$$\begin{aligned}\vec{Y}_{(\bullet)}(\omega) \in \Delta \;&\Leftrightarrow y_k - \varepsilon \le Y_{(k)}(\omega) \le y_k + \varepsilon \text{ for all } k,\\ &\Leftrightarrow \text{for all } k,\ \exists\, j \text{ such that } y_k - \varepsilon \le Y_j(\omega) \le y_k + \varepsilon\,.\end{aligned}$$

We illustrate this point for $n = 3$, $Vol(\Delta) = 1/8$, $y_1 = 2.6, y_2 = 4.2, y_3 = 7.8$. $\varepsilon = (1/8^3)/2 = 0.25$. This is small enough for the intervals $y_j \pm 0.25$ to be disjoint.

There are $3! = 6$ different ways to have $\vec{Y}(\omega) \in \Delta$. They are:

(1) $2.35 \le Y_1(\omega) \le 2.85,\ 3.95 \le Y_2(\omega) \le 4.45,\ 7.55 \le Y_3(\omega) \le 8.05,$
(2) $2.35 \le Y_1(\omega) \le 2.85,\ 3.95 \le Y_3(\omega) \le 4.45,\ 7.55 \le Y_2(\omega) \le 8.05,$
(3) $2.35 \le Y_2(\omega) \le 2.85,\ 3.95 \le Y_1(\omega) \le 4.45,\ 7.55 \le Y_3(\omega) \le 8.05,$
(4) $2.35 \le Y_2(\omega) \le 2.85,\ 3.95 \le Y_3(\omega) \le 4.45,\ 7.55 \le Y_1(\omega) \le 8.05,$
(5) $2.35 \le Y_3(\omega) \le 2.85,\ 3.95 \le Y_1(\omega) \le 4.45,\ 7.55 \le Y_2(\omega) \le 8.05,$
(6) $2.35 \le Y_3(\omega) \le 2.85,\ 3.95 \le Y_2(\omega) \le 4.45,\ 7.55 \le Y_1(\omega) \le 8.05,$

Let us assume that $k = 2$, i.e., we consider the interval $[3.95, 4.45]$.

In **(2)** and **(4)**, we choose $j = 3$ to obtain $Y_j \in [3.95, 4.45]$.

On the other hand, in **(1)** and **(6)**, we choose $j = 2$ to obtain $Y_j \in [3.95, 4.45]$.

We refer you again to Remark 11.16 on p.282 to understand that **(A)** shows that

(B)
$$\begin{aligned}\vec{Y}_{(\bullet)}(\omega) \in \Delta \;&\Leftrightarrow \textbf{some} \text{ permutation of } \vec{Y}(\omega) \in \Delta\\ &\Leftrightarrow \textbf{each} \text{ permutation of } \vec{Y}(\omega) \in \Delta\,.\end{aligned}$$

- Since a list of n items has $n!$ permutations, there are $n!$ such (disjoint) events: There are $n!$ permutations $(k_1, k_2, \ldots, k_n)$ of $(1, 2, \ldots, n)$ with corresponding event
$$\{y_1 - \varepsilon \le Y_{k_1} \le y_1 + \varepsilon\} \cap \{y_2 - \varepsilon \le Y_{k_2} \le y_2 + \varepsilon\} \cap \cdots \cap \{y_n - \varepsilon \le Y_{k_n} \le y_n + \varepsilon\}.$$
- Since the Y_j are iid and $P\{y_i - \varepsilon \le Y_{k_j} \le y_i + \varepsilon\} \approx 2\varepsilon \cdot f(y_i)$ for each i and j, each such event has probability $\approx \prod_{j=1}^n f(y_j) \cdot (2\varepsilon)^n$.
- Thus, $f_{\vec{Y}_{(\bullet)}}(\vec{y}) \cdot Vol(\Delta) \approx n! \cdot \prod_{j=1}^n f(y_j) \cdot Vol(\Delta)$
- As $\Delta \to 0$, "$\approx$" becomes "$=$" and then $f_{\vec{Y}_{(\bullet)}}(\vec{y}) = n! \cdot \prod_{j=1}^n f(y_j)$. ■

ALTERNATE PROOF:

(a) We may assume that $\vec{y}$ satisfies $y_1 < y_2 < \cdots < y_n$, since $f_{\vec{Y}_{(\bullet)}}(\vec{y}) = 0$ otherwise.
- For small enough dt_1, dt_2, dt_n, the intervals $[y_j, y_j + d\,t_j]$ are disjoint.

(b) Thus, $\big[\, y_j \le Y_{(j)}(\omega) \le y_j + d\,t_j \text{ for all } j \,\big] \Leftrightarrow \big[$ there is a permutation $i_1, i_2, \ldots, i_n$ of the indices $1, 2, \ldots, n$ such that $y_j \le Y_{i_j}(\omega) \le y_j + d\,t_j$ for all j $\big]$

(c) Thus, $\big[\, y_j \le Y_{(j)}(\omega) \le y_j + d\,t_j \text{ for all } j \,\big] \;\Leftrightarrow\; [\,\text{among the } X_i(\omega), \text{exactly one is in } [y_1, y_1 + d\,t_1],$ exactly one is in $[y_2, y_2 + d\,t_2]$, ..., exactly one is in $[y_n, y_n + d\,t_n]$. (Thus, NONE are outside the union of those intervals.)

(d) This can be interpreted as the counts of the outcomes of a multinomial sequence $X_1, \ldots, X_n$, where $X_k(\omega)$ results in outcome $\#j$, if $y_j \le Y_k \le y_j + d\,t_j$.

- The probabilities $p_j = P\{X_k \text{ results in } \#j\}$ are, for small enough dt_j, equal to
$p_j = P\{Y_i \in [y_j, y + dt_j]\,\} = \int_{y_j}^{y+dt_j} f(t)\,dt \;\approx\; f(t_j)\,dt_j.$

(e) From **(b)**, **(c)**, **(d)**:

$$\begin{aligned} f_{\vec{Y}_{(\bullet)}}(\vec{y})\,dt_1 \cdots dt_n &= P\{y_j \le Y_{(j)}(\omega) \le y_j + d\,t_j \text{ for all } j\} \\ &= P\{\text{ there is a permutation } i_1, i_2, \ldots, i_n \text{ of the indices } 1, 2, \ldots, n \\ &\qquad \text{such that } y_j \le Y_{i_j} \le y_j + d\,t_j \text{ for all } j\} \\ &= P\{\text{ each } X_k \text{ has exactly one outcome } \#j \text{ for each } j = 1, \ldots, n\} \\ &= \binom{n}{1, 1, \ldots, 1} p_1^1 p_2^1 \cdots p_n^1 \;=\; \frac{n!}{1! \cdots 1!} \prod_j \big(f(t_j)\,dt_j\big)\,. \end{aligned}$$

Thus, $f_{\vec{Y}_{(\bullet)}}(\vec{y})\,dt_1 \cdots dt_n \;=\; n!\,\prod_j f(t_j)\,(dt_1 \cdots dt_n)$.

(f) We cancel $dt_1 \cdots dt_n$ on both sides and obtain $f_{\vec{Y}_{(\bullet)}}(\vec{y})\,dt_1 \;=\; n!\,\prod_j f(t_j)$. ■

Example 11.8. Find the formula for the joint density of $Y_{(1)}$ and $Y_{(n)}$.

Solution:

(a) Note that, since the Y_j are continuous, "$<$" and "$\le$" can be interchanged and the same is true for "$>$" and "$\ge$" when computing probabilities.

(b) Also, applying $A = (A \cap B) \uplus A \cap B^{\complement}$ with $A = \{Y_{(n)} \le y_n\}$ and $B = \{Y_{(1)} \le y_1\}$ yields
$P\{Y_{(n)} \le y_n\} \;=\; P\{Y_{(n)} \le y_n,\, Y_{(1)} \le y_1\} \;+\; P\{Y_{(n)} \le y_n,\, Y_{(1)} > y_1\}.$

We find the CDF as follows:

$$\begin{aligned} F_{Y_{(1)},Y_{(n)}}(y_1, y_n) &\overset{\text{(b)}}{=} P\{Y_{(n)} \le y_n\} \;-\; P\{Y_{(1)} > y_1, Y_{(n)} \le y_n\} \\ &= P\{Y_j \le y_n \text{ for all } j\} \;-\; P\{y_1 < Y_j \le y_n \text{ for all } j\} \\ &= \prod_{j=1}^{n} P\{Y_j \le y_n\} \;-\; \prod_{j=1}^{n} P\{y_1 < Y_j \le y_n\} \;=\; \big[F(y_n)\big]^n \;-\; \big[F(y_n) - F(y_1)\big]^n\,. \end{aligned}$$

We used first independence, then identical distribution in the last line.

Differentiation of the above then gives us $f_{Y_{(1)},Y_{(n)}}(y_1, y_n)$ as follows:

For convenience, we define $G(y_1, y_n) \;:=\; F_{Y_{(1)},Y_{(n)}}(y_1, y_n)$. Then,

$$G(y_1, y_n) \;=\; \big[F(y_n)\big]^n \;-\; \big[F(y_n) - F(y_1)\big]^n$$

Thus,

$$\frac{\partial G}{\partial y_1} \;=\; 0 \;-\; n\big[F(y_n) - F(y_1)\big]^{n-1} f(y_1) \;=\; n \cdot f(y_1)\,\big[F(y_n) - F(y_1)\big]^{n-1}$$

Thus,

$$f_{Y_{(1)},Y_{(n)}}(y_1,y_n) \;=\; \frac{\partial^2 G}{\partial y_1\,\partial y_n} \;= n\cdot f(y_1)\cdot(n-1)\left[F(y_n)-F(y_1)\right]^{n-2}\cdot f(y_n)$$
$$= n(n-1)\cdot f(y_1)\,f(y_n)\cdot\left[F(y_n)-F(y_1)\right]^{n-2}$$

Alternate solution:

The PDF can be found by interpreting certain events related to finding the density as the outcomes of the following multinomial sequence, $\vec{X}=(X_1,\ldots,X_n)$,

(c) For a given j, the outcomes ω_i' and associated probabilities p_i for X_j are

$\boxdot\ \omega_1'$: Y_j is close to $y_1 \;\Rightarrow\; p_1 = f(y_1)\,dy_1$ $\quad\boxdot\ \omega_2'$: Y_j is close to y_n; $\Rightarrow\; p_2 = f(y_n)\,dy_n$

$\boxdot\ \omega_3'$: Y_j strictly inbetween y_1 and $y_n \;\Rightarrow\; y_1 < Y_j < y_n \;\Rightarrow\; p_3 = F(y_n)-F(y_1)$.

Note that it is impossible that none of ω_1', ω_2', ω_3' happens and $Y_j < y_1$ or $Y_j > y_n$.

(d) We denote by W_i the count of indices j such that $X_j=\omega_i'$.
Then $\vec{W}=(W_1,W_2,W_3)\sim$ multinomial [110] with joint PMF $p_{\vec{W}}(\vec{w})$ given by

$$p_{\vec{W}}(\vec{w}) \;=\; \binom{n}{w_1,w_2,w_3}\,p_1^{w_1}\,p_2^{w_2}\,p_k^{w_3}\,.$$

- Similar to what was done in the proofs of theorems 11.22 (CDF and PDF of the jth order statistic) and 11.23 (Joint PDF of the order statistic), we conclude from **(c)** and **(d)** that

(e) $f_{Y_{(1)},Y_{(n)}}(y_1,y_n)\,d\,y_n d\,y_n \;=\; P\{Y_{(1)}$ is "dy_1 close" to y_1 and $Y_{(n)}$ is "dy_n close" to $y_n\ \}$

$= P\{$ exactly one Y_j is "dy_1 close" to y_1 and exactly one Y_j is "dy_n close" to y_n
and the other Y_j (there are $n-2$ left) are between y_1 and y_n

$= P\{W_1=1,W_2=1,W_3=n-2\} \;=\; p_{\vec{W}}(1,1,n-2) \;=\; \binom{n}{1,1,n-2}\,p_1^1\,p_2^1\,p_k^{n-2}\,.$

$= n(n-1)\cdot f(y_1)\,dy_1\cdot f(y_n)\,dy_n\cdot\left[F(y_n)-F(y_1)\right]^{-2}.$

(f) Thus, $f_{Y_{(1)},Y_{(n)}}(y_1,y_n)\,dy_1\,dy_n \overset{\text{(e)}}{=} n(n-1)\cdot f(y_1)\cdot f(y_n)\cdot\left[F(y_n)-F(y_1)\right]^{n-2}dy_1\,dy_n$.

- We cancel $dy_1\,dy_n$ in that last equation and obtain

(g) $$f_{Y_{(1)},Y_{(n)}}(y_1,y_n) \;=\; n(n-1)\cdot f(y_1)\cdot f(y_n)\cdot\left[F(y_n)-F(y_1)\right]^{n-2}.$$

We have obtained the same result for the joint PDF of $Y_{(1)}$ and $Y_{(n)}$ as in the first solution. □

Remark 11.17 (Sample median)**.** Recall from Definition 10.4 (pth quantile) on p.216 that the median of a random variable U with CDF $F_U(\cdot)$ was its 0.5th quantile

$$\phi_{0.5} \;=\; \min\{u\in\mathbb{R} : F_U(u)\geq 0.5\}\,.$$

If U is continuous with a strictly increasing CDF, then $\phi_{0.5}$ is that unique value u, for which $F_U(u)=0.5$. Thus, half of the area under the density $f_U(\cdot)$ is to the left of $\phi_{0.5}$ and the other half is to the right of $\phi_{0.5}$.

[110] See Definition 11.19 (Multinomial distribution) on p.274.

Assume that $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$ describes the action of picking a sample of n real numbers. In other words, each Y_j is a random variable and each invocation $\vec{Y}(\omega)$ results in the specific sample $\vec{y} = \big(y_1, \ldots, y_n\big)$, where

$$y_1 \,=\, Y_1(\omega),\; y_2 \,=\, Y_2(\omega),\; \ldots\; y_n \,=\, Y_n(\omega)\,.$$

Further assume that the Y_j are continuous. Then we can assume that all sample picks $Y_1, \ldots, Y_n$ are distinct, so that the order statistic satisfies strict inequalities

(A) $$Y_{(1)} \,<\, Y_{(2)} \,<\, \cdots\, Y_{(n)}\,.$$

The **sample median** of $\vec{Y}$ is defined as follows.

(a) If $n = 2k+1$ is odd, then the sample median of $\vec{Y}$ is is the $(k+1)$th order statistic $Y_{(k+1)}$.

(b) If $n = 2k$ is even, then the sample median of $\vec{Y}$ is is the (random) average $\dfrac{Y_k + Y_{k+1}}{2}$.

Two examples:

(1) If $n = 7$, then the sample median is $Y_{(n+1)} = Y_{(4)}$. Three of the Y_j are to the left of $Y_{(4)}$ and the same number are to the right.

(2) If $n \,=\, 8$, then the sample median of $\vec{Y}$ is is the average $\dfrac{Y_4 + Y_5}{2}$. Since we have strict inequalities in **(A)**, four of the Y_j are to the left of the sample median and the same number are to the right.

The point to remember is that the sample median of an odd–sized sample is an order statistic, whereas that of an even sized one is not.

Example: Let us assume that the the sample picks of an odd sized sample $\vec{Y} = \big(Y_1, \ldots, Y_{2n+1}\big)$ are continuous and iid random variables. We can compute the PDF of the sample median as that of $Y_{(n+1)}$ This time we do so by associating a multinomial random vector with three outcomes: Either Y_j is near y_{n+1} or it is near one of the n values to the left or it is near one of the n values to the right. In that manner we obtain

$$f_{Y_{(n+1)}}(y) \;=\; \binom{2n+1}{n, 1, n}[F(y)]^n \cdot f(y) \cdot [1 - F(y)]^n\,. \;\square$$

Remark 11.18. Here are two observations about n iid random variables $Y_1, \ldots, Y_n$.

(a) Assume that $Y_{k_1}, \ldots, Y_{k_n}$ is a permutation (ANY permutation!!) of $Y_1, \ldots, Y_n$. Then the symmetry that results from iid implies that

$$P\{Y_1 < Y_2 < \cdots < Y_n\} \;=\; P\{Y_{k_1} < Y_{k_2} < \cdots < Y_{k_n}\}\,.$$

Since there are $n!$ permutations, each one of those probabilities equals $\dfrac{1}{n!}$.

(b) Fix an arbitrary $k \in [1, k]_{\mathbb{Z}}$. Then

$$P\{Y_k = Y_{(1)}\} \;=\; P\{Y_k = Y_{(2)}\} \;=\; \ldots\, P\{Y_k = Y_{(n)}\}\,.$$

Since there are n such arrangements, each one of those probabilities equals $\dfrac{1}{n}$. $\square$

11.10 The Bivariate Normal Distribution (Optional)

Definition 11.21 (Bivariate normal distribution). ⋆

We say that two continuous random variables Y_1 and Y_2 have a **bivariate normal distribution**, or that they have a **joint normal distribution**, if their joint PDF is

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \qquad -\infty < y_1 < \infty,\ -\infty < y_2 < \infty, \tag{11.73}$$

where $Q = \frac{1}{1-\rho^2}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right]$.
We then also write $(Y_1,Y_2) \sim \mathcal{N}(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\rho)$. □

Whereas we have marked this definition as optional, you should remember the following theorem.

Theorem 11.24.

If two random variables Y_1 and Y_2 are $\mathcal{N}(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\rho)$, then
(a) *$Y_1 \sim \mathcal{N}(\mu_1,\sigma_1^2$ and $Y_1 \sim \mathcal{N}(\mu_2,\sigma_2^2$.*
Thus, $E[Y_1]=\mu_1$, $Var[Y_1]=\sigma_1^2$, $E[Y_2]=\mu_2$, $Var[Y_2]=\sigma_2^2$.
(b) *$Cov[Y_1,Y_2] = \sigma_1\,\sigma_2\,\rho$. Thus, ρ is the correlation coefficient of Y_1 and Y_2.*

PROOF (outline):
One proves **(a)** by showing that the marginal densities are

$$f_{Y_1}(y) = \frac{1}{\sigma_1\sqrt{2\pi}}e^{-(y-\mu_1)^2/(2\sigma_1^2)}, \qquad f_{Y_2}(y) = \frac{1}{\sigma_2\sqrt{2\pi}}e^{-(y-\mu_2)^2/(2\sigma_2^2)}.$$

See (10.37) on p.230.
For the proof of **(b)**, see Casella, Berger [3]. ■

Theorem 11.25.

If two jointly normal random variables Y_1 and Y_2 are uncorrelated, then they are independent.

PROOF: If $\rho = 0$, the joint PDF of Y_1 and Y_2 which was given in (11.73) is

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2},$$

where $Q = \frac{(y_1-\mu_1)^2}{\sigma_1^2} - 0 + \frac{(y_2-\mu_2)^2}{\sigma_2^2}$. Thus,

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{(\sqrt{2\pi}\sigma_1)(\sqrt{2\pi}\sigma_2)}\exp\left\{-\frac{(y_1-\mu_1)^2}{2\sigma_1^2} - \frac{(y_2-\mu_2)^2}{2\sigma_2^2}\right\} \\ &= \left(\frac{1}{(\sqrt{2\pi}\sigma_1)}\exp\left\{-\frac{(y_1-\mu_1)^2}{2\sigma_1^2}\right\}\right)\left(\frac{1}{(\sqrt{2\pi}\sigma_2)}\exp\left\{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}\right\}\right) \end{aligned}$$

It follows from Theorem 11.24**(a)** that $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1)\, f_{Y_2}(y_2)$. The independence of Y_1 and Y_2 follows from Theorem 11.4 on p.248. ■

Remark 11.19. The concept of joint normality can be extended from two random variables to an arbitrary number of random variables $Y_1, \ldots, Y_n$. However, the definition of their joint PDF utilizes $n \times n$ matrices and their determinants. This requires some background in linear algebra and that is not a prerequisite for this course. □

12 Functions of Random Variables and their Distribution

This chapter essentially only contains enough material to serve as a reference and review "sheet". You will not be able to properly understand the techniques noted here if you do not work through the many examples of the WMS text!

12.1 The Method of Distribution Functions

The Method of Distribution Functions is best explained by some examples.

Example 12.1. Find the CDF and PDF for $U := 2Y - 6$, where the density of the random variable Y is

$$f_Y(y) \;=\; \begin{cases} 8y\,, & \text{if } 0 \le y \le 1/2\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{12.1}$$

Solution: Applying the distribution function method means the following:

- ⊡ Find the CDF $F_U(u)$ of U ⊡ Find the PDF $f_U(u)$ of U by differentiating $F_U(u)$
- ⊡ Do this with help of the relation $U \;=\; 2Y - 6 \;\Leftrightarrow\; Y \;=\; \dfrac{U+6}{2}$.

We obtain

$$F_U(u) \;=\; P\{U \le u\} \;=\; P\{2Y - 6 \le u\} \;=\; P\left\{Y \;\le\; \frac{u+6}{2}\right\} \;=\; F_Y\left(\frac{u+6}{2}\right).$$

Note that

$$0 \;\le y \;\le\; \frac{1}{2} \;\Leftrightarrow\; 0 \;\le\; \frac{u+6}{2} \;\le \frac{1}{2} \;\Leftrightarrow\; -6 \;\le\; u \;\le -5$$

Thus, $F_U(u) = 0$ for $u < -6$ and $F_U(u) = 1$ for $u > -5$.

For $-6 \le u \le -5$, i.e., $0 \le y \le \frac{1}{2}$, we must integrate:

$$P\left\{Y \;\le\; \frac{u+6}{2}\right\} \;=\; \int_0^{(u+6)/2} f_Y(t)\,dt \;=\; \int_0^{(u+6)/2} 8t\,dt \;=\; \frac{8}{2}\left(\frac{u+6}{2}\right)^2.$$

We combine the cases $u < -6;\ -6 \le u \le -5;\ u > -5$ and obtain

$$F_U(u) \;=\; \begin{cases} 0\,, & \text{if } u < -6\,, \\ (u+6)^2\,, & \text{if } -6 \le u \le -5\,, \\ 1\,, & \text{if } u > -5\,. \end{cases}$$

We differentiate this CDF and otain the density function for U:

$$f_U(u) \;=\; \frac{dF_U(u)}{du} \;=\; \begin{cases} 2(u+6)\,, & \text{if } -6 \le u \le -5\,, \\ 0\,, & \text{else}\,. \end{cases} \;\square$$

Example 12.2 (WMS Ch.06.3, Example 6.3)**.** The following is Example 6.3 of the WMS text. Its proof has been substantially rewritten.

Let (Y_1, Y_2) denote a random sample of size $n = 2$ from the uniform distribution on the interval $(0, 1)$. In other words, we assume that Y_1 and Y_2 are jointly continuous and have a joint PDF which is constant and not zero on the unit square.

The issue is to find the probability density function for $U := Y_1 + Y_2$.

Solution: It follows from the assumptions that Y_1 and Y_2 possess the same mariginal PDF The density function for each Y_i is

$$f(y) := f_{Y_1}(y) = f_{Y_2}(y) = \begin{cases} 1\,, & 0 \le y \le 1\,, \\ 0\,, & \text{elsewhere}\,. \end{cases}$$

Since Y_1 and Y_2 are independent,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2) = f(y_1) f(y_2) = \begin{cases} 1\,, & 0 \le y_1 \le 1\,, 0 \le y_2 \le 1\,, \\ 0\,, & \text{elsewhere}\,. \end{cases}$$

Thus, $F_U(u) = P\{Y_1 + Y_2 \le u\} = \iint_B f(y_1) f(y_2)\, dy_1\, dy_2$, where, for a fixed u, the region of integration is

(A) $$B := \big([0,1] \times [0,1]\big) \cap \{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \le u\}\,.$$

We will separately treat the cases • $u \le 0$ or $u \ge 2$ • $0 < u \le 1$ • $1 < u < 2$.

Case 1: $u \le 0$ or $u \ge 2$.

If $u \le 0$, then $[0,1] \times [0,1]$ and $\{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \le u\}$ are disjoint. Thus, $B = \emptyset$ and $\iint_B \cdots = 0$ and thus, $F_U(u) = 0$.

If $u \ge 2$, then $[0,1] \times [0,1] \subseteq \{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \le u\}$. Thus, $\iint_B \cdots = \int_0^1 \int_0^1 \cdots$ and thus, $F_U(u) = 1$.

Case 2: • $0 < u \le 1$.

The graph of $y_1 + y_2 = u$ in the (y_1, y_2) plane is a straight line which intersects the vertical coordinate axis, $y_1 = 0$, at $y_2 = u$ and the horizontal coordinate axis, $y_2 = 0$, at $y_1 = u$. Thus, B is the triangle bounded by the coordinate axes and the line $y_1 + y_2 = u$. since it is half of a square with side length u, its area is $u^2/2$.

Of course, this also follows from the fact that $\iint_B \ldots$ is achieved by first integrating, for $0 \le y_1 \le u$, over the vertical slice of B at y_1 and then integrating those integrals. Since the vertical slice of B at y_1 extends from $y_2 = 0$ to $y_1 + y_2 = u$, i.e., to $y_2 = u - y_1$

$$\begin{aligned} F_U(u) &= \iint_B 1\, dy_1\, dy_2 = \int_0^u \int_0^{u-y_1} 1\, dy_2\, dy_1 \\ &= \int_0^u (u - y_1)\, dy_1 = \left(u y_1 - \frac{u^2}{2} \right) \Big|_0^u = u^2 - \frac{u^2}{2} = \frac{u^2}{2}\,. \end{aligned}$$

Case 3: • $1 < u < 2$.

Let $\widetilde{B} := \big([0,1]\times[0,1]\big) \setminus \{(y_1,y_2)\in\mathbb{R}^2 : y_1+y_2 \geq u\}$. Then

(B) $$\widetilde{B} \;= \big([0,1]\times[0,1]\big) \cap \{(y_1,y_2)\in\mathbb{R}^2 : y_1+y_2 \leq u\}\,,$$

(C) $$F_U(u) \;= 1 \,-\, P\{Y_1+Y_2\geq u\} \;=\; 1 \,-\, \iint_{\widetilde{B}} 1\,dy_1\,dy_2$$

Now, the graph of $y_1+y_2=u$ in the (y_1,y_2) plane is a straight line which intersects the vertical line, $y_1=1$, at $y_2=u-1$ and the horizontal line, $y_2=0$, at $y_1=u-1$.
$\widetilde{B}$ is the right angle triangle bounded by the lines $y_1=1, y_2=1$ and $y_1+y_2=u$.
Its legs have length $1-(u-1)=2-u$. Thus, its area is half that of a square with side length $2-u$. Thus, the area of $\widetilde{B}$ is $(2-u)^2/2$. It follows from **(C)** that

$$F_U(u) \;=\; 1 \,-\, \text{area}(\widetilde{B}) \;=\; 1 \,-\, \frac{4-4u+u^2}{2} \;=\; -1 \,+\, 2u \,-\, \frac{u^2}{2}\,.$$

This also could have been computed by iterated integration. In this case,

$$\begin{aligned} 1 \,-\, F_U(u) \;&=\; \iint_{\widetilde{B}} 1\,dy_1\,dy_2 \;=\; \int_{u-1}^{1}\int_{u-y_1}^{1} 1\,dy_2\,dy_1 \\ &= \int_{u-y_1}^{1} (1-u+y_1)\,dy_1 \;=\; \left((1\,-\,u) \,+\, \frac{y_1^2}{2}\right)\Big|_{u-1}^{1} \\ &= (1-u)(2-u) \,+\, \frac{1}{2} \,-\, \frac{(u-1)^2}{2} \;=\; 2 \,-\, 2u \,+\, \frac{u^2}{2}\,. \end{aligned}$$

We thus obtain, as before, $F_U(u) \;=\; 1-\big(2+2u-u^2/2\big) \;=\; -1+2u-u^2/2$. $\square$

The problem of the next example is that of WMS Ch.6.4, Example 6.8. This instructor does not understand the reasoning given there and has provided a completely different proof. You find this example here rather than in the next section (section 12.2: The Method of Transformations in One Dimension), because it is solved with the techniques of this section.

Example 12.3. Let Y_1 and Y_2 be jointly continuous random variables with density function

$$f_{Y_1,Y_2}(y_1,y_2) \;=\; \begin{cases} e^{-(y_1+y_2)}\,, & 0\leq y_1\,, 0\leq y_2\,, \\ 0\,, & \text{else}\,. \end{cases}$$

What are the CDF andPDF of $U := Y_1+Y_2$?

Solution:

$$P\{U\leq u\} \;= P\{Y_1+Y_2\leq u\} \;=\; \iint_R e^{-y_1-y_2}\,d\vec{y}$$

where R = triangle with vertices $(0,u),(0,0),(u,0)$. Thus, for $u>0$,

$$\begin{aligned} P\{U\leq u\} \;&=\; \int_0^u \left[\int_0^{u-y_1} e^{-y_1-y_2}\,dy_2\right] dy_1 \;=\; \int_0^u e^{-y_1}\left[-e^{-y_2}\Big|_0^{u-y_1}\right] dy_1 \\ &= \int_0^u e^{-y_1}\left[1\,-\,e^{-(u-y_1)}\right] dy_1 \;=\; \int_0^u e^{-y_1}\left[1\,-\,e^{y_1}e^{-u}\right] dy_1 \\ &= \int_0^u e^{-y_1}\,dy_1 \,-\, \int_0^u e^{-u}\,dy_1 \;=\; -e^{-y_1}\Big|_0^u \,-\, u\,e^{-u} \\ &= -\,(e^{-u}-1) \,-\, u\,e^{-u} \;=\; 1 \,-\, (1+u)\,e^{-u}\,. \end{aligned}$$

The derivative is (for $u > 0$)

$$\begin{aligned} f_U(u) &= \frac{d}{du}\big(1 - (1+u)\,e^{-u}\big) = -(1+u)'\,e^{-u} - (1+u)\big(e^{-u}\big)' \\ &= -e^{-u} - (1+u)\big(-e^{-u}\big) = -e^{-u} + e^{-u} + u\,e^{-u} = u\,e^{-u}. \end{aligned}$$

Thus, the CDF is $F_U(u) = \begin{cases} 1 - (1+u)\,e^{-u}, & \text{if } u > 0, \\ 0, & \text{else} \end{cases}$

and the PDF is $f_U(u) = \begin{cases} u\,e^{-u}, & \text{if } u > 0, \\ 0, & \text{else}. \end{cases}$

The latter agrees with the WMS result. □

Remark 12.1. In the following we use the arrow notation $\vec{y} = \big(y_1, \ldots, y_n\big)$, $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$, ...

Summary of the Distribution Function Method

Goal: Find the PDF $f_U(u)$ for $U = g(\vec{Y})$, where $g : D \to \mathbb{R}$ has a domain $D \subseteq \mathbb{R}^n$ large enough to hold all arguments $\vec{y}$ that are relevant for the problem.

(1) Find the region R = $\{g \le u\} = g^{-1}\big(]-\infty, u]\big)$. (Thus, $R \subseteq \mathbb{R}^n$.)
(2) Find the "boundary" $R^* = \{g = u\}$ of the region R.
(3) Find the CDF $F_U(u) = P\{U \le u\}$ by integrating $f(\vec{y})$ over the region R.
(4) Find the the PDF $f_U(u) = \frac{dF_U(u)}{du}$ by differentiating $F_U(u)$.

Note for the above that, since g may not be invertible, g^{-1} denotes the preimage $g^{-1}(B) = \{\vec{y} : g(\vec{y}) \in B\}$, where $B \subseteq \mathbb{R}$. If, e.g., $B =]-\infty, u]$, then $R = g^{-1}\big(]-\infty, u]\big)$, and **(3)** expresses

$$\begin{aligned} F_U(u) &= P\{U \le u\} = P\{g\big(\vec{Y}\big) \le u\} = P\big\{\omega : \vec{Y}(\omega) = \vec{y} \text{ such that } g(\vec{y}) \le u\big\} \\ &= P\{Y \in R\} = \iint \cdots \int_R f_{\vec{Y}}(\vec{y})\, d\vec{y}. \ \square \end{aligned} \tag{12.2}$$

The next remark really should be considered another example for the distribution method. It has been marked as optional, so it will not be part of any exam or quiz. Nevertheless, you are strongly encouraged to work through its proof and increase your skills with respect to applying the distribution method.

Remark 12.2. ⋆ Let Y be a continuous random variable with PDF $f_Y(y)$ and let $h : \mathbb{R} \to \mathbb{R}$ be a **symmetrical function** (also, **symmetric function**), i.e., $h(-y) = h(y)$ for all y. Also, assume that

(1) $y \mapsto h(y)$ is differentiable (hence, continuous) everywhere.
(2) $y \mapsto h(y)$ is injective for $y \ge 0$, i.e., $0 \le y < y' \ \Rightarrow\ h(y) \ne h(y')$. (Thus, by symmetry, $h(y)$ also is injective for $y < 0$).

Continuous functions of a real variable are either strictly increasing or strictly decreasing on any subset of the domain where they are injective. (Draw a picture!) Thus, there are two possibilities.

(1) h is strictly increasing on $[0, \infty[$ (and then, by symmetry, h is strictly decreasing on $[-\infty, 0[$). Also, h attains its global minimum at $y = 0$.
(2) h is strictly decreasing on $[0, \infty[$ (and then, by symmetry, h is strictly increasing on $[-\infty, 0[$). Also, h attains its global maximum at $y = 0$.

In either case, there are no jumps for the continuous $h(\cdot)$. We will determine the CDF and PDF of the random variable $U := h \circ Y$ under the following assumptions: For any given $u \in \mathbb{R}$,

(3) h is strictly increasing on $[0, \infty[$

(4) $h(0) = 0$ and thus, $h(y) \geq 0$ for all y. Note that then $P\{U > 0\} = 1$ and $P\{U \leq 0\} = 0$.

(6) Thus, if $u > 0$, then $U\big(\omega\big) \leq u \;\Leftrightarrow\; |Y(\omega)| \leq y = h^{-1}(u)$. Thus,

$$\begin{aligned} F_U(u) \;&= P\{U \leq u\} \;=\; P\{|Y| \leq h^{-1}(u)\} \;=\; P\{-h^{-1}(u) \leq Y \leq h^{-1}(u)\} \\ &= F_Y\big(h^{-1}(u)\big) \;-\; F_Y\big(-h^{-1}(u)\big) \quad \text{if } u > 0, \text{ i.e.,} \end{aligned}$$

$$F_U(u) \;=\; \begin{cases} 1\,, & \text{if } h(y) < u \text{ for all } y\,, \\ F_Y\big(h^{-1}(u)\big) \;-\; F_Y\big(-h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{if } u \leq 0\,. \end{cases} \tag{12.3}$$

We differentiate $\dfrac{d}{du}$ to obtain the density. We write $h^{-1\prime}(u) = \dfrac{dh^{-1}(u)}{du}$:

- $f_U(u) \;=\; h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big) \;-\; (-1) h^{-1\prime}(u) f_Y\big(-h^{-1}(u)\big)$

Thus,

$$f_U(u) \;=\; \begin{cases} h^{-1\prime}(u)\, \big[\, f_Y\big(h^{-1}(u)\big) \;+\; f_Y\big(-h^{-1}(u)\big)\,\big]\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{12.4}$$

□

Example 12.4. As an example for that last remark, let us consider the function $h(y) = y^2$. [111] h is strictly increasing on $[0, \infty[$ and its minimum is $h(0) = 0$. Thus, h satisfies the assumptions **(3)** and **(4)** of Remark12.2. Since $\lim\limits_{y \to \infty} y^2 = \infty$, the condition "if $h(y) < u$ for all y" of (12.3) is never satisfied. Further, the condition "if there is $y = h^{-1}(u)$" of (12.3) and (12.4) becomes "$u \geq 0$".

Thus, if $U = Y^2$, then $h^{-1}(u) = \sqrt{u}$ for $u \geq 0$ and $h^{-1\prime}(u) \;=\; 1/(2\sqrt{u})$. We obtain

$$f_U(u) \;=\; \begin{cases} \dfrac{1}{2\sqrt{u}}\, \big[\, f_Y\big(\sqrt{u}\big) \;+\; f_Y\big(-\sqrt{u}\big)\,\big]\,, & \text{if } u > 0\,, \\ 0\,, & \text{else}\,. \end{cases}$$

□

Example 12.5. Assume that the random variable Y is $\mathcal{N}(0,1)$, i.e., Y is standard normal. What is the distribution of $U := Y^2$?

For this example, let

$$\phi(y) \;:= f_Y(y) \;=\; \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,, \tag{12.5}$$

$$\Phi(y) \;:= \int_{-\infty}^{y} \phi(t)\, dt\,. \tag{12.6}$$

In other words, ϕ is the PDF of Y and Φ is the CDF of Y.

Since $U \geq 0$, we have $f_U(u) = F_U(u) = 0$ for $u < 0$. Thus, we may assume that $u \geq 0$.

[111] That is WMS Example 6.4.

Then, $F_U(u) = P\{-\sqrt{u} \le Y \le \sqrt{u}\} = \Phi(\sqrt{u}) - \Phi(-\sqrt{u})$ and thus,

$$f_U(u) = F_U'(u) = \frac{d}{du}\left[\,\Phi(\sqrt{u}) - \Phi(-\sqrt{u})\,\right]$$
$$= \phi(\sqrt{u})\,\frac{1}{2\sqrt{u}} + \phi(-\sqrt{u})\,\frac{1}{2\sqrt{u}} = \phi(\sqrt{u})\,\frac{1}{\sqrt{u}} = \frac{1}{\sqrt{2\pi}}\,e^{-(\sqrt{u})^2/2}\,\frac{1}{\sqrt{u}}$$

Above, we used symmetry $\phi(-\sqrt{u}) = \phi(\sqrt{u})$ to obtain the equation before the last. Thus,

$$f_U(u) = \frac{1}{\sqrt{2\pi}}\,e^{-u/2}\,u^{-1/2} = \frac{u^{1/2-1}}{2^{1/2}\sqrt{\pi}}\,e^{-u/2}$$

One can show that $\Gamma(1/2) = \sqrt{\pi}$. [112] We use that result without attempting to prove it and obtain, setting $\alpha := 1/2$ and $\beta := 2$,

$$f_U(u) = \frac{u^{1/2-1}\,e^{-u/2}}{2^{1/2}\Gamma(1/2)} = \frac{u^{\alpha-1}\,e^{-u/\beta}}{\beta^\alpha\Gamma(\alpha)}\,.$$

We finally remember that all this was done for $u \ge 0$ and that $f_U(u) = 0$ for $u < 0$.

$$f_U(u) = \begin{cases} \dfrac{u^{\alpha-1}\,e^{-u/\beta}}{\beta^\alpha\Gamma(\alpha)}\,, & \text{if } u \ge 0\,, \\ 0\,, & \text{else}\,. \end{cases}$$

It follows that the square of a $\mathcal{N}(0,1)$ variable has a gamma$(1/2, 2)$ distribution. Equivalently, it has a chi–square distribution with one degree of freedom. □

Example 12.6. It is important that you recognize when there are significant shortcuts. It might be possible to obtain $F_U(u) = F_U\big(g^{-1}(y)\big)$ without having to integrate the PDF. Here is an example. Let the random variable Y be expon(1). Find the CDF and PDF of $U := 2Y - 4$.

Solution:

(1) Here, $u = g(y) = 2y - 4$ has inverse $y = g^{-1}(u) = (u+4)/2$.

(2) The CDF of Y is $F_Y(y) = \begin{cases} 1 - e^{-y}\,, & \text{if } y \ge 0\,, \\ 0\,, & \text{else}\,. \end{cases}$

(3) Thus, $F_U(u) = P\{U \le u\} = P\{2Y - 4 \le u\} = P\left\{Y \le \dfrac{u+4}{2}\right\} = F_Y\left(\dfrac{u+4}{2}\right)$.

(4) From **(2)**: $F_U(u) = \begin{cases} 1 - e^{-\frac{u+4}{2}}\,, & \text{if } \frac{u+4}{2} \ge 0\,, \\ 0\,, & \text{else}\,. \end{cases}$

(5) Thus, $F_U(u) = \begin{cases} 1 - e^{-\frac{u+4}{2}}\,, & \text{if } u \ge -4\,, \\ 0\,, & \text{else}\,. \end{cases}$

(6) We have obtained $F_U(u)$ without integrating a PDF.

(7) The density is $f_U(u) = F_U'(u) = \begin{cases} \frac{1}{2}\,e^{-\frac{u+4}{2}}\,, & \text{if } u \ge -4\,, \\ 0\,, & \text{else}\,. \end{cases}$

□

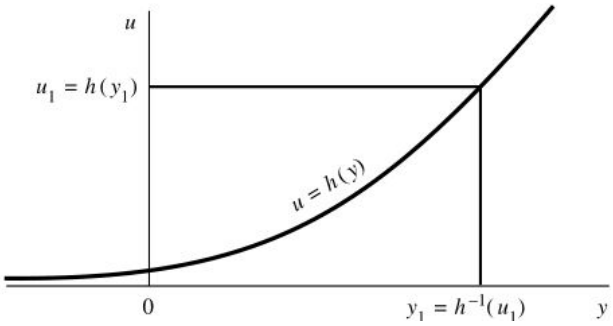
[112] See, e.g., https://en.wikipedia.org/wiki/Gamma_function or Shilov, G. [9].

12.2 The Method of Transformations in One Dimension

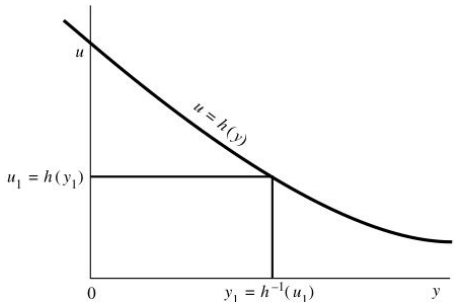
Introduction 12.1. We already encountered the method of transformations in Remark 12.2 on p.293. There we computed the CDF and PDF of the random variable $U = h(Y)$ for a continuous random variable Y and a symmetric and differentiable function $h(y)$ which was injective on the interval $B_1 = [0, \infty[$. (By symmetry, h also had those characteristics on $B_2 =]-\infty, 0[$.)

At the heart of the calculations was the fact that injectivity allowed us to compute, for a given u, a unique $y = h{-}1(u)$ such that $h(y) = u$.

Since differentiable functions are continuous, injectivity on an interval B implies that h is either strictly increasing or strictly decreasing on B. See figures 12.1 and 12.2 below.



12.1 (Figure). **Strictly increasing function.**
Source: WMS Ch.6.4



12.2 (Figure). **Strictly decreasing function.**
Source: WMS Ch.6.4

Those figures illustrate the following.

(1) If h is strictly increasing, then $h(y) \le u_1 \Leftrightarrow y \le h^{-1}(u_1)$. Thus,

$$P\{U \le u\} = P\{h(Y) \le u\} = P\{h^{-1}[h(Y)] \le h^{-1}(u)\} = = P\{Y \le h^{-1}(u)\}, \quad \text{i.e.,} \quad F_U(u) = F_Y\big(h^{-1}(u)\big). \tag{12.7}$$

(2) If h is strictly decreasing, then $h(y) \le u_1 \Leftrightarrow y \ge h^{-1}(u_1)$. Thus,

$$P\{U \le u\} = P\{h(Y) \le u\} = P\{Y \ge h^{-1}(u)\} = 1 - P\{Y \le h^{-1}(u)\}, \quad \text{i.e.,} \quad F_U(u) = 1 - F_Y\big(h^{-1}(u)\big). \tag{12.8}$$

Case I: h is strictly increasing

We differentiate (12.7) with respect to u and write $h^{-1\prime}(u)$ for $\dfrac{dh^{-1}(u)}{du}$. Then

$$f_U(u) = \frac{dF_U(u)}{du} = \frac{dF_Y\big(h^{-1}(u)\big)}{du} = f_Y\big(h^{-1}(u)\big) \cdot h^{-1\prime}(u).$$

Since h is strictly increasing, $h^{-1\prime}(u) > 0$. Thus, $h^{-1\prime}(u) = |h^{-1\prime}(u)|$. Thus,

$$f_U(u) = f_Y\big(h^{-1}(u)\big) \cdot |h^{-1\prime}(u)|. \tag{12.9}$$

Case II: h is strictly decreasing

We differentiate (12.8) with respect to u. Then

$$f_U(u) \;=\; -\frac{dF_Y\big(h^{-1}(u)\big)}{du} \;=\; f_Y\big(h^{-1}(u)\big)\cdot\big(-h^{-1\prime}(u)\big)\,.$$

Since h is strictly decreasing, $h^{-1\prime}(u) < 0$. Thus, $-h^{-1\prime}(u) = |h^{-1\prime}(u)|$. Thus,

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big)\cdot|h^{-1\prime}(u)|\,. \tag{12.10}$$

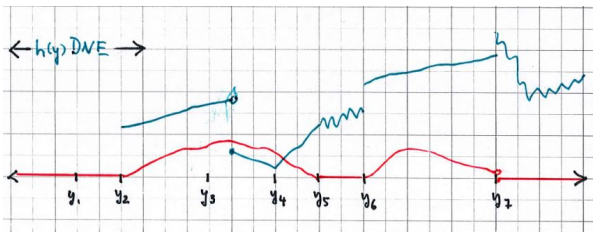
(3) We compare (12.9) and (12.10) and see that they are equal. Thus, as long as h is eiher strictly increasing everywhere or strictly decreasing everywhere, (i.e., as long as f is invertible everywhere,)

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big)\cdot|h^{-1\prime}(u)| \;=\; f_Y\big(h^{-1}(u)\big)\cdot\left|\frac{d[h^{-1}(u)]}{du}\right|\,. \tag{12.11}$$

Since $\displaystyle\int_a^b f_Y(t)\,dt \;=\; \int\limits_{[a,b]\cap\{\tilde{y}:f(\tilde{y})\neq 0\}} f_Y(t)\,dt$ for any interval $[a,b]$, we only need to worry about the behavior of h for arguments belonging to

$$\text{suppt}(f_Y) \;=\; \{\tilde{y} : f_Y(\tilde{y}) \neq 0\}\,.^{113}$$

- $\text{suppt}(f_Y) =]y_2, y_5[\cup]y_6, y_7[$. It does not matter what $h(y)$ does outside $\text{suppt}(f_Y)$.
- h must be injective on the support of f_Y.
- h changes direction at y_3 and y_4, so the pieces $]y_2, y_3[$, $]y_3, y_4[$, $]y_4, y_5[$, must be treated separately. □



The following theorem summarizes the observations of those introductory results:

Theorem 12.1.

Given are a continuous random variable Y with density $f_Y(y)$ and a differentiable function $h(y)$ which is either strictly increasing or strictly decreasing for all $y \in \text{suppt}(f_Y)$, i.e., for all y that satisfy $f_Y(y) > 0$. Then the PDF of $U := h(Y)$ is

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big)\cdot|h^{-1\prime}(u)| \;=\; f_Y\big(h^{-1}(u)\big)\cdot\left|\frac{d[h^{-1}(u)]}{du}\right|\,. \tag{12.12}$$

PROOF: See the introduction 12.1. ■

Example 12.7 (Increasing function)**.** Given is a random variable Y with the following PDF:

$$f_Y(y) \;=\; \begin{cases} 2y\,, & \text{if } 0 \le y \le 1\,,\\ 0\,, & \text{else}\,. \end{cases}$$

[113] ⋆ See Definition 4.8 (Support of a real–valued function) on p.99

Let $U := 4Y - 3$. Find the PDF for U by means of the transformation method.

Solution: We apply the transformation method with the strictly increasing function $u = h(y) = 4y - 3$. Then the inverse of h is $y = h^{-1}(u) = (u+3)/4$, for all $u \in \mathbb{R}$.

(1) We apply the transformation method with $u = h(y) = 4y - 3$ (strictly increasing).
(2) Then the inverse of h is $y = h^{-1}(u) = (u+3)/4$, for all $u \in \mathbb{R}$.
(3) Further, $h^{-1\prime}(u) = 1/4$. Since $0 \le (u+3)/4 \le 1 \;\Leftrightarrow\; -3 \le u \le 1$,

$$f_U(u) \;=\; \begin{cases} \dfrac{2(u+3)}{4} \cdot \dfrac{1}{4}\,, & \text{if } -3 \le u \le 1\,, \\ 0\,, & \text{else}\,. \end{cases} \;=\; \begin{cases} \dfrac{u+3}{8}\,, & \text{if } -3 \le u \le 1\,, \\ 0\,, & \text{else}\,. \end{cases} \quad \square$$

Example 12.8 (Decreasing function)**.** Given is a random variable Y with the same PDF as in Example 12.7:

$$f_Y(y) \;=\; \begin{cases} 2y\,, & \text{if } 0 \le y \le 1\,, \\ 0\,, & \text{else}\,. \end{cases}$$

Let $U := -3Y + 2$. Find the PDF for U by means of the transformation method.

Solution: We apply the transformation method with the strictly decreasing function $u = h(y) = 2 - 3y$. Then the inverse of h is $y = h^{-1}(u) = (2-u)/3$, for all $u \in \mathbb{R}$.

(1) We apply the transformation method with $u = h(y) = 2 - 3y$ (strictly decreasing).
(2) Then the inverse of h is $y = h^{-1}(u) = (2-u)/3$, for all $u \in \mathbb{R}$.
(3) Further, $h^{-1\prime}(u) = -1/3$. Since $0 \le (2-u)/3 \le 1 \;\Leftrightarrow\; 0 \ge (u-2) \ge -3 \;\Leftrightarrow\; -1 \le u \le 2$,

$$f_U(u) \;=\; \begin{cases} \dfrac{2(2-u)}{3} \cdot \left|\dfrac{-1}{3}\right|\,, & \text{if } -1 \le u \le 2\,, \\ 0\,, & \text{else}\,. \end{cases} \;=\; \begin{cases} \dfrac{4-2u}{9}\,, & \text{if } -3 \le u \le 1\,, \\ 0\,, & \text{else}\,. \end{cases} \quad \square$$

Example 12.9 (Distribution function method with two variables)**.** Given are two jointly continuous random variables with uniform distribution on the triangle

$$B \;:=\; \{(y_1, y_2) : 0 < y_2 < 1 - y_1 < 1\}\,.$$

Find the CDF of $U = Y_1 + Y_2$.

(1) The joint PDF of (Y_1, Y_2) is $\quad f_{Y_1,Y_2}(y_1, y_2) \;=\; \begin{cases} 2\,, & \text{if } 0 < y_2 < 1 - y_1 < 1\,, \\ 0\,, & \text{else}\,. \end{cases}$
(2) $F_U(u) \;=\; P\{U \le u\} \;=\; P\{Y_1 + Y_2 \le u\} \;=\; \iint\limits_{B \cap C} 2\, d\vec{y}$, where $C = \{(y_1, y_2) : y_1 + y_2 \le u\}$.
(3) $(y_1, y_2) \in B \;\Rightarrow\; 0 < 1 - y_1 < 1 \;\Rightarrow\; 0 > y_1 - 1 > -1 \;\Rightarrow\; 0 < y_1 < 1$.
$0 < y_2 < 1$ is obvious. Thus, $u \le 0 \;\Rightarrow\; P\{U \le u\} = 0$.
(4) B is the triangle with vertices $(0,0), (0,1)$ and $(1,0)$.
For $u > 0$, C is the triangle with vertices $(0,0), (0,u)$ and $(u,0)$

(5) Thus, $0 < u < 1 \;\Rightarrow\; B \cap C = C \;\Rightarrow\; \iint\limits_{B\cap C} 2\, d\vec{y} \;=\; 2 \iint_C d\vec{y}$

(6) Thus, from **(5)** & **(2)**, $0 < u < 1 \;\Rightarrow\; B \cap C = C \;\Rightarrow\; F_U(u) = 2 \iint_C d\vec{y}$.
$\iint_C \cdots d\vec{y}$ is done by integrating, for each fixed $0 < y_1 < u$, over that part of the vertical line $\{y_2 : y_2 = y_1\}$ that is within C. That is the segment $0 < y_2 < u - y_1$.

(7) Thus, $0 < u < 1 \;\Rightarrow\; F_U(u) = 2 \displaystyle\int_0^u \int_0^{u-y_1} dy_2\, dy_1$

$$= 2 \int_0^u (u - y_1 - 0)\, dy_1 \;=\; 2u^2 \;-\; 2\, \frac{y_1^2}{2}\Big|_0^u \;=\; u^2\,.$$

(8) From **(4)**, $u \geq 1 \;\Rightarrow\; B \cap C = B \;=\; \text{suppt}(f_U) \;\Rightarrow\; F_U(u) = 1$.

(9) Thus, from **(3)** & **(7)** & **(8)**, $F_U(u) \;=\; \begin{cases} 0\,, & \text{if } u \leq 0\,, \\ u^2\,, & \text{if } 0 < u < 1\,, \\ 1\,, & \text{if } u \geq 1\,. \end{cases}$

Differentiation yields $\quad f_U(u) \;=\; \begin{cases} 2u\,, & \text{if } 0 < u < 1\,, \\ 0\,, & \text{if } u \leq 0 \text{ or } u \geq 1\,. \end{cases}$ □

Remark 12.3. In the following we use the arrow notation $\vec{y} = \big(y_1, \ldots, y_n\big)$, $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$, ...

Summary of the Transformation Method

Goal: Find the PDF $f_U(u)$ for $U = h(Y)$, where

- $h : R \to \mathbb{R}$ has a domain $R \subseteq \mathbb{R}$ large enough to hold all arguments y that are relevant for the problem. That requires that R contains the support of the PDF f_Y (the set where f_Y is not zero).
- h is invertible on R. In other words, h is injective on R: If $y \in R$ and $u = h(y)$, then there is no $\tilde{y} \in R$ such that $\tilde{y} \neq y$ and $h(\tilde{y}) = u$.
- Thus h has an inverse $u \mapsto h^{-1}(u)$ which maps any u that is a function value $u = h(y)$ back to y. Do not confuse this genuine inverse function of $h(\cdot)$ with the preimage function $B \mapsto h^{-1}(B) = \{y \in Y : h(y) \in B\}$! That one maps **sets** to **sets**!
- We require that h is either strictly increasing or strictly decreasing for those $y \in R$ where $f_Y(y) > 0$. This assumption guarantees that h is injective and its inverse $u \mapsto h^{-1}(u)$ exists on the support of f_Y.

To find the PDF $f_U(u)$ for $U = h(Y)$, proceed as follows:

(1) Find the inverse function, $y = h^{-1}(u)$, for those u that correspond to y with $f_Y(y) \neq 0$.

(2) Find the derivative $\quad \dfrac{dh^{-1}}{du} \;=\; \dfrac{dh^{-1}(u)}{du} \;=\; h^{-1\prime}(u)$.

(3) Finally, compute $f_U(u)$ as follows: $\quad f_U(u) \;=\; f_Y\big(h^{-1}(u)\big) \left| \dfrac{dh^{-1}(u)}{du} \right|$. □

Remark 12.4. The transformation method still works if h is not either strictly increasing or decreasing on suppt(g),as long as h is injective and R can be subdivided by intervals on which h is either strictly increasing or strictly decreasing.

As an example, consider $u := h(y) := \begin{cases} y, & \text{if } y \le 0\,, \\ e^{-y}, & \text{if } y > 0\,. \end{cases}$

- On $]-\infty, 0]$, h is strictly increasing with inverse $y = h^{-1}(u) = u$. This inverse has derivative $h^{-1\prime}(u) = 1 > 0$.
- On $]0, \infty[$, h is strictly decreasing with inverse $y = h^{-1}(u) = -\ln(u)$. This inverse has derivative $h^{-1\prime}(u) = -1/u < 0$.
- Obviously if $y \le 0$, then $y \le 0 \Leftrightarrow u \le 0$. Moreover, $y > 0 \Leftrightarrow 0 < u = e^{-y} < 1$.
- Thus, $f_U(u) = \begin{cases} f_Y(h^{-1}(u) \cdot |1| = f_Y(u), & \text{if } u \le 0\,, \\ f_Y(h^{-1}(u)) \cdot |-1/u| = \dfrac{f_Y\big(-\ln(u)\big)}{u}, & \text{if } 0 < u < 1\,, \\ 0, & \text{else}\,. \end{cases}$ □

12.3 The Method of Transformations in Multiple Dimension

Introduction 12.2. In Chapter 12.2 (The Method of Transformations in Multiple Dimension), we looked for ways to compute the density f_U(u) of the transform $U = h(Y)$ of a continuous random variable Y by means of a function h which maps real numbers y to real numbers $u = h(y)$. Theorem 12.1 on p.297 provided us with an explicit formula for the PDF $f_U(u)$ of the transformed random variable $U = h(Y)$:

$$f_U(u) = f_Y\big(h^{-1}(u)\big) \cdot |h^{-1\prime}(u)| = f_Y\big(h^{-1}(u)\big) \cdot \left| \frac{d[h^{-1}(u)]}{du} \right|. \tag{12.13}$$

(1) Since $|h^{-1\prime}(u)|$ appears in that formula, $h^{-1}(u)$ must exist and be differentiable.
(2) That in turn requires that h is differentiable, in particular continuous.
(3) Moreover, neither $h'(y)$ nor $h^{-1\prime}(u)$ can be zero, since $h'(y) \cdot h^{-1\prime}(u) = 1$.

Existence of $h^{-1}(u)$ requires h to be injective on the support of the PDF f_Y:

(4) If u_0 is the function value $u_0 = h(y)$ of some argument y that satisfies $f_Y(y) > 0$,
• then there is no other argument $\tilde{y}$ that also satisfies $u_0 = h(\tilde{y})$ and $f_Y(\tilde{y}) > 0$.

Since h is continuous, **(4)** is satisfied if h is either strictly increasing or strictly decreasing for all y in the support of h, so we replaced **(4)** with that simpler assumption.

We now look for an n–dimensional analogue. If you have attended a linear algebra course, you are knowledgeable about $n \times n$ matrices and their determinants. If your background about those subjects is limited to a course in multivariable calculus, then assume that $n = 2$ or $n = 3$. We study

- random vectors $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$, where each coordinate Y_j is a random variable.
- functions $\vec{u} = \vec{h}\big(\vec{y}\big)$ that map n–dimensional arguments $\vec{y}$ to n–dimensional function values $\vec{y}$, have continuous partial derivatives $\dfrac{\partial h_i}{y_j}$ for $i, j \in [1, n]_{\mathbb{Z}}$ and that satisfy a multidimensional analogue of **(4)**:

(5) If the vector $\vec{u}_0$ is a function value $\vec{u}_0 = \vec{h}\big(\vec{y}\big)$ of some argument $\vec{y}$ that satisfies $f_{\vec{Y}}\big(\vec{y}\big) > 0$, (here, $f_{\vec{Y}}\big(\vec{y}\big)$) is the PDF of the jointly continuous random variables $Y_1, \ldots, Y_n$),
• then there is no other argument $\vec{\tilde{y}}$ that also satisfies $\vec{u}_0 = h(\vec{\tilde{y}})$ and $f_Y(\vec{\tilde{y}}) > 0$.

These two conditions guarantee the invertibility of the function $\vec{y} \mapsto \vec{u} = \vec{h}(\vec{y})$: This inverse function $\vec{h}^{-1}(\cdot)$ is defined by the relation

$$\vec{u} \;=\; \vec{h}(\vec{y}) \;\Leftrightarrow\; \vec{y} \;=\; \vec{h}^{-1}(\vec{u})\,.$$

Since the function values $\vec{y} = \vec{h}^{-1}(\vec{u})$ belong to $\mathbb{R}^n$, $\vec{h}^{-1}(\cdot)$ consists of n coordinate functions $h_1^{-1}(\cdot)$, $h_2^{-1}(\cdot), \ldots, h_n^{-1}(\cdot)$. They are defined by the equations

$$h_1^{-1}(\vec{u}) \;=\; y_1, \quad h_2^{-1}(\vec{u}) \;=\; y_2, \;\ldots,\; h_n^{-1}(\vec{u}) \;=\; y_n. \tag{12.14}$$

In the onedimensional case, the existence of continuous $\dfrac{dh}{du}$ which satisfies $\left|\dfrac{dh}{du}\right| \neq 0$ implies that of a continuous and non–zero derivative $\dfrac{dh^{-1}}{dy}$. Further,

$$\frac{dh^{-1}}{dy} \;=\; 1\Big/\frac{dh}{du}\,. \tag{12.15}$$

In the n–dimensional case, we must replace the condition $\left|\dfrac{dh}{du}\right| \neq 0$ with the condition

(5) $$J^{-1} \;:=\; \det \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \ldots & \frac{\partial h_1}{\partial y_n} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \ldots & \frac{\partial h_2}{\partial y_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial h_n}{\partial y_1} & \frac{\partial h_n}{\partial y_2} & \ldots & \frac{\partial h_n}{\partial y_n} \end{bmatrix} \;\neq\; 0\,.$$

The choice of the symbol J^{-1} for this determinant will become clear in a moment. The assumptions**(5)** and **(6)** are sufficient for the existence of all partial derivatives $\dfrac{\partial h_i^{-1}}{u_j}$ and their continuity. They form an $n \times n$ matrix and one can show that it's determinant, which we denote by J, also does not vanish. In other words,

$$J \;=\; \det \begin{bmatrix} \frac{\partial h_1^{-1}}{\partial u_1} & \frac{\partial h_1^{-1}}{\partial u_2} & \ldots & \frac{\partial h_1^{-1}}{\partial u_n} \\ \frac{\partial h_2^{-1}}{\partial u_1} & \frac{\partial h_2^{-1}}{\partial u_2} & \ldots & \frac{\partial h_2^{-1}}{\partial u_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial h_n^{-1}}{\partial u_1} & \frac{\partial h_n^{-1}}{\partial u_2} & \ldots & \frac{\partial h_n^{-1}}{\partial u_n} \end{bmatrix} \;\neq\; 0\,. \tag{12.16}$$

Moreover, the determinants J^{-1} and J satisfy the analogue of (12.15):

$$J^{-1} \;=\; \frac{1}{J}\,. \;\square \tag{12.17}$$

Before we examine how this material about the matrices of the partial derivatives and their determinants can be used to compute the joint PDF of the random vector $\vec{U}(\omega) \;=\; \vec{h}\big(\vec{Y}(\omega)\big)$ and before state our findings as a formal theorem, we illustrate the above with the following example.

Example 12.10 (The joint PDF of two independent, exponential random variables – Part 1)**.** In this twodimensional example, the function $\vec{h} = (h_1, h_2)$ is defined as follows:

$$u_1 := h_1(y_1, y_2) := 2y_1 + y_2\,, \tag{12.18}$$

$$u_2 := h_2(y_1, y_2) := y_1 - 2y_2\,. \tag{12.19}$$

(1) We show that this function can be inverted by solving these equations for $\vec{y} = (y_1, y_2)$.

- $u_1 - 2u_2 \overset{(12.18)}{=} y_2 + 4y_2 = 5y_2 \Rightarrow y_2 = u_1/5 - 2u_2/5\,.$
- Thus, $y_1 \overset{(12.19)}{=} u_2 + 2y_2 = u_2 + (1/5)[2u_1 - 4u_2] = (2u_1)/5 + u_2/5.$

We have found the inverse function $\vec{h}^{-1} = \left(h_1^{-1}, h_2^{-1}\right)$ to be

$$h_1^{-1}(u_1, u_2) = y_1 = \frac{1}{5}(2u_1 + u_2)\,, \tag{12.20}$$

$$h_2^{-1}(u_1, u_2) = y_2 = \frac{1}{5}(u_1 - 2u_2)\,. \tag{12.21}$$

We will continue in Example12.11 on p.304. □

In the introduction, we informally discussed the following result from multivariable calculus which we are rephrasing here in the language of joint PDFs of continuous random variables and which is at the heart of this section. It is so lengthy that we spread it over several boxes. As mentioned before, assume that $n \leq 3$ if you do not have sufficient knowledge of linear algebra.

Theorem 12.2.

- *Let $\vec{Y} = \left(Y_1, \ldots, Y_n\right)$ be a vector of randomvariables with joint PDF $f_{\vec{Y}}(\vec{y})$ and let R be a "nice" subset of $\mathbb{R}^n$ which is so big that it hosts all outcomes $\vec{Y}(\omega)$ of $\vec{Y}$.*
- *Let the function $\vec{h} : R \to \mathbb{R}^n;\ \vec{y} \mapsto \vec{u} = \vec{h}(\vec{y})$ satisfy the following.*

⊡ *$\vec{h}$ has continuous partial derivatives $\dfrac{\partial h_i}{y_j}$ for all $1 \leq i, j \leq n$.*

⊡ *If the vector $\vec{u}$ is a function value $\vec{u} = \vec{h}(\vec{y})$ of some argument $\vec{y}$ that satisfies $f_{\vec{Y}}(\vec{y}) > 0$, then there is no other argument $\tilde{\vec{y}}$ that satisfies all those conditions.*

Then $\vec{h}$ has an inverse $\vec{h}^{-1} = h_1^{-1}, h_2^{-1}, \ldots, h_n^{-1}$ which is defined by the relation

$$\vec{u} = \vec{h}(\vec{y}) \Leftrightarrow \vec{y} = \vec{h}^{-1}(\vec{u})\,.$$

We can write this for the coordinate functions $h_i(\cdot)$ and $h_j^{-1}(\cdot)$ as follows:

$$u_1 = h_1(\vec{y}),\ \ldots,\ u_n = h_n(\vec{y}) \quad \text{and} \quad y_1 = h_1^{-1}(\vec{u}),\ \ldots,\ y_n = h_n^{-1}(\vec{u}). \tag{12.22}$$

Also, all partial derivatives $\dfrac{\partial h_i^{-1}}{u_j}$ exist and are continuous for $1 \leq i, j \leq n$.

(12.23) *Let*
$$\frac{d\vec{h}}{d\vec{y}} := \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \dots & \frac{\partial h_1}{\partial y_n} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \dots & \frac{\partial h_2}{\partial y_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n}{\partial y_1} & \frac{\partial h_n}{\partial y_2} & \dots & \frac{\partial h_n}{\partial y_n} \end{bmatrix}, \qquad \frac{d\vec{h^{-1}}}{d\vec{u}} := \begin{bmatrix} \frac{\partial h_1^{-1}}{\partial u_1} & \frac{\partial h_1^{-1}}{\partial u_2} & \dots & \frac{\partial h_1^{-1}}{\partial u_n} \\ \frac{\partial h_2^{-1}}{\partial u_1} & \frac{\partial h_2^{-1}}{\partial u_2} & \dots & \frac{\partial h_2^{-1}}{\partial u_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n^{-1}}{\partial u_1} & \frac{\partial h_n^{-1}}{\partial u_2} & \dots & \frac{\partial h_n^{-1}}{\partial u_n} \end{bmatrix}.$$

(12.24) *Let*
$$J^{-1} := J^{-1}(\vec{y}) := det\left(\frac{d\vec{h}}{d\vec{y}}\right), \qquad J := J(\vec{u}) := det\left(\frac{d\vec{h^{-1}}}{d\vec{u}}\right).$$

- *We add another assumption:* $J^{-1}(\vec{y}) \neq 0$ *for all* y *that satisfy* $f_{\vec{Y}}(\vec{y}) > 0$.

(12.25) *Then* $J\big(h(\vec{y})\big) \neq 0$ *and* $J\big(h(\vec{y})\big) = 1\Big/J^{-1}(\vec{y})$.

Further, the density of the transform $\vec{U} = h\big(\vec{Y}\big)$ *is computed as*

(12.26) $$f_{\vec{U}}(\vec{u})\big) = f_{\vec{Y}}\big(h^{-1}(\vec{u})\big) \cdot |J(\vec{u})|.$$

PROOF: Beyond the scope of this course. It needs knowledge not only of linear algebra, but also of the so called implicit function theorem. ■

Before we give some examples to illustrate this theorem, we make a remark about some of the notation introduced there and then give a name to the determinant J^{-1} of the matrix $\frac{d\vec{h}}{d\vec{y}}$ of the partial derivatives of h.

Remark 12.5. In the onedimensional case ($n = 1$), the situation is as follows.

- $\mathbb{R}^n$ is the set $\mathbb{R}$ of real numbers, • $\vec{u} = \vec{h}\big(\vec{y}\big)$ becomes $u = h(y)$ for real numbers y and u,
- the 1×1 "matrix" of "partial" derivatives is $h'(y) = \frac{dh}{dy}$.

Considering that last point, it seems natural to write $\frac{d\vec{h}}{d\vec{y}}$ for the $n \times n$ matrix of partial derivatives $\frac{\partial h_i}{\partial y_j}$ and this author chose to do so. However, you will find either different notation [114] or, like in the WMS text, no dedicated symbols at all. That works well enough with 2×2 matrices. □

Definition 12.1 (Jacobian and Jacobian matrix)**.**

[114] For example, Williamson, Richard E. and Trotter, Hale [14] uses the notation $\vec{h}'(\vec{y})$, the multidimensional analogue of $h'(y)$.

The matrix $\frac{d\vec{h}}{d\vec{y}}$ of the partial derivatives of the function $\vec{y} \mapsto \vec{h}(\vec{y})$ is called the **Jacobian matrix** of $\vec{h}(\cdot)$. We refer to its determinant, $J^{-1}(\vec{y}) = \det\left(\frac{d\vec{h}}{d\vec{y}}\right)$, as the **Jacobian**, sometimes also the **Jacobian determinant**, of $\vec{h}(\cdot)$. □

Notation 12.1 (Jacobian)**.**

- Stewart writes $\frac{\partial(u_1,\ldots,u_n)}{\partial(y_1,\ldots,y_n)} := \det\left(\frac{d\vec{h}^{-1}}{d\vec{u}}\right)$ and $\frac{\partial(y_1,\ldots,y_n)}{\partial(u_1,\ldots,u_n)} := \det\left(\frac{d\vec{h}^{-1}}{d\vec{u}}\right)$
- Thus, the expression $J = J(\vec{u}) = \det\left(\frac{d\vec{h}^{-1}}{d\vec{u}}\right)$, which appears in (12.26) $f_{\vec{U}}(\vec{u})\big) = f_{\vec{Y}}\big(h^{-1}(\vec{u})\big) \cdot |J(\vec{u})|$, is the Jacobian of $h^{-1}(\vec{u})$ and not of $h(\vec{y})$.
- This author follows the great majority of books on multivariable calculus in defining the the Jacobian as the determinant of $\frac{d\vec{h}}{d\vec{y}}$.
- Be aware that WMS chooses instead to call $J = \det\frac{d\vec{h}^{-1}}{d\vec{u}}$ the Jacobian.
- The reason seems to be that most books on probability and statistics agree on using the letter J for $\det\frac{d\vec{h}^{-1}}{d\vec{u}}$ (without giving a name to that determinant) and WMS does not want to use the somewhat lengthy "the reciprocal of the Jacobian" in its frequent references to J

□

Example 12.11 (The joint PDF of two independent, exponential random variables – Part 2)**.** In Example 12.10 on p.302, we defined $\vec{u} = \vec{h}(\vec{y})$ as follows:

$$u_1 = h_1(y_1, y_2) = 2y_1 + y_2\,, \qquad u_2 = h_2(y_1, y_2) = y_1 - 2y_2\,.$$

We computed its inverse $\vec{u} = \vec{h}^{-1}(\vec{u}) =$ and obtained

$$y_1 = h_1^{-1}(u_1, u_2) = \frac{1}{5}(2u_1 + u_2)\,, \qquad y_2 = h_1^{-1}(u_1, u_2) = \frac{1}{5}(u_1 - 2u_2)\,.$$

Observe that both $\vec{h}$ and $\vec{h}^{-1}$ are defined for all points in $\mathbb{R}^2$.

The partial derivatives of $\vec{h}$ are

$$\frac{\partial h_1}{\partial y_1} = 2\,, \quad \frac{\partial h_1}{\partial y_2} = 1\,, \quad \frac{\partial h_2}{\partial y_1} = 1\,, \quad \frac{\partial h_2}{\partial y_2} = -2\,.$$

Those of $\vec{h}^{-1}$ are

$$\frac{\partial h_1^{-1}}{\partial u_1} = \frac{2}{5}\,, \quad \frac{\partial h_1^{-1}}{\partial u_2} = \frac{1}{5}\,, \quad \frac{\partial h_2^{-1}}{\partial u_1} = \frac{1}{5}\,, \quad \frac{\partial h_2^{-1}}{\partial u_2} = \frac{-2}{5}\,.$$

Further,

$$\frac{d\vec{h}}{d\vec{y}} = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}, \qquad \frac{d\vec{h}^{-1}}{d\vec{u}} = \begin{bmatrix} \frac{2}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{-2}{5} \end{bmatrix},$$

Since the determinant of a 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, is $ad - bc$, we obtain

$$J^{-1} = (2)(-2) - (1)(1) = -5, \qquad J = \left(\frac{2}{5}\right)\left(\frac{-2}{5}\right) - \left(\frac{1}{5}\right)\left(\frac{1}{5}\right) = \frac{-4-1}{25} = \frac{-1}{5},$$

Observe that $J = \dfrac{1}{J^{-1}}$, validates what was stated in (12.25) on p.303.
We will continue in Example12.12. □

Example 12.12 (The joint PDF of two independent, exponential random variables – Part 3)**.** In Example 12.10 on p.302, we defined $\vec{u} = \vec{h}(\vec{y})$ as follows:

(12.27) $$u_1 = h_1(y_1, y_2) = 2y_1 + y_2\,, \qquad u_2 = h_2(y_1, y_2) = y_1 - 2y_2\,.$$

In its continuation, Example 12.11 above, we obtained $J = \text{const} = \dfrac{-1}{5}$ for the reciprocal of the Jacobian of $\vec{h}$.
We are ready to specify the random variables that we wish to transform by means of $\vec{h}(\cdot)$.

- Assume that Y_1 and Y_2 are independent expon(2) random variables.
- Let $U_1 := h_1(\vec{Y}) = 2Y_1 + Y_2$, $\quad U_2 := h_2(\vec{Y}) = Y_1 + 2Y_2$.
- Apply Theorem 12.2 on p.302 to compute the joint density $f_{\vec{U}}(u_1, u_2)$ of $\vec{U} = \vec{h}(\vec{Y})$.

Solution:

(a) $$f_{\vec{Y}}(\vec{y}) = f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2) = \begin{cases} \frac{1}{4}\, e^{-(y_1+y_2)/2}, & \text{if } y_1, y_2 > 0\,, \\ 0, & \text{else}\,. \end{cases}$$

(b) We recall that $y_1 = \dfrac{1}{5}(2u_1 + u_2)$ and $y_2 = \dfrac{1}{5}(u_1 - 2u_2)$. Thus,

$$\begin{aligned} f_{\vec{U}}(\vec{u}) &= f_{U_1,U_2}(u_1, u_2) = \frac{1}{4}\, \exp\left\{ -\left(\frac{1}{5}(2u_1 + u_2) + \frac{1}{5}(u_1 - 2u_2) \right) \Big/ 2 \right\} \cdot \left| -\frac{1}{5} \right| \\ &= \frac{1}{20}\, \exp\left\{ \frac{-1}{10}\,(2u_1 + u_2 + u_1 - 2u_2) \right\} = \frac{1}{20}\, \exp\left\{ \frac{3u_1 - u_2}{-10} \right\} = \frac{1}{20}\, \exp\left\{ \frac{u_2 - 3u_1}{10} \right\}. \end{aligned}$$

- **BUT ONLY IF** $y_1 = h_1^{-1}(\vec{u}) \geq 0$ **AND** $y_2 = h_2^{-1}(\vec{u}) \geq 0$! What are those vectors $\vec{u}$?

(c) $y_1 \geq 0$ and $y_2 \geq 0 \Leftrightarrow 2u_1 + u_2 \geq 0$ and $u_1 - 2u_2 \geq 0$

(d) $y_1 \geq 0$ and $y_2 \geq 0 \overset{(12.27)}{\Rightarrow} u_1 = 2y_1 + y_2 \geq 0$.

(e) From **(c)**: $2u_1 + u_2 \geq 0 \Rightarrow u_2 \geq -2u_1$

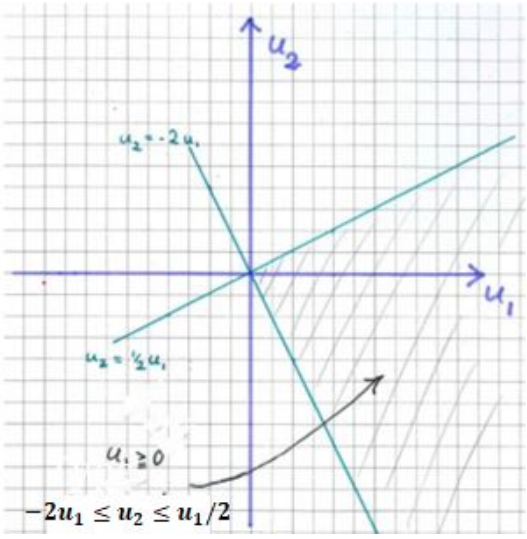
(f) From **(c)**: $u_1 - 2u_2 \geq 0 \Rightarrow u_1 \geq 2u_2$ $\Rightarrow u_2 \leq \dfrac{u_1}{2}$

(g) From **(d)**, **(e)**, **(f)**: $h_1^{-1}(\vec{u}) \geq 0$ and $h_2^{-1}(\vec{u}) \geq 0 \Leftrightarrow$ $u_1 \geq 0$ and $-2u_1 \leq u_2 \leq \dfrac{u_1}{2}$.

- The figure to the right shows that those are the points enclosed by the quadrant which is obtained when rotating the first quadrant clockwise, by an angle of $60°$

(h) Thus, if we denote this quadrant by R,

$$f_{\vec{U}}(\vec{u}) \;=\; \begin{cases} \dfrac{1}{20}\, e^{(u_2-3u_1)/10}, & \text{if } \vec{u} \in R\,, \\ 0, & \text{else}\,. \end{cases}$$



where $\mathbf{h_1^{-1}(u_1, u_2) > 0}$ and $\mathbf{h_2^{-1}(u_1, u_2) > 0}$

At this point we know how to integrate with respect to the PDF of $\vec{U} = \vec{h}(\vec{Y})$. We can replace the integral $d\vec{u}$ over the region R by an iterated integral $du_2\, du_1$ as follows.

For a fixed $u_1 > 0$, the integration bounds for u_2 are $-2u_1 \leq u_2 \leq \dfrac{u_2}{2}$. (See **(g)**). Thus,

$$\iint_{\mathbb{R}^2} \cdots f_{\vec{U}}(\vec{U})\, d\vec{u} \;=\; \iint_R \cdots \tfrac{1}{20}\, e^{(u_2-3u_1)/10}\, d\vec{u} \;=\; \int_0^\infty \int_{-2u_1}^{u_2/2} \cdots \tfrac{1}{20}\, e^{(u_2-3u_1)/10}\, du_2\, du_1$$

For example, if $w = g(\vec{U}) = g(u_1, u_2)$ is a real–valued function of $(u_1, u_2) \in \mathbb{R}^2$, then

$$E[g(\vec{U})] \;=\; =\; \int_0^\infty \int_{-2u_1}^{u_2/2} g(\vec{u})\tfrac{1}{20}\, e^{(u_2-3u_1)/10}\, du_2\, du_1 \;\square$$

12.4 The Method of moment–generating Functions

Assumption 12.1. Unless stated otherwise, we will assume in this entire section that

(a) $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$ denotes a list of n random variables ($n \in \mathbb{N}$).
- Either all Y_j are discrete, or they all are continuous random variables.

(b) $h : D \to \mathbb{R}; \quad \vec{y} \mapsto u = h(\vec{y}) = h(y_1, \ldots, y_n)$
is a function with domain $D \subseteq \mathbb{R}^n$ (this covers $\mathbb{R} = \mathbb{R}^1$ for $n = 1$), such that
- there is no issue with the existence of the PMF or PDF of $U := h(\vec{Y})$.
- All MGFs, $m_{Y_j}(t) = E\big[e^{tY_j}\big]$ and $m_U(t) = E\big[e^{tU}\big]$ exist if $|t|$ is small enough, i.e., there is some $\delta > 0$ such that those MGFs exist for $-\delta < t < \delta$.

(c) Those assumptions also hold for differently named (vectors of) random variables and functions, e.g. $V = g\big(\vec{\widetilde{Y}}\big) = g\big(\widetilde{Y}_1, \ldots, \widetilde{Y}_k\big)$. $\square$

Introduction 12.3. The moment–generating function method for finding the probability distribution of a function of random variables $Y_1, Y_2, \ldots, Y_n$ is based on Proposition 9.5 on p.208 (Section 9.5: Moments, Central Moments and Moment Generating Functions). It was stated without proof and asserts that the following is true under the conditions stated in Assumption 12.1:

Assume that two random variables Y and $\widetilde{Y}$ possess identical kth moments about the origin for all $k = 1, 2, \ldots$. In other words, assume that

$$E[Y^1] = E[\widetilde{Y}^1],\ E[Y^2] = E[\widetilde{Y}^2],\ E[Y^3] = E[\widetilde{Y}^3], \ldots$$

Then $P_Y = P\widetilde{Y}$, i.e., Y and $\widetilde{Y}$ have the same distribution. □

We have the following uniqueness theorem.

Theorem 12.3 (The MGF determines the distribution)**.**

Given are two random variables Y and $\widetilde{Y}$. If their moment–generating functions $m_Y(t)$ and $m_{\widetilde{Y}}(t)$ exist and coincide in a small interval that is centered at $t = 0$,

- *Then $P_Y = P_{\widetilde{Y}}$, i.e., Y and $\widetilde{Y}$ have the same probability distribution.*

PROOF:

Theorems 9.18 on p.209 and 10.9 on p.225 allow us to conclude that

$$E[Y^k] = \frac{d^k}{dt^k} m_Y(t)\Big|_{t=0} = \frac{d^k}{dt^k} m_{\widetilde{Y}}(t)\Big|_{t=0} = E[\widetilde{Y}^k] \text{ for all } k \in \mathbb{N}.$$

It follows from Proposition 9.5 on p.208 that $P_Y = P_{\widetilde{Y}}$ ■

Remark 12.6.

To find the distribution of $U = h(\vec{Y}) = h\big(Y_1, Y_2, \ldots, Y_n\big)$ by means of the MGF method, proceed as follows:

- Compute the MGF $m_U(t) = E\big[e^{tU}\big]$ of U
- Does this MGF match that of a random variable V with a known distribution? You may want to consult a list of MGFs like the one in Appendix 2 of [13] Wackerly, Mendenhall, Scheaffer, R.L.
- Then you are done, since Theorem 12.3 (The MGF determines the distribution) guarantees that $P_U = P_V$.

Of course, the devil is in the details. In most cases, you will not succeed in finding that matching MGF, unless one or both of the following are satisfied:

- U is a linear function $U = a_1 Y_1 + \cdots + a_n Y_n$, with constant $a_j \in \mathbb{R}$.
- The random variables $Y_1, \ldots, Y_n$ are independent and $h(\vec{y}) = h_1(y_1) \cdot h_2(y_2) \cdots h_n(y_n)$, for suitable functions $h_j(y)$.

We will examine some very important and general cases that illustrate all this. □

Example 12.13 (WMS Ch.06.5, Example 6.10)**.** Suppose that Y is a normally distributed random variable with mean μ and variance σ^2. Show that

$$Z := \frac{Y - \mu}{\sigma}$$

has a standard normal distribution, i.e., $Z \sim \mathcal{N}(0,1)$.

Solution:

(a) According to Proposition 10.6 on p.231, $m_Y(t) = e^{\mu t + (\sigma^2 t^2)/2}$.

(b) Any random variable W is independent from any constant (real number) a.

(c) Thus, according to Theorem 11.10 on p.256, the random variables $h_1(W) = e^{tW}$ and $h_2(a) = e^{-at}$ are independent, and $E[e^{tW} \cdot e^{-at}] = E[e^{tW}] \cdot e^{-at}$.

(d) Thus if $U = Y - \mu$, then $m_U(t) = E\big[e^{tY-t\mu}\big] = E\big[e^{tY}\, e^{-t\mu}\big] = E\big[e^{tY}\big] \cdot e^{-t\mu}$.

- Thus, $m_U(t) = m_Y(t)\, e^{-t\mu} \overset{\text{(a)}}{=} e^{\mu t + (\sigma^2 t^2)/2} \cdot e^{-t\mu} = e^{(\sigma^2 t^2)/2}$.
- Since $Z = U/\sigma$, $m_Z(t) = m_U(t/\sigma) = e^{(\sigma^2 (t/\sigma)^2/2} = e^{t^2/2}$.

(e) We use Proposition 10.6 once more and see that $t \mapsto e^{t^2/2}$ is the MGF of a standard normal random variable. Thus, $Z \sim \mathcal{N}(0,1)$. □

Example 12.14 (WMS Ch.06.5, Example 6.11)**.** Let Z be a normally distributed random variable with mean 0 and variance 1. Use the method of moment–generating functions to find the probability distribution of Z^2.

Solution:

The moment–generating function for Z^2 is

$$\textbf{(A)} \qquad \begin{aligned} m_{Z^2}(t) = E(e^{tZ^2}) &= \int_{-\infty}^{\infty} e^{tz^2} f(z)\, dz = \int_{-\infty}^{\infty} e^{tz^2} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z^2/2)(1-2t)}\, dz = \int_{-\infty}^{\infty} \psi(z)\, dz\,, \end{aligned}$$

where

$$\begin{aligned} \psi(z) &= \exp\left[-\left(\frac{z^2}{2}\right)(1-2t)\right] \Big/ \sqrt{2\pi} \\ &= \exp\left[-\left(\frac{z^2}{2}\right) \Big/ (1-2t)^{-1}\right] \Big/ \left(\sqrt{2\pi}(1-2t)^{-1/2} \cdot \frac{1}{(1-2t)^{-1/2}}\right). \end{aligned}$$

We define $\sigma := (1-2t)^{-1/2}$ and obtain

$$\psi(z) = \exp\left[-\left(\frac{z^2}{2}\right) \Big/ \sigma^2\right] \Big/ \left(\sqrt{2\pi}\,\sigma \cdot \frac{1}{\sigma}\right) = e^{-z^2/(2\sigma^2)} \cdot \frac{\sigma}{\sqrt{2\pi}\,\sigma} = \sigma\, \varphi(z)\,,$$

where $\varphi(z)$ is the density of a $\mathcal{N}(0,\sigma)$ random variable. Thus, $\int_{-\infty}^{\infty} \varphi(z)\, dz = 1$. It follows from **(A)** and $\psi(z) = \sigma\, \varphi(z)$ and $\sigma := (1-2t)^{-1/2}$ that

$$m_{Z^2}(t) = \int_{-\infty}^{\infty} \psi(z)\, dz = \int_{-\infty}^{\infty} (1-2t)^{-1/2} \varphi(z)\, dz = \frac{1}{(1-2t)^{1/2}} \int_{-\infty}^{\infty} \varphi(z)\, dz = \frac{1}{(1-2t)^{1/2}}\,.$$

According to Proposition 10.8 on p.234, $t \mapsto \dfrac{1}{(1-2t)^{1/2}}$ is the MGF of a random variable which follows a gamma$(1/2, 2)$ distribution which is, by definition 10.11 on p.235, also known as a χ^2 distribution with one degree of freedom. We obtained this result previously in Example 12.5 on p.294 by the method of distribution functions. □

Theorem 12.4 (MGF of a sum of functions of independent variables)**.**

Given are n independent random variables $Y_1, Y_2, \ldots, Y_n$ with MGFs $m_{Y_1}(t), m_{Y_2}(t), \ldots, m_{Y_n}(t)$. and n real–valued functions $h_1(y_1), \ldots, h_n(y_n)$ of real numbers $y_1, \ldots, y_n$.
Let $U := h_1(Y_1) + h_2(Y_2) + \cdots + h_n(Y_n)$. Then (under the conditions of Assumption 12.1 on 306)

$$m_U(t) = m_{h_1(Y_1)+\cdots+h_n(Y_n)} = \prod_{j=1}^{n} m_{h_j(Y_j)}(t)\,. \tag{12.28}$$

PROOF:

For each $j = 1, \ldots, n$, let $g_j(y) := e^{th_j(y)}$. Consider a fixed t. Since functions of independent random variables are independent random variables, the random variables $V_j := g_j(Y_j) = e^{th_j(Y_j)}$ are independent. We apply Theorem 11.10 on p.256 and obtain

$$\begin{aligned} m_U(t) &= E\left[e^{t(V_1+V_2+\cdots+V_n)}\right] \\ &= E\left[e^{tV_1}\right]\cdots E\left[e^{tV_n}\right]. \;=\; E\left[e^{th_1(Y_1)}\right]\cdots E\left[e^{th_n(Y_n)}\right] \\ &= m_{h_1(Y_1)}(t)\cdot m_{h_1(Y_1)}(t)\cdots m_{h_1(Y_n)}(t)\,. \;\blacksquare \end{aligned}$$

Corollary 12.1 (WMS Ch.06.5, Theorem 6.2)**.**

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with moment–generating functions $m_{Y_1}(t), m_{Y_2}(t), \ldots, m_{Y_n}(t)$, respectively. Then

$$m_{Y_1+\cdots+Y_n}(t) = \prod_{j=1}^{n} m_{Y_j}(t) = m_{Y_1}(t)\cdot m_{Y_2}(t)\cdots m_{Y_n}(t)\,. \tag{12.29}$$

PROOF:

This follows from applying Theorem 12.4 to the functions $h_j(y_j) = y_j$. $\blacksquare$

Next, we generalize But its great importance gives it the status of a theorem.

Theorem 12.5 (Linear combinations of uncorrelated normal variables are normal)**.**

Given are n uncorrelated, $\mathcal{N}(\mu_j, \sigma_j^2)$ random variables Y_j, $(j = 1, \ldots, n$. In other words, each Y_j is normal with expectation μ_j and standard deviation σ_j. Let $a_1, \ldots, a_n \in \mathbb{R}$. Then

$$\sum_{j=1}^{n} a_j Y_j \;\sim\; \mathcal{N}\left(\sum_{j=1}^{n} a_j\mu_j\,, \sum_{j=1}^{n} a_j^2\sigma_j^2\right)\,. \tag{12.30}$$

Thus, the linear combination of uncorrelated normal random variables is normal with expectation and variance being the linear combinations of the indivicual expectations and variances.

PROOF:

First off, we recall that one of the special properties of normal random variables is that they are uncorrelated if and only if they are independent. Thus we can use everything that applies to independent random variables.

Consider a fixed t and define

$$U := \sum_{j=1}^{n} a_j Y_j \,.$$

We apply Theorem 12.4 (MGF of a sum of functions of independent variables) on p.309 with the functions $h_j(y_j) = a_j y_j$ and obtain

$$\begin{aligned} m_U(t) &= \prod_{j=1}^{n} m_{a_j Y_j}(t) = \prod_{j=1}^{n} m_{Y_j}(a_j t) \\ &= \prod_{j=1}^{n} \exp\left\{(\sigma_j^2/2)(a_j t)^2 + \mu_j (a_j t)\right\} \end{aligned}$$

Here we used that a $\mathscr{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ variable has MGF $e^{\widetilde{\sigma}^2 t^2/2 + \widetilde{\mu} t}$. See Proposition 10.6 on p.231. Thus,

$$\begin{aligned} m_U(t) &= \exp\left\{\sum_{j=1}^{n} (\sigma_j^2/2)(a_j t)^2 + \mu_j (a_j t)\right\} \\ &= \exp\left\{\left(\sum_{j=1}^{n} (\sigma_j^2 a_j^2/2)\, t^2\right) + \left(\sum_{j=1}^{n} (\mu_j a_j)\, t\right)\right\} \\ &= \exp\left\{\left(\sum_{j=1}^{n} (a_j^2 \sigma_j^2)\right) \Big/ 2 \cdot t^2 + \left(\sum_{j=1}^{n} (a_j \mu_j)\right) \cdot t\right\} \end{aligned}$$

By Proposition 10.6, the last expression is the MGF of a $\mathscr{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ variable with

$$\widetilde{\mu} = \sum_{j=1}^{n} (a_j \mu_j)\,, \qquad \widetilde{\sigma}^2 = \sum_{j=1}^{n} (a_j^2 \sigma_j^2)\,.$$

Since distributions of random variables are determined by their MGFs,

$$U \sim \mathscr{N}\left(\sum_{j=1}^{n} a_j \mu_j \,, \sum_{j=1}^{n} a_j^2 \sigma_j^2\right) . \quad \blacksquare$$

Remark 12.7. It is a consequence of Theorem 12.5 that the sum of two independent random variables also is normal. In the next example we construct two normal variables U and W which are not independent, such that their sum is not normal. It shows that we cannot drop the assumption of independence in Theorem 12.5. This example is given in many books on probability. It can be found, e.g., in [7] Pishro-Nik, Hossein: Introduction to Probability, Statistics, and Random Processes.

Assume that U and V are independent random variables with distributions

- $U \sim \mathcal{N}(0,1)$,
- $V \sim \text{binom}(n=1, p=0.5)$.

Let

$$W(\omega) \;:=\; \begin{cases} U(\omega), & \text{if } V(\omega)=1, \\ -U(\omega), & \text{if } V(\omega)=0\,. \end{cases}$$

(a) Show that $W \sim \mathcal{N}(0,1)$.
(b) Let $Y := U + W$. Show that Y is not a continuous random variable.

It follows from **(b)** that Y is not normal, since normal random variables are continuous. □

Solution to (a): Note that the PDF of U is symmetric, i.e., $f_U(u) = f_U(-u)$ for all $u \in \mathbb{R}$. Thus, for all u,

$$P\{U \le u\} \;=\; \int_{-\infty}^{u} f_U(t)dt \;=\; \int_{-u}^{\infty} f_U(t)dt \;=\; P\{U \ge -u\} \;=\; P\{-U \le u\}\,.$$

It follows that U and $-U$ have the same distribution and thus, $-U \sim \mathcal{N}(0,1)$. [115]
Now, we show that $W \sim \mathcal{N}(0,1)$. Let $w \in \mathbb{R}$. Then,

$$\begin{aligned} P\{W \le w\} &= P\{W \le w, V=0\} \;+\; P\{W \le w, V=1\} \\ &= P\{W \le w \mid V=0\}\, P\{V=0\} \;+\; P\{W \le w \mid V=1\}\, P\{V=1\} \\ &= \frac{1}{2}\, P\{-U \le w \mid V=0\} \;+\; \frac{1}{2}\, P\{U \le w \mid V=1\} \end{aligned}$$

We use the independence of U and V followed by $U \sim -U$ and obtain

$$P\{W \le w\} \;=\frac{1}{2}\left(P\{-U \le w\} \;+\; P\{U \le w\}\right) \;=\frac{1}{2}\left(P\{U \le w\} \;+\; P\{U \le w\}\right) \;=\; P\{U \le w\}\,.$$

Thus, $W \sim U$. Since U is standard normal, so is W. We have proven **(a)**.

Solution to (b): It follows from the definition of W and $Y := U + W$, that

$$Y(\omega) \;:=\; \begin{cases} 2U(\omega), & \text{if } V(\omega)=1, \\ 0, & \text{if } V(\omega)=0\,. \end{cases}$$

Since U is a continuous random variable, $P\{2U=0\} = 0$.

$$\begin{aligned} &\text{Thus,} \qquad P\{2U=0, V=1\} \;\le\; P\{2U=0\} \;=\; 0\,. \\ &\text{Thus,} \qquad P\{Y=0\} \;=\; P\{2U=0, V=1\} \;+\; P\{V=0\} \;=\; P\{V=0\} \;=\; \frac{1}{2}\,. \end{aligned}$$

It follows that the CDF F_Y of Y has a jump

- $F_Y(0) \;-\; F_Y(0-) \;=\; P\{Y=0\} \;=\; 1/2$

at $y=0$. Thus, Y is not a continuous random variable and we have shown **(b)**. ■

[115] This result should not come as a surprise since, for $n=1$ and $a_1=-1$, Theorem 12.5 on p.309 states the following: If $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$, then $-Y_1 \sim \mathcal{N}(-\mu, \sigma^2)$. Note though, that the proof given here shows that U and $-U$ have the same distribution whenever U has a symmetric PDF. Also note that $U \sim -U$ holds if U is discrete with a symmetric PMF, i.e., $p_U(u) = P\{U=u\} = P\{U=-u\} = p_U(-u)$, for all u.

Theorem 12.6.

Given are n independent, gamma(α_j, β) random variables Y_j, $(j = 1, \ldots, n$. In other words, each Y_j is gamma with the same scale parameter β. Then

$$\sum_{j=1}^{n} Y_j \sim \text{gamma}\left(\sum_{j=1}^{n} \alpha_j, \beta\right). \tag{12.31}$$

Thus, the sum of independent gamma random variables with the same scale parameter β is gamma with the shape parameter being the sum of the shape parameters, and scale parameter β.

PROOF:

Consider a fixed t and define

$$U := \sum_{j=1}^{n} Y_j .$$

We apply Theorem 12.4 (MGF of a sum of functions of independent variables) on p.309 and recall that the MGF of a gamma$(\widetilde{\alpha}, \widetilde{\beta})$ variable $\widetilde{Y}$ is, according to Proposition 10.8 on p.234, $m_{\widetilde{Y}} = (1-\widetilde{\beta}t)^{\widetilde{\alpha}}$. We obtain

$$\begin{aligned} m_U(t) &= m_{\sum_j Y_j}(t) = \prod_{j=1}^{n} m_{Y_j}(t) \\ &= \prod_{j=1}^{n} \frac{1}{(1-\beta t)^{\alpha_j}} = \frac{1}{(1-\beta t)^{\sum_{j=1}^{n} \alpha_j}} . \end{aligned}$$

Since distributions of random variables are determined by their MGFs,

$$U \sim \text{gamma}\left(\sum_{j=1}^{n} \alpha_j, \beta\right). \quad \blacksquare$$

Corollary 12.2.

Let $Y_1, Y_2, \ldots, Y_n$ be independent χ^2 variables such that each Y_j has ν_j degrees of freedom. Then

$$m_{Y_1 + \cdots + Y_n}(t) \sim \chi^2\left(\sum_{j=1}^{n} \nu_j \; df\right). \tag{12.32}$$

PROOF:

This follows immediately from Theorem 12.6, Since χ^2 variables with ν_j df are gamma$(\nu_j/2, 2)$. $\blacksquare$

13 Limit Theorems

Introduction 13.1. In this section we will discuss the ways in which a sequence Y_n of random variables can have a random variable Y as its limit. Before we go there, let us quickly review convergence of a sequence $(y_n)_n$ of real numbers and of a sequence of functions $f_n : A \to \mathbb{R}$, with all members f_n defined on a subset A of $\mathbb{R}^k$, where $k = 1, 2, \ldots$. Note that $k = 1$ covers the situation where the arguments are real numbers. Some examples of number sequences:

- If $y_n = \dfrac{3-2n}{5+n^2-6n}$, then $\lim\limits_{n\to\infty} y_n = \dfrac{3}{5}$, and the sequence converges to $\dfrac{3}{5}$.
- If $y_n = (-1)^n$, then $\lim\limits_{n\to\infty} y_n$ does not exist.
- If $y_n = \sum\limits_{j=1}^{n} n$, then $\lim\limits_{n\to\infty} y_n = \infty$. Recall that convergence only happens if the limit is a real number. Thus, $(y_n)_n$ does not "converge to ∞". Rather, this sequence diverges. [116]

For the following examples of function sequences, let us agree that, if $f_n, f : A \to \mathbb{R}$, where $A \subseteq \mathbb{R}$, then "pointwise convergence" [117] of the functions f_n to the function f simply means that

$$\lim_{n\to\infty} f_n(a) = f(a) \qquad \text{for all } a \in A\,. \tag{13.1}$$

- Let $f_n, f, g, h : [0,1] \to \mathbb{R}$ be the functions

$$\boxdot\, f_n(x) := x^n \quad \boxdot\, f(x) := \begin{cases} 0\,, & \text{if } 0 \le x < 1, \\ 1\,, & \text{if } x = 1\,, \end{cases} \quad \boxdot\, g(x) := 0, \quad \boxdot h(x) := x. \tag{13.2}$$

The situation with respect to pointwise convergence is as follows:

- f is the pointwise limit of the sequence f_n.
- Even though g is the pointwise limit of the sequence f_n on $[0,1[$, it is not the pointwise limit on $[0,1]$, since $\lim\limits_{n\to\infty} f_n(x) = g(x) = 0$, for $0 \le x < 1$, but $\lim\limits_{n\to\infty} f_n(1) = 1$, whereas $g(1) = 0$.
- h is not the pointwise limit of the sequence f_n (except on $\{0,1\}$).

Did you notice that no use was made of the fact that the domain $[0,1]$ of those functions is a set of numbers?

- Assume instead that Ω is some arbitrary, nonempty set (not necessarily a probability space). Further assume that there are functions $f_n, f : \Omega \to \mathbb{R}$. We still have the notion of pointwise convergence of the functions f_n to the function f: (13.1) becomes

  $$\lim_{n\to\infty} f_n(\omega) = f(\omega) \qquad \text{for all } \omega \in \Omega \tag{13.3}$$

  and one certainly can examine whether or not the above is true for any kind of domain, i.e., for any nonempty set Ω.

We will not discuss vector–valued sequences. However, for completeness sake, we give the following example.

[116] There is no such thing as divergence to $\pm\infty$. Thus, you must say that (y_n) diverges, **not** that (y_n) diverges to ∞.

[117] The formal definition of pointwise limits will be given in Section 13.1 (Four Kinds of Limits for Sequences of Random Variables).

- If $\vec{y}_n = \big((-1)^n, \cos(2/n)\big)$, then $\lim_{n\to\infty} \vec{y}_n$ does not exist, since the limit of a vector–valued sequence is, by definition, the vector of the limits of the coordinates. The second coordinate sequence, $y_n = \cos(2/n)\big)$, converges to the number 1. Since the first coordinate sequence, $y_n = (-1)^n$, does not have a limit, neither does $\big(\vec{y}_n\big)_n$. Thus this sequence does not converge.

After these preliminary remarks, let us consider sequences of random variables. We recall that all random variables Y are functions

$$Y : (\Omega, \mathfrak{F}, P) \to \mathbb{R} \qquad \omega \mapsto Y(\omega)\,.$$

They take their arguments ω in a probability space $(\Omega, \mathfrak{F}, P)$ and map them to numeric outcomes $y = Y(\omega)$.

- The σ–algebra is of no significance in this chapter, so we keep ignoring it and simply consider the probability space (Ω, P).
- On the other hand, the arguments ω play an essential role and we will often replace "Y" with "$\omega \mapsto Y(\omega)$" to remind the reader that we are dealing with functions of ω.
- If $(Y_n)_n$ is a sequence of random variables $(\Omega, P) \to \mathbb{R}$. Then each $\omega \in \Omega$ comes with its own sequence $\big(Y_n(\omega)\big)_n$ of real numbers.
- One obvious question to ask about those sequences $Y_n(\omega)$ of real numbers is this one:
  ⊡ Does $\lim_{n\to\infty} Y_n(\omega)$ exist and will it be a real number (rather than $\pm\infty$) for all $\omega \in \Omega$?
  ⊡ If so, then the assignment $\omega \mapsto Y(\omega) := \lim_{n\to\infty} Y_n(\omega)$ defines a real–valued function $Y : (\Omega, P) \to \mathbb{R}$, i.e., another random variable. What are its properties?
- Not quite so obvious: ⊡ Does the presence of the probability measure P on Ω give additional insight about the convergence behavior of the functions $\omega \mapsto Y_n(\omega)$?
- In contrast to the deterministic case where the only mode of convergence available to us is pointwise convergence, [118] we will see in Section 13.1 (Four Kinds of Limits for Sequences of Random Variables) that the presence of a probability P allows us to consider additional modes of convergence:
  ⊡ convergence almost surely,
  ⊡ convergence in probability measure,
  ⊡ convergence in distribution. □

13.1 Four Kinds of Limits for Sequences of Random Variables

The following definition is a central place for all the different convergence modes of sequences of random variables that are of interest to us. We will examine each one in detail.

Definition 13.1 (Convergence of Random Variables)**.**

[118] This is not entirely true: If Ω is a subset of $\mathbb{R}$ or of $\mathbb{R}^k$. then there is the notion of **uniform convergence**, $f_n(\cdot) \to f(\cdot)$. We will not be concerned with uniform convergence in this course.

Let Y_n $(n \in \mathbb{N})$ and Y be random variables on a probability space (Ω, P). We define

(13.4) $Y_n \overset{\text{pw}}{\to} Y$ or $\text{pw} - \lim_{n\to\infty} Y_n = Y$, if $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$, for all $\omega \in \Omega$,

(13.5) $Y_n \overset{\text{a.s.}}{\to} Y$ or $\text{a.s.} - \lim_{n\to\infty} Y_n = Y$, if $P\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\} = 1$,

(13.6) $Y_n \overset{\text{P}}{\to} Y$ or $\text{P} - \lim_{n\to\infty} Y_n = Y$, if $\forall\, \varepsilon > 0 \ \lim_{n\to\infty} P\{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \varepsilon\} = 0$,

(13.7) $Y_n \overset{\text{D}}{\to} Y$, if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$, $\forall\, y \in \mathbb{R}$ where the CDF F_Y of Y is continuous.

We also say:
If $Y_n \overset{\text{pw}}{\to} Y$, Y is the **pointwise limit** of the Y_n, or: Y_n **converges pointwise** to Y.
If $Y_n \overset{\text{a.s.}}{\to} Y$, Y is the **almost sure limit** of the Y_n, or: Y_n **converges almost surely** to Y.
If $Y_n \overset{\text{P}}{\to} Y$, Y is the **limit in probability**; of the Y_n, or: Y_n **converges in probability** to Y.
If $Y_n \overset{\text{D}}{\to} Y$, Y is the **limit in distribution** of the Y_n, or: Y_n **converges in distribution** to Y.

Example 13.1. Consider $\Omega := [0,1]$ as a probability space (Ω, P) by defining

$$P(]a,b]) := b - a, \text{ for } 0 \le a < b \le 1.$$

In other words, P is the uniform distribution on $[0,1]$.

We rename the functions f_n, f, g, h of (13.2) in the introduction to Y_n, Y, U, V, since doing so will make it less confusing to examine the convergence behavior of the sequence. This particularly applies to converges in probability and in distribution. Accordingly, we define

$$Y_n(\omega) := \omega^n, \ U(\omega) = 0, \ V(\omega) := \omega, \ (\text{for } 0 \le \omega \le 1) \quad Y(\omega) := \begin{cases} 0, & \text{if } 0 \le \omega < 1, \\ 1, & \text{if } \omega = 1. \end{cases}$$

Part I: Pointwise and a.s convergence

Pointwise convergence behavior of the Y_n corrresponds to that of (13.2):

- Y is the pointwise limit of the sequence Y_n,
- U is the pointwise limit of the Y_n on $[0,1[$ only, but not on Ω,
- V is not the pointwise limit of the Y_n (except for $\omega = 0$) or $\omega = 1$).

With respect to almost sure convergence, we see that

- $Y_n \overset{\text{a.s.}}{\to} Y$, since $\{\lim_{n\to\infty} Y_n = Y\} = [0,1] = \Omega$, and $P(\Omega) = 1$.
- $Y_n \overset{\text{a.s.}}{\to} U$, since $\{\lim_{n\to\infty} Y_n \ne U\} = \{1\}$, and $P(\{1\}) = 0$.
- $(Y_n)_n$ does not converge to V a.s., since $P\{\lim_{n\to\infty} Y_n = V\} = P\{0,1\} = 0 \ne 1$.

Part II: Convergence in probability

Next, we examine convergence in probability. We will see that a sequence of random variables can have more than one P–limit by showing the following: The sequence $\omega \mapsto Y_n(\omega) = \omega^n$ has both $\omega \mapsto U(\omega) = 0$ and $\omega \mapsto Y(\omega) = 1$ if $\omega = 1$ and 0 else as P–limits.

By definition of $P\text{–}\lim_{n\to\infty} Y_n = \widetilde{Y}$, we must prove that, for any fixed, but arbitrary $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{\, |Y_n - \widetilde{Y}| > \varepsilon \,\} \;=\; 0. \qquad \text{See (13.6).}$$

Since this probability decreases as ε increases and we must show that it approaches 0 as $n \to \infty$, we only need to worry about the very small ε. Thus, we may assume that $0 < \varepsilon < 1$.
We observe that, for $Y_n(\omega) = \omega^n$ and $0 < \varepsilon < 1$,

(A)
$$\begin{aligned} &\big[\, |Y_n(\omega)| \geq \varepsilon \;\Leftrightarrow\; \omega^n \geq \varepsilon \;\Leftrightarrow\; \omega \geq \varepsilon^{1/n} \,\big] \\ \Rightarrow\; &\big[\, P\{\, |Y_n| \geq \varepsilon \,\} \;=\; P\big(\,[\varepsilon^{1/n}, 1]\,\big) \;=\; 1 \,-\, \varepsilon^{1/n} \,\big]\,. \end{aligned}$$

(B)
$$0 < \varepsilon < 1 \;\Rightarrow\; \lim_{n\to\infty} \varepsilon^{1/n} \;=\; 1 \;\Rightarrow\; \lim_{n\to\infty} \big(1 - \varepsilon^{1/n}\big) \;=\; 0.$$

Part II (1): We now prove that $P\text{–}\lim_{n\to\infty} Y_n = Y$:

(a)
$$\begin{aligned} &\big[\, |Y_n(\omega) \,-\, Y(\omega)| \geq \varepsilon \;\Leftrightarrow\; |Y_n(\omega)| \geq \varepsilon \text{ and } \omega \neq 1 \,\big] \\ \Rightarrow\; &\big[\, P\{|Y_n \,-\, Y| \geq \varepsilon\} \;\leq\; P\{|Y_n| \geq \varepsilon\} \overset{\textbf{(A)}}{=} 1 \,-\, \varepsilon^{1/n} \overset{\textbf{(B)}}{\to} 0, \text{ as } n \to \infty. \,\big]\,. \end{aligned}$$

Thus, $\lim_{n\to\infty} P\{|Y_n \,-\, Y| \geq \varepsilon\} \;=\; 0$.

Part II (2): We now prove that $P\text{–}\lim_{n\to\infty} Y_n = U$:

- We could repeat the proof for the P–convergence of Y_n to Y with very minor modifications and the reader is encouraged to do so. Instead, we will use that result to show that $P\text{–}\lim_{n\to\infty} Y_n = U$
- Since the outcome $\{1\}$ has probability zero and $Y(\omega) = U(\omega)$ for $\omega \neq 1$,

$$\begin{aligned} P\{|Y_n \,-\, Y| \geq \varepsilon\} \;&= P\{|Y_n \,-\, Y| \geq \varepsilon \text{ and } \omega \neq 1\} \\ &= P\{|Y_n \,-\, U| \geq \varepsilon \text{ and } \omega \neq 1\} \;=\; P\{|Y_n \,-\, U| \geq \varepsilon\}\,. \end{aligned}$$

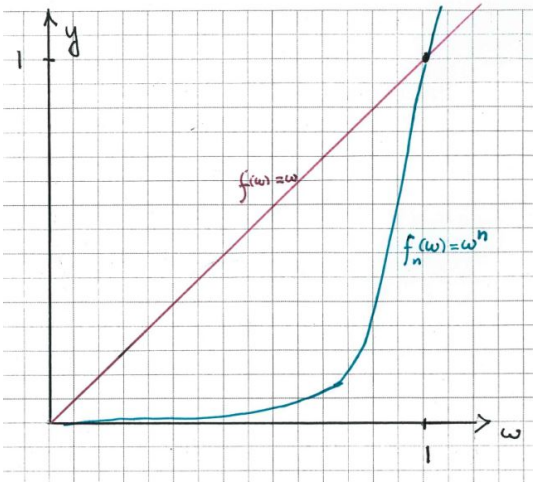
- Since $\lim_{n\to\infty} P\{|Y_n \,-\, Y| \geq \varepsilon\} \;=\; 0$,

$$\lim_{n\to\infty} P\{|Y_n \,-\, U| \geq \varepsilon\} \;=\; \lim_{n\to\infty} P\{|Y_n \,-\, Y| \geq \varepsilon\} \;=\; 0.$$

Thus, $P\text{–}\lim_{n\to\infty} Y_n = U$.

Part II (3): Next, we show that it is not true that $(Y_n)_n$ converges in probability to V.
We argue by picture rather than giving an exact proof, since that would require some very tedious of terms containing $\ln(k)$.

- The picture makes it very clear that $\varepsilon = 1/10 \Rightarrow \omega - \omega^n > \varepsilon$ for $\frac{49}{100} \leq \omega \leq \frac{51}{100}$ and $n \geq 100$.
  Thus, $P\{|Y_n - V| \geq \varepsilon\} \geq \varepsilon \cdot \left(\frac{51}{100} - \frac{49}{100}\right) = \frac{2}{1000}$.
  Thus, $\lim_{n\to\infty} P\{|Y_n - V| \geq \varepsilon\} = 0$ is not true.
- Since $\lim_{n\to\infty} P\{|Y_n - V| \geq \varepsilon\} = 0$ must hold for **ALL** ε and we showed that this is not so for $\varepsilon = \dfrac{1}{10}$, it follows that $(Y_n)_n$ does not converge in probability to V.



Part III: Convergence in distribution

We will show that Y_n does not converge to V in distribution as follows.

- Let $0 < y < 1$. We recall that $P]a,b] = b - a$, for all $0 \leq a < b \leq 1$.
- From $V(\omega) = \omega$, we get $F_V(y) = P\{V \leq y\} = P\{\omega \in \Omega : V(\omega) \leq y\} = P]0,y] = y$.
- Since $Y_n(\omega) = \omega^n$, $F_{Y_n}(y) = P\{Y_n \leq y\} = P\{\omega \in \Omega : \omega^n \leq y\} = P]0, y^{1/n}] = y^{1/n}$.
- We note that $0 < y < 1 \Rightarrow \lim_{n\to\infty} y^{1/n} = 1$.
  Thus, $F_V(y) = y$, whereas, $\lim_{n\to\infty} F_{Y_n}(y) = 1$ for $0 < y < 1$.
  Thus, $\lim_{n\to\infty} F_{Y_n}(y) \neq F_V(y)$ for $0 < y < 1$.
- Since all those y are points of continuity for F_V, it follows that $(Y_n)_n$ does not converge in distribution to V.

On the other hand, the theorem that follows now shows that $(Y_n)_n$ converges in distribution to Y and U, since we have shown convergence in probability to those random variables. □

Theorem 13.1 (Relationship between the modes of convergence)**.**

Let Y and $Y_1, Y_2, \ldots$ be random variables on a probability space (Ω, P). Then,

(13.8) $$Y_n \overset{pw}{\to} Y \;\Rightarrow\; Y_n \overset{a.s.}{\to} Y \;\Rightarrow\; Y_n \overset{P}{\to} Y \;\Rightarrow\; Y_n \overset{D}{\to} Y\,.$$

PROOF:

I: It is obvious that $Y_n \overset{\mathbf{pw}}{\to} Y \;\Rightarrow\; Y_n \overset{\mathbf{a.s.}}{\to} Y$ for the following reason:

- Let $A := \{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) \neq Y(\omega)\}$.
- Then, $Y_n \overset{\mathbf{pw}}{\to} Y \;\Rightarrow\; A = \emptyset \;\Rightarrow\; P(A) = 0 \;\Rightarrow\; Y_n \overset{\mathbf{a.s.}}{\to} Y$.

II: The proofs that $Y_n \overset{\mathbf{a.s.}}{\to} Y \;\Rightarrow\; Y_n \overset{\mathbf{P}}{\to} Y$ and $Y_n \overset{\mathbf{P}}{\to} Y \;\Rightarrow\; Y_n \overset{\mathbf{D}}{\to} Y$ are outside the scope of this course. Fairly accessible proofs for those who can work with sets like

$$\bigcap_{n\geq 1}\left(\bigcup_{j\geq n}\{\omega\in\Omega : |Y_j(\omega)-Y(\omega)| \geq \varepsilon\}\right)$$

and are familiar with the exact definition of convergence of sequences [119] can be found at this Wikipedia link. ■

There are many theorems concerning the convergence of random variables. We only mention here the following two which will be used later in this chapter.

Theorem 13.2 (Slutsky's Theorem). ⋆

Let $Y_1, Y_2, \dots)$ and $U_1, U_2, \dots$ be two sequences of random variables. Let Y be another random variable and c a constant such that

- $Y_n \xrightarrow{D} Y$ *(convergence in distribution)*
- $U_n \xrightarrow{P} c$ *(convergence in probability)*

Then,

$$Y_n + U_n \xrightarrow{D} Y + c\,, \tag{13.9}$$

$$Y_n \cdot U_n \xrightarrow{D} cY\,, \tag{13.10}$$

$$\frac{Y_n}{U_n} \xrightarrow{D} \frac{Y_n}{c}\,, \quad \text{assuming that } c\neq 0. \tag{13.11}$$

PROOF: Omitted. See, e.g., [1] Bickel and Doksum: Mathematical Statistics.

Theorem 13.3 (Convergence is maintained under continuous transformations). ⋆

Let $Y_1, Y_2, \dots)$ and Y be random variables on some probability space (Ω, P). Let $f : \mathbb{R} \to \mathbb{R}$ be continuous. Then,

$$Y_n \xrightarrow{a.s.} Y \;\Rightarrow\; f\circ Y_n \xrightarrow{a.s.} f\circ Y\,.$$
$$Y_n \xrightarrow{P} Y \;\Rightarrow\; f\circ Y_n \xrightarrow{P} f\circ Y\,.$$
$$Y_n \xrightarrow{D} Y \;\Rightarrow\; f\circ Y_n \xrightarrow{D} f\circ Y\,.$$

PROOF: Omitted. [120] ■

Example 13.2 (Convergence in probability but not a.s.). ⋆

Consider the "sliding hump" example. [121] As our probability space we choose $\Omega := [0,1]$, the unit interval in $\mathbb{R}$, with the probability measure defined by $P\big(]a,b]\big) := b-a$.

[119] x_n converges to $x \;\Leftrightarrow$ for all $\varepsilon>0$ one can find $N\in\mathbb{N}$ such that $|x_n - x|<\varepsilon$ whenever $n\geq N$.

[120] A proof can be found at this Convergence of random variables (Mann–Wald theorem, general transformation theorem) Wikipedia link.

[121] See this StackExchange link.

(a) We partition Ω into the two intervals $I_1 = [0, 1/2]$ and $I_2 =]1/2, 1]$.

- For $n = 1, 2$, let $Y_n(\omega) := \begin{cases} 1\,, & \text{if } \omega \in I_n, \\ 0\,, & \text{else}\,. \end{cases}$

(b) We partition Ω into the three intervals $I_3 = [0, 1/3], I_4 =]1/3, 2/3]$, and $I_5 =]2/3, 1]$, then into $I_6 = [0, 1/4]$, $I_7 =]1/4, 2/4]$, $I_8 =]2/4, 3/4]$, and $I_9 =]3/4, 1]$, and so on .....

- We define random variables Y_n as in **(a)**: For $n \in \mathbb{N}$, let $Y_n(\omega) := \begin{cases} 1\,, & \text{if } \omega \in I_n, \\ 0\,, & \text{else}\,. \end{cases}$

(c) Then the sequence Y_n converges in probability to the (deterministic) random variable $\omega \mapsto Y(\omega) := 0$. A proof is given directly after this example.

(d) But this sequence of random variables does not converge almost surely. In fact, there is no $0 \le \omega \le 1$ for which $\lim\limits_{n\to\infty} Y_n(\omega)$ exist:

- Fix $\omega \in [0, 1]$. By construction, there are indices $n_1 = n_1(\omega) \;<\; n_2 = n_2(\omega) < n_3 = n_3(\omega) < \cdots$, such that $\omega \in I_{n_k}$ and I_{n_k} has length $1/k$. (Thus, $P(I_{n_k}) = 1/k$.)

(e) Let $\omega' \in [0, 1]; \omega' \neq \omega$. The subsequences $n_k(\omega)$ and $n_k(\omega')$ will differ for all k so large that $\dfrac{1}{k} < \dfrac{|\omega - \omega'|}{2}$, i.e., $\dfrac{2}{k} < |\omega - \omega'|$, since $\omega \in I_{n_k(\omega)}$ and $\omega' \in I_{n_k(\omega')} \Rightarrow I_{n_k(\omega)} \cap I_{n_k(\omega')} = \emptyset$. (Draw a picture!)

(f) It follows for such big k, that $Y_{n_k(\omega)}(\omega) = 1$ and $Y_{n_k(\omega)}(\omega') = 0$.
On the other hand, $Y_{n_k(\omega')}(\omega) = 0$ and $Y_{n_k(\omega')}(\omega') = 1$.
Thus, the full sequences $Y_n(\omega)$ does not have a limit, since it would have to be 1 along the subsequence $n_k(\omega)$ and 0 along the subsequence $n_k(\omega')$.

(g) ω is arbitrary in $\Omega = [0, 1]$. This shows that there is no $\omega \in \Omega$ for which $\lim\limits_{n\to\infty} Y_n(\omega)$ exists. $\square$

PROOF that (Y_n) converges in probability:
If we write $|I_n|$ for the length of the interval I_n, then

(h) $\boxdot\, |I_n| = 1 \;\Leftrightarrow\; n = 1 \;\; \boxdot\, |I_n| = 1/2 \;\Leftrightarrow\; n = 2, 3 \;\; \boxdot\, |I_n| = 1/3 \;\Leftrightarrow\; n = 4, 5, 6.$
Thus, if $s_1 = 1$, $s_2 = s_1 + 2$, $s_3 = s_2 + 3, \ldots, s_k = s_{k-1} + k = \sum\limits_{j=1}^{k} j = \dfrac{k \cdot (k+2)}{2}, \ldots,$

(i) then $I_n = 1/k \;\Leftrightarrow\; n = s_{k-1} + 1,\, s_{k-1} + 2, \ldots, s_{k-1} + k \;\Leftrightarrow\; s_{k-1} < n \le s_k$.

(j) It should be clear that $\big[\, n \to \infty \,\big] \big[\, k \to \infty \,\big]$ For a proof: $\boxdot$ "$\Leftarrow$" follows from $n \ge k$.
$\boxdot$For the other direction, we observe that $n \overset{\text{(i)}}{\le} 2s_k = 2k(k+1) < 2(k+1)^2$,
i.e., $\sqrt{n/2} - 1 \;<\; k$. Thus, $\big[\, n \to \infty \,\big] \;\Rightarrow\; \big[\, k \to \infty \,\big]$ and "$\Rightarrow$" follows.

(k) Since $Y_n(\omega) := \begin{cases} 1\,, & \text{if } \omega \in I_n, \\ 0\,, & \text{else} \end{cases}$ for $n \in \mathbb{N}$, we obtain $P\{|Y_n| \ge \varepsilon\} = 0$ for $\varepsilon \le 1$ and, with n_k as defined in **(k)**, $P\{|Y_{n_k}| \ge \varepsilon\} = \dfrac{1}{k}$ for $0 < \varepsilon \ge 1$. Thus, $P\{|Y_{n_k}| \ge \varepsilon\} \le \dfrac{1}{k}$ for $\varepsilon > 0$.

(l) Fix $\varepsilon > 0$ and $k \in \mathbb{N}$. $|I_n|$ and hence, $P\{|Y_n| > \varepsilon\}$ is nonincreasing with n. Thus, $n \ge n_k \;\Rightarrow\; P\{|Y_n| > \varepsilon\} \;\le\; P\{|Y_{n_k}| > \varepsilon\} \;=\; \dfrac{1}{k}$. Since $\big[\, n \to \infty \,\big] \overset{\text{(j)}}{\Rightarrow} \big[\, k \to \infty \,\big]$, it follows that $\lim\limits_{n\to\infty} P\{|Y_n| > \varepsilon\} = 0$ and this shows that $Y_n \overset{P}{\to} 0$. ■

□

13.2 Two Laws of Large Numbers

Our knowledge of convergence in probability and almost surely enables us to understand the weak law and the strong law of large numbers. Recall that the "id" part of any iid sequence (Y_n) implies that $E[Y_1] = E[Y_2] = \cdots$ and $Var[Y_1] = Var[Y_2] = \cdots$.

Theorem 13.4 (Weak Law of Large Numbers)**.**

Let $Y_1, Y_2, \ldots$ be an iid sequence of random variables on a probability space (Ω, P). with finite variance: $\sigma^2 := var[Y_n] < \infty$. Let $\mu := E[Y_n]$. Then,

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \text{ converges in probability to } \mu, \text{ i.e.,}$$

$$[\,\varepsilon > 0\,] \;\Rightarrow\; \left[\lim_{n\to\infty} P\left\{ \left| \frac{1}{n}\sum_{j=1}^{n} Y_j - \mu \right| > \varepsilon \right\} = 0. \right] \tag{13.12}$$

PROOF: Let

$$\omega \mapsto \bar{Y}_n(\omega) := \frac{Y_1(\omega) + Y_2(\omega) + \cdots + Y_n(\omega)}{n} = \frac{1}{n}\sum_{j=1}^{n} Y_j(\omega).$$

We have seen in Example 11.5 (Variance of the sample mean) on p.262, that

$$\mu_{\bar{Y}_n} = E[\bar{Y}_n] = \mu\,, \qquad \text{and} \qquad \sigma^2_{\bar{Y}_n} = Var[\bar{Y}_n] = \frac{\sigma^2}{n}\,. \tag{A}$$

We apply Tchebysheff's inequality 10.53 on p.238 with $k = \varepsilon\sqrt{n}/\sigma$ and obtain from **(A)**, that

$$P\left\{|\bar{Y}_n - \mu| > \varepsilon\right\} \le \frac{1}{(n\,\varepsilon^2/\sigma)^2} = \frac{\sigma^2}{n\,\varepsilon^2} \to 0, \text{ as } n \to \infty$$

This proves that $P\text{–}\lim_{n\to\infty} \bar{Y}_n = \mu$. ■

Remark 13.1. We have previously encountered the random variable $\bar{Y}_n$ under the name $\bar{Y}$, as the sample mean of a sample of size n. See Example 11.5 (Variance of the sample mean) on p.262.

It is considered bad form to use a subscript for the sample mean. We chose to do so in this section about the laws of large numbers anyway, since we are not dealing with this sample mean in the context of samples of a fixed size, but we are examining what happens as this size approaches infinity. □

Remark 13.2. We have learned in Theorem 13.1 (Relationship between the modes of convergence) on p.317, that almost sure convergence implies convergence in probability. One can interpret this in the following manner:

- It is harder to establish almost sure convergence, since it is a more powerful tool for proving that some mathematical property is true.
- Accordingly, it would be wonderful if one could strengthen a theorem that proves convergence in probability for some sequence of random variables, to show that this convergence actually happens almost surely.
- It turns out that this is possible for the weak law of large numbers (Theorem 13.4 on p.320. It is called the **weak** law of large numbers because there also is a **strong** law of large numbers which replaces the conclusion $P\text{–}\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j = \mu$ with a.s.–$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j = \mu$. We will study that next. □

Theorem 13.5 (Strong Law of Large Numbers)**.**

Let $Y_1, Y_2, \ldots$ be an iid sequence of random variables on a probability space (Ω, P).
Let $\mu := E[Y_n]$. Then,

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \text{ converges almost surely to } \mu, \text{ i.e.,}$$

$$P\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j \neq \mu\right\} = 0\,. \tag{13.13}$$

PROOF:

Outside the scope of these lecture notes. ■

Example 13.3 (Infinite Monkey Theorem)**.** A monkey has been granted eternal life. It is continually hitting at random the keys of a wordprocessor that will never break down.

The keyboard has a customized layout that makes it equally likely for each key, at any given key stroke, to be selected by the monkey. (For example, there is no CAPS key. Rather, there are separate keys for "a" and "A", "b" and "B", ...)

What is the probability that, in this infinite sequence of letters, there is a contiguous block that constitutes the collected work of William Shakespeare? We expect a flawless result: No typos, correct punctuation, CAPS exactly when required, ...

Solution:

- There are K different keys that are being hit, at each stroke, with equal probability.
- Only one of them is correct at any given time and the others are failures.
- Thus, the sequence $X_1, X_2, \ldots$ of key strokes is a sequence of independent random items with constant success probability $p_j = p = 1/K$.
- We consider the indices $1, 2, 3, \ldots$ as points in time, so X_{753} is the key that was hit at time $j = 753$.
- The author does not know how many letters Shakespeares collected work ("S-C-W") consists of, but this certainly is a finite number. Let us denote it by N.

Let $Y_1 := 1$, if $X_1, X_2, \ldots, X_N$ form S-C-W. Let $Y_1 := 0$, else.
Let $Y_2 := 1$, if $X_{N+1}, X_{N+2}, \ldots, X_{2N}$ form S-C-W. Let $Y_2 := 0$, else.
- -
Let $Y_j := 1$, if $X_{(j-1)N+1}, X_{(j-1)N+2}, \ldots, X_{jN}$ form S-C-W. Let $Y_j := 0$, else.
- -

- If $i \neq j$, then Y_i and Y_j depend on "disjoint" chunks $\big(X_{(i-1)N+1}, X_{(i-1)N+2}, \ldots, X_{iN}\big)$ and $\big(X_{(j-1)N+1}, X_{(j-1)N+2}, \ldots, X_{jN}\big)$ of the independent X_k. Thus, Y_i and Y_j are independent.
- Also, both are binom$(1, (1/K)^N)$ (Bernoulli trials).
- Thus, $(Y_n)_n$ is an iid sequence with expectations $\mu = (1/K)^N$.
- By the strong law of large numbers, there is an event $A \subseteq \Omega$ such that $P(A) = 1$ and

$$\omega \in A \;\Rightarrow\; \lim_{n\to\infty} \sum_{j=1}^{n} Y_j(\omega) \Big/ n \;=\; \mu \;=\; \left(\frac{1}{K}\right)^N \;>\; 0.$$

- Since we divide the sum by n, the limit is zero if only finitely many $Y_j(\omega)$ are not zero, i.e., if only finitely many $Y_j(\omega)$ are 1. Thus,

$$\omega \in A \;\Rightarrow\; Y_j(\omega) \;=\; 1\,, \quad \text{infinitely often!}$$

- Since $P(A) = 1$ and Y_j denotes the completion of the jth collection of Shakespeare's works:
- With probability 1, the monkey will produce an infinite number of Shakespeare's entire collection! □

13.3 Sampling Distributions

Introduction 13.2. Back in Chapter 8.2 (Sampling and Urn Models With and Without Replacement), we gave Definition 8.2 (Sampling as a Random element) on p.185 of a sample.

- A sample of size n was nothing but a vector $\vec{X} = \big(X_1, X_2, \ldots, X_n\big)$ of random elements, or a realization $\vec{x} = \vec{X}(\omega)$ of those random elements. What makes this vector a sample is the interpretation of $\omega \mapsto X_j(\omega)$ as the jth pick of an item from a population of interest and the intent to use the outcomes $x_j = X_j(\omega)$ for inferences about that population.

These sample picks may happen with or without replacement. Sampling with replacement is desirable from a mathematical point of view, since this allows us to assume that the sample picks have identical distribution. Thus, if the X_j are real–valued, their cumulative distribution functions satisfy

$$F_{X_1}(x) \;=\; F_{X_2}(x) \;= \cdots = \; F_{X_n}(x) \;\; (x \in \mathbb{R})\,;$$

This in turn implies that, if the sample picks are real-valued functions of ω i.e., they are random variables, they all have the same expectation, variance, MGF, and so on.

Also, those sample picks may or may not be independent. independence would be extremely desirable from a mathematical perspective. For example, if the X_j are jointly continuous and independent random variables, knowledge of the marginal densities yields the joint density, because,

$$f_{\vec{X}}(\vec{x}) \;=\; f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n) \;\; (\vec{x} \in \mathbb{R}^n)\,.$$

Unfortunately, identical distribution and independence are simplifications of the real world. This is even true when one considers n rolls of a die. [122] The surface on which the die is rolled is not perfectly even, so that negates identical distribution. If several people take turns, then the different ways in which they throw the die creates a dependency. Of course, it is very likely that those differences, if we are able to detect them, are so minuscule that they can be ignored.

But there are many examples where those deviations are so large that we cannot work under the iid assumption. This need not necessarily occur in a real world application. It can also be part of the probabilistic models we create: Whenever we assume that we sample without replacement from a finite population, the probabilistic makeup of the items remaining in that population changes with every item we happen to pick for our sample.

Consider sampling at random from an urn that initially contains R red and $N - R$ black balls. If X_j is red, then there will be less of a probability of X_{j+1} being red, than if X_j was black. Hence, the X_j are neither independent, nor identically distributed.

However, those sample picks constitute a simple random sample according to Definition 8.3 (Simple Random Sample) on p.186:

- A sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ of size n from a population of size $N \geq n$ is called a simple random sample (SRS), if it is done without replacement and if each one of the potential outcomes $\vec{x} = \vec{X}(\omega)$ has equal chance of being selected.

If the sample size of an SRS is large, but small when compared to the size of the population, then treating it as iid will result in insignificant domputational differences. [123] This observation is one of the reasons that even the more restrictive definition of an SRS is of a generality we are not looking for in this chapter. We follow [5] Hogg, McKean, Craig: Introduction to Mathematical Statistics.

A typical statistical problem can be described as follows: We have a random variable Y that we know about, but we do not know its distribution, given by its CDF $F_Y(y)$.

Our insufficient knowledge of Y can manifest itself in two different ways:

(I) We know the type of distribution, but not all of its parameters. For example, we may know that Y is normal with $\sigma^2 = 3.65$, but its mean μ is unknown.

(II) We do not even know the type of distribution: Does Y follow a Poisson distribution or is it normal or exponential or .....?

We deal in this section with problem **(I)**. □

Example 13.4. Some more problem **(I)** examples are the following:

(a) $Y \sim$ binom$(64, p)$, with unknown success probability p. We write $p_Y(y; p)$ for the PMF.

(b) $Y \sim \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We write $f_Y(y; \mu, \sigma)$ for the PDF.

(c) $Y \sim$ expon(β), with unknown β. We write $f_Y(y; \beta)$ for the PDF.

(d) $Y \sim$ gamma$(\alpha, 3)$, with unknown α. We write $f_Y(y; \alpha)$ for the PDF. □

Remark 13.3. The examples just given suggest now to handle the general case. Since the random variable Y is given and we know its distribution except for one or several parameters, we know its PMF $p_Y(y)$ in the discrete case or PDF $f_Y(y)$ in the continuous case. It is customary to write θ or

[122] Interpret X_j as the jth pick from the population of all rolls of that die.

[123] We mentioned this in Remark 8.2 on p.185.

$\vec{\theta}$ for the unknown parameter or **parameters of the distribution** and to write Θ for the **parameter space**, i.e., the set of all parameters we consider for the problem. [124]

Thus, in Example 13.4**(a)**, $\Theta = [0, 1]$. In Example 13.4**(b)**, $\Theta =]-\infty, \infty[\times[0, \infty[$.

Problem **(I)** can now be formulated as follows:

- Given is a random variable Y of which we know its distribution except for one or several parameters.
  ⊡ We know the PMF $p_Y(y;\theta)$ if Y is discrete. ⊡ We know the PDF $f_Y(y;\theta)$ if Y is continuous.
- What is a good, possibly optimal, procedure for the estimation of θ from the sample that we have drawn or intend to draw from the population?

It seems obvious enough that this estimate must be a (deterministic) function

$$\theta \;=\; T(\vec{y}) \;=\; T(y_1, \ldots, y_n) \;=\; T\big(\vec{Y}(\omega)\big) \;=\; T\big(Y_1(\omega), \ldots, Y_n(\omega)\big)\,.$$

of the potential outcomes (realizations) of the sample. □

We had stated in the introduction that only iid samples are considered in this section.

Definition 13.2 (Random samples from a distribution)**.**

Let Y be a random variable on a probability space (Ω, P). Let $n \in \mathbb{N}$. We call a vector $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$ a **random sampling action of size** n **on** (or **from) the distribution of** Y, if

- the random variables $Y_1, \ldots, Y_n$ are iid with distribution P_Y.

The following are alternate names for this kind of sampling action:

- **random sampling action of size** n **on** (or **from)** Y
- "random sampling action" can be shortened to "**random sample**"
- **random sample** also refers to a realization $\vec{y} = \vec{Y}(\omega)$ of a random sampling action.

Note that the last two bulleted items are consistent with earlier definitions of sampling where we also use "sample" both for a sampling action and a realization of such an action. □

That definition allows us to restate the essence of Remark 13.3 as follows: We expect a procedure to estimate the parameter θ of a PMF $p_Y(y;\theta)$ or PDF $f_Y(y;\theta)$ to be a random variable $\omega \mapsto T\big(\vec{Y}(\omega)\big)$.

There is a special name for transforms $\vec{y} \mapsto T(\vec{y})$ of a random sample on Y.

Definition 13.3 (Statistic)**.**

[124] It is unfortunate that this standard notation for parameters to be estimated is at odds with the other standard which uses the CAPS version of a letter to denote a random item and the corresponding small letter to denote an outcome of this random element. (For example, $y = Y(\omega)$).

Let Y be a random variable on a probability space (Ω, P) and $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$ a random sampling action on Y. Let

$$T : \mathbb{R}^n \mapsto \mathbb{R}\,; \qquad \vec{y} \mapsto T(\vec{y})$$

be some function that can be applied to the sampling action $\vec{Y}$. We call the random variable

$$\omega \mapsto T\big(\vec{Y}(\omega)\big)$$

a **statistic** of that sampling action. We call the distribution of that random variable,

$$B \;\mapsto\; P_{T\circ\vec{Y}}(B) \;=\; P\{T(\vec{Y}) \in (B)\} \;=\; P\{\omega \in \Omega : T\big(\vec{Y}(\omega)\big) \in B\} \tag{13.14}$$

its **sampling distribution**. Once the sampling action has been performed and the corresponding realization $\vec{y} = Y(\vec{\omega})$ has been obtained, we call $t \;=\; T\big(\vec{Y}(\omega)\big)$ the realization of the statistic. □

Theorem 13.6.

Let Y be a random variable on a probability space (Ω, P) and $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$ a random sampling action on Y. Let $T_1, T_2, \ldots, T_k : \mathbb{R}^n \mapsto \mathbb{R}$ be statistics for that sample action. Let

$$T^* : \mathbb{R}^k \mapsto \mathbb{R}\,; \qquad (t_1, \ldots, t_k) \mapsto T^*(t_1, \ldots, t_k)\,.$$

Then, setting $\vec{t} = (t_1, \ldots, t_k)$ and $\vec{T} = (T_1, \ldots, T_k)$, the composition

$$T^* \circ \vec{T} \circ \vec{Y} \;:\; \omega \mapsto T^*\big(\vec{T}[\,\vec{Y}(\omega)\,]\big) \;=\; T^*\big(T_1[\,\vec{Y}(\omega)\,], \ldots, T_k[\,\vec{Y}(\omega)\,]\big)$$

also is a statistic of $\vec{Y}$.

PROOF:

Left as an exercise which is very easy for the student who has had exposure to functions $\mathbb{R}^n \to \mathbb{R}^k$ with dimensions n and/or k that can exceed the value 3. ■

The last theorem can be stated succinctly and without mathematical symbols as follows:

A function of a function of the data is a function of the data.

Here is an example of a statistic which is so important that it deserves its own definition. It also is used to illustrate Theorem 13.6.

Definition 13.4 (Sample variance)**.**

Let $\vec{Y} = \big(Y_1, \ldots, Y_n\big)$ be a random sample action on a random variable Y.
The **sample variance** is defined as the random variable

$$\omega \mapsto S^2(\omega) \;:=\; \frac{1}{n-1}\sum_{j=1}^{n}\big(Y_j(\omega) - \bar{Y}(\omega)\big)^2\,. \tag{13.15}$$

We further call $\omega \mapsto S(\omega) := \sqrt{S^2(\omega)}$ the The **sample standard deviation**.
We will often write s^2 and s for the realizations $S^2(\omega)$ and $S(\omega)$ that result from creating the sample.

We write S_n, S_n^2, s_n, s_n^2 for S, S^2, s, s^2, if we want to keep track of the sample size. That will be the case, e.g., if we consider the sample variance of the first n picks of a sample of infinite size. □

Example 13.5. For the following examples assume that $\vec{Y} = (Y_1, \ldots, Y_n)$ is a random sample on a random variable Y.

(a) In Example 11.5 (Variance of the sample mean) on p.262, we considered the sample mean $\omega \mapsto \bar{Y}(\omega) = \frac{1}{n}\sum_{j=1}^{n} Y_j(\omega)$. $\bar{Y}$ is a statistic: The transform is $T(\vec{Y}) = \frac{1}{n}\sum_{j=1}^{n} Y_j$.
We also mentioned that this statistic is an obvious choice for estimating the parameter $\mu = E[Y]$ of the underlying random variable Y.

(b) Sample variance S^2 and sample standard deviation S which were defined above are statistics. This can be shown with the help of Theorem 13.6 on p.325 as follows. Let

$$t_1 = T_1(\vec{y}) = y_1,\ t_2 = T_2(\vec{y}) = y_2, \ldots, t_n = T_n(\vec{y}) = y_n,\ t_{n+1} = T_{n+1}(\vec{y}) = \bar{y}\,.$$
$$T^*(t_1, \ldots, t_n, t_{n+1}) = \frac{1}{n-1}\sum_{j=1}^{n}\left(t_j - t_{n+1}\right)^2$$

Then $S^2 = T^*\big(T_1(\vec{Y}), \ldots, T_n(\vec{Y})\big), T_{n+1}(\vec{Y})\big)$. By Theorem 13.6, S^2 is a statistic for $\vec{Y}$. We apply this theorem again to the function $T^{**} : t^* \mapsto \sqrt{t^*}$ and obtain that the standard deviation S is a statistic, since $S = T^{**}(S^2)$.

(c) The jth order statistic, $Y_{(j)}$ is indeed a statistic, since knowledge of all values of a list $y_1, \ldots, y_n$ of real numbers uniquely determines which one is the jth largest value in that list.

(d) The **sample range**, $R = Y_{(n)} - Y_{(1)}$, is a statistic, since it is a function (the difference) of the two statistics $Y_{(n)}$ and $Y_{(1)}$. □

Example 13.6 (WMS Ch.07.1, Example 7.1)**.** Example 7.1 of the WMS text discusses in quite big detail the sampling distribution of the statistic $\bar{Y}$ for a sample of three independent rolls of a balanced die. You are strongly encouraged to study it. □

Theorem 13.7 (WMS Ch.07.2, Theorem 7.1)**.**

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e., we sample on a random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then the sample mean $\bar{Y}$ follows a normal distribution with mean μ and variance σ^2/n.

PROOF: That is an immediate consequence of Theorem 12.5 (Linear combinations of uncorrelated normal variables are normal) on p.309. ■

Theorem 13.8 (WMS Ch.07.2, Theorem 7.2)**.**

Let $\vec{Y} = (Y_1, \ldots, Y_n)$ be a random sample on $Y \sim \mathcal{N}(\mu, \sigma^2)$. Let $Z_j = (Y_j - \mu)/\sigma$ for $j = 1, 2, \ldots, n$. Then $\vec{Z} = (Z_1, \ldots, Z_n)$ is a random sample on a standard normal variable. (In particular, the Z_j are iid.) Further,

$$\sum_{j=1}^{n} Z_i^2 = \sum_{j=1}^{n} \left(\frac{Y_j - \mu}{\sigma} \right)^2 \tag{13.16}$$

follows a χ^2 distribution with n degrees of freedom.

PROOF: It follows from Theorem 12.5 (Linear combinations of uncorrelated normal variables are normal) on p.309 that the linear combination $Z_j = (Y_j - \mu/\sigma)$ is standard normal. It follows from Theorem 12.4 (MGF of a sum of functions of independent variables) on p.309 that the Z_j are iid. It follows from Theorem 12.6 on p.312 that $\sum_{j=1}^{n} Z_i^2 \sim \chi^2(\text{df} = n)$. ■

The following is Example Example 6.13 of the WMS text.

Proposition 13.1. ⋆

Let Y_1 and Y_2 be independent standard normal random variables. Then $Y_1 + Y_2$ and $Y_1 - Y_2$ are independent and normally distributed, both with mean 0 and variance 2.

PROOF: See WMS Ch.06.6, Example 6.13. ■

Theorem 13.9 (Independence of sample mean and sample variance in normal populations)**.**

Let $\vec{Y} = (Y_1, \ldots, Y_n)$ be a random sample on $Y \sim \mathcal{N}(\mu, \sigma^2)$. Let $Z_j = (Y_j - \mu)/\sigma$ for $j = 1, \ldots, n$. Then, $\vec{Z} = (Z_1, \ldots, Z_n)$ is a random sample on a standard normal variable. Moreover,

(a) $$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \sim \chi^2(df = n-1)$$

(b) *$\bar{Y}$ and S^2 are independent random variables.*

PROOF: ⋆ See the proof of WMS Ch.07.2, Theorem 7.3 for the case $n = 2$. ■

- The sample mean $\bar{Y}$ was a natural choice to estimate the mean $\mu = E[Y]$ of a random variable X.
- It seems just as natural to use the sample variance S^2 to estimate $\sigma^2 = Var[Y]$. We will see that, if Y follows a normal distribution, this choice turns out to be mathematically sound.

The t distribution which we define next is a means towards that end.

Definition 13.5 (Student's t–distribution [125])**.**

[125] Named after the English statistician William S. Gosset (1876 – 1937). Georg Ferdinand Ludwig Philipp Cantor (1845 – 1918), Gosset was Head Brewer of the Guinness Brewery in Dublin, Ireland and published his papers under the pseudonym "Student".

Let Z and W be independent random variables such that Z is standard normal and$W$ is χ^2 with ν df. Let

$$T = \frac{Z}{\sqrt{W/\nu}} \tag{13.17}$$

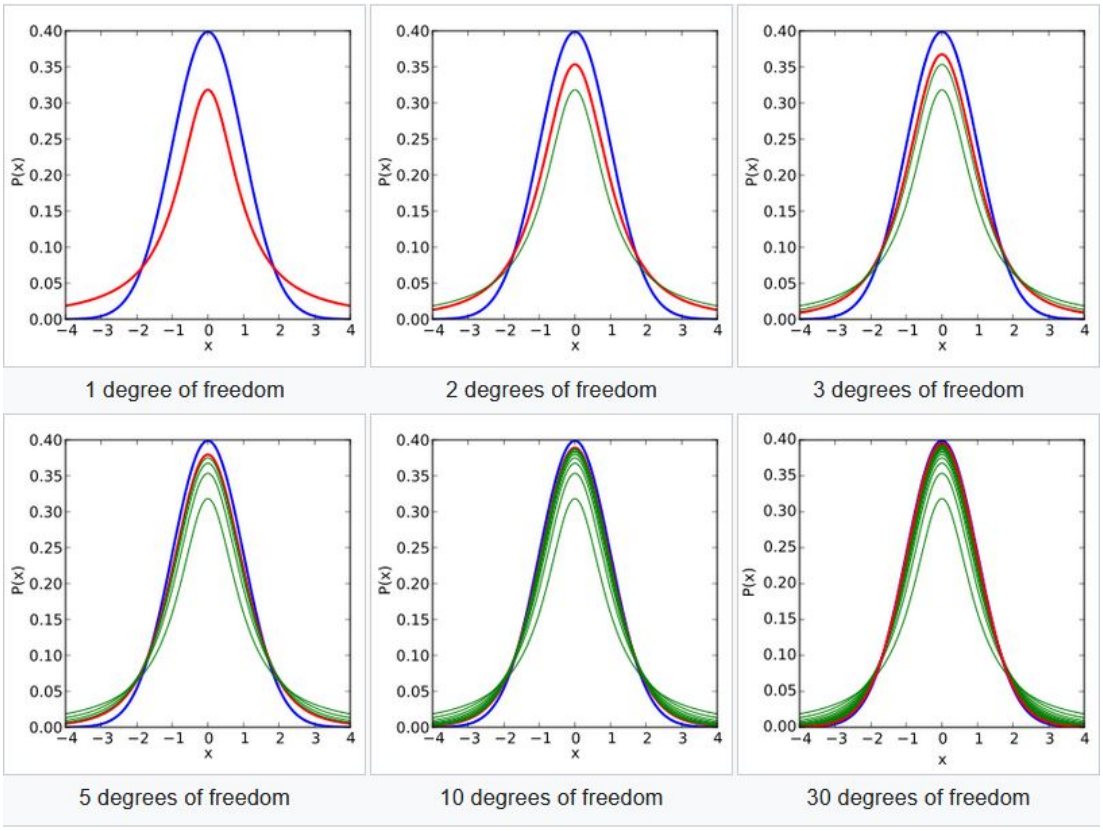
Then we refer to the distribution P_T of T as a **t–distribution** or **Student's t–distribution** with ν df. We also write that as $T \sim t(\nu)$ or $T \sim t(\text{df } = \nu)$. □

Remark 13.4.

- One can prove that $E[T] = 0$ for any ν, and $Var[T] = \frac{\nu}{\nu - 2}$ for $\nu > 2$.

The density of the t–distribution looks very similar to that of a normal density. Both have a symmetrical, bell shaped graph. But note the following:

- Since it does not depend on ν, $E[T] = 0$ is not a parameter of the t–distribution.
- Since $\frac{\nu}{\nu - 2} > 1$, the tails are fatter than those of a $\mathcal{N}(0, 1)$ variable. See Figure 13.1. □



13.1 (Figure). **densities of the standard normal and t distribution.** Source: Wikipedia.

Remark 13.5. The following looks somewhat strange. Assume that Z_1 and Z_2 are independent and standard normal. Since $Z_2^2 \sim \chi^2(\text{df} = 1)$ and $|Z_2| = \sqrt{Z_2^2}$, the random variable $Z_1/|Z_2|$ follows a $t(\text{df} = 1)$ distribution! □

Theorem 13.10.

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $\vec{Y} = (Y_1, \ldots, Y_n)$ be a random sample on Y. Let

$$T := \frac{\bar{Y} - \mu}{S/\sqrt{n}}. \tag{13.18}$$

Then T follows a t–distribution with $(n-1)$ df.

PROOF: Let

$$\textbf{(A)} \qquad Z := \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad W := \frac{(n-1)S^2}{\sigma^2}.$$

We have seen that $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2(\text{df } = n-1)$. Since $\bar{Y}$ and S^2 are independent by Theorem 13.9 on p.327, Z as a function of $\bar{Y}$ only and W as a function of S^2 only also are independent. Now,

$$\begin{aligned} T &\overset{(13.18)}{=} \frac{\bar{Y} - \mu}{S/\sqrt{n}} = \frac{(\bar{Y} - \mu)/(\sigma/\sqrt{n})}{S/\sqrt{n}/(\sigma/\sqrt{n})} \overset{\textbf{(A)}}{=} \frac{Z}{S/\sigma} \\ &= \frac{Z}{(\sqrt{n-1}/\sqrt{n-1}) \cdot (\sqrt{S^2}/\sqrt{\sigma^2})} \\ &= \frac{Z}{\sqrt{\left[(n-1)S^2\right]/\sigma^2}\,/\,\sqrt{n-1}} \overset{\textbf{(A)}}{=} \frac{Z}{\sqrt{W}/\sqrt{n-1}} = \frac{Z}{\sqrt{W/(n-1)}}. \end{aligned}$$

By definition, of the t–distribution, $\dfrac{Z}{\sqrt{W/(n-1)}} \sim t(\text{df } = n-1)$. ■

Example 13.7 (WMS Ch.07.2, Example 7.6)**.** Example 7.6 of the WMS text discusses a practical example of the Student's t–distribution that discusses how to estimate the unknown variance of a normal random variable from a sample. You are strongly encouraged to study it. □

The next and last distribution tied to random sampling on a normal variable that we give in this section allows us to compare the variances of two random samples on normal random variables that represent two independent populations. This is used in the so called analysis of variance (ANOVA) to decide whether the means of several independent normal populations all coincide or whether at least two of them are different.

Definition 13.6 (F–distribution)**.**

Given are two independent random variables $W_1 \sim \chi^2(\text{df } = \nu_1)$ and $W_2 \sim \chi^2(\text{df } = \nu_2)$. with ν_1 and ν_2 df, respectively. Then we say that

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

follows an **F distribution** with ν_1 **numerator degrees of freedom** and ν_2 **denominator degrees of freedom**. □

Remark 13.6. ⋆ One can show that

- $\nu_2 > 2 \;\Rightarrow\; E[F] \;=\; \dfrac{\nu_2}{\nu_2 - 2}$,
- $\nu_2 > 4 \;\Rightarrow\; Var[F] \;=\; \dfrac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$. □

Theorem 13.11.

Consider two random samples of sizes n_1 and n_2 from two independent populations, on random variables $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with sample variances S_1^2 and S_2^2. Let

$$F \;:=\; \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\,. \tag{13.19}$$

Then F follows an F distribution with $(n_1 - 1)$ numerator df and $(n_2 - 1)$ denominator df.

PROOF: Let

$$W_1 \;:=\; \frac{(n_1 - 1)S_1^2}{\sigma_1^2}\,, \quad W_2 \;:=\; \frac{(n_2 - 1)S_2^2}{\sigma_2^2}\,.$$

Since the random samples are independent, so are their sample variances S_1^2 and S_2^2, and so are the transforms W_1 of S_1^2 and W_2 of S_2^2. By Theorem 13.9 (Independence of sample mean and sample variance in normal populations) on p.327,

$$W_1 \;\sim\; \chi^2(\text{df} = n_1 - 1)\,, \qquad \text{and} \qquad W_2 \;\sim\; \chi^2(\text{df} = n_2 - 1)\,.$$

According to Definition 13.6 of an F distribution,

$$\frac{W_1/\nu_1}{W_2/\nu_2} \;=\; \frac{\left[(n_1 - 1)S_1^2/\sigma_1^2\right]/(n_1 - 1)}{\left[(n_2 - 1)S_2^2/\sigma_2^2\right]/(n_2 - 1)} \;=\; \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows an F distribution with $(n_1 - 1)$ numerator df and $(n_2 - 1)$ denominator df. ■

Example 13.8 (WMS Ch.07.2, Example 7.7)**.** Example 7.6 of the WMS text discusses another practical example of the Student's F distribution. You are strongly encouraged to study it. □

13.4 The Central Limit Theorem

Introduction 13.3. In section13.3 (Sampling Distributions) we were able to determine the sampling distributions of some very important statistics that can be computed from the realization of a random sample $\vec{Y}$ on some random variable Y. But there was very restrictive assumption on that underlying random variable

- Y had to follow a normal distribution.

We will find a solution for determining the sampling distribution of the sample mean, $\bar{Y} \;=\; \frac{1}{n}\sum_{j=1}^{n} Y_j$, even if Y is not normal.

- It is an **asymptotic solution**, i.e., its comes in form of a $U = \lim_{n\to\infty} U_n$ theorem.
- Here, U_n is a statistic $T_n \circ \vec{Y}$, which we can compute from (the realization of) $\vec{Y}$ and $\bar{Y}_n := \frac{1}{n}\sum_{j=1}^{n} Y_j$, a very natural approximation of $\bar{Y}$, can also be computed from U_n
- n denotes the sample size. Thus, the sample must be sufficiently large to allow us to ignore the discrepancy between U_n and U.

We have learned that there are four different kinds of limits which occur in connection with a sequence of random variables. We will discuss in this chapter the central limit theorem. It allows us to show the existence of the least desirable of those four limits, the limit in distribution. But that is not as bad as it sounds for the following reason.

- For large enough n, the CDF of U_n is close to that of U. Since the CDF determines the probabilities of all important events B, we can approximate $P\{U_n \in B\} \approx P\{U \in B\}$, □

We will state and prove the limit theorem which was mentioned in the introduction above, after the following important theorem that relates convergence in distribution, $Y_n \xrightarrow{\mathbf{D}} Y$, to (pointwise) convergence, $m_{Y_n}(t) \to m_Y(t)$ of the associated MGFs.

Theorem 13.12 (Lévy–Cramér continuity theorem). ⋆

Let $Y_1, Y_2, \dots$) be a sequence of random variables (iid is not assumed) with associated CDFs $F_{Y_1}, F_{Y_2}, \dots$) and MGFs $m_{Y_1}(t), m_{Y_2}(t), \dots$).
Let Y be a random variable with associated CDF F_Y and MGF $m_Y(t)$. Then,

$$\begin{aligned} &\big[\, m_{Y_n}(t) \to m_Y(t) \text{ as } n \to \infty, \quad \text{for all } t \in \mathbb{R} \,\big] \\ \Rightarrow\ &\big[\, F_{Y_n}(y) \to F_Y(y) \text{ as } n \to \infty, \quad \text{for all } y \text{ where } F_Y(\cdot) \text{ is continuous.} \,\big] \end{aligned} \tag{13.20}$$

PROOF: Outside the scope of this course. ■

Theorem 13.13 (Central Limit Theorem).

Central Limit Theorem:
Let $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ be a vector of iid random variables with common expectation $E[Y_j] = \mu$ and finite variance $Var[Y_j] = \sigma^2$. Let Z be a standard normal variable and

$$U_n := \frac{\sum_{j=1}^{n} Y_j - n\mu}{\sigma \cdot \sqrt{n}} = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}, \qquad \text{where } n \in \mathbb{N},\ \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i .$$

Then, U_n converges to Z in distribution as $n \to \infty$. In other words,

$$\lim_{n\to\infty} P\{U_n \le u\} = P\{Z \le u\} = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad \text{for all } u .$$

PROOF:

(1) Let $\widetilde{Y}_n := Y_n - \mu$. The $\widetilde{Y}_n$ are iid, with $E[\widetilde{Y}_j] = 0$, $Var[\widetilde{Y}_j] = \sigma^2$ and MGF $m(t) := m_{\widetilde{Y}_n}(t)$. By Corollary 12.1 on p.309, $m_{\widetilde{Y}_1+\dots\widetilde{Y}_n}(t) = \big[m(t)\big]^n$. Thus.

(2) $$m_{U_n}(t) \;= E\left[\exp\left\{\sum_{j=1}^{n}\widetilde{Y}_j\cdot\frac{t}{\sigma\sqrt{n}}\right\}\right] \;= m_{\widetilde{Y}_1+\cdots\widetilde{Y}_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \;= \left[m\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n .$$

(3) According to Theorem 13.12 (Lévy–Cramér continuity theorem), it suffices to show that $\lim\limits_{n\to\infty} m_{U_n}(t) \;=\; m_Z(t) \;=\; e^{t^2/2}$.
Equivalently, since $x\mapsto e^x$ is continuous, it suffices to show that

(4) $$\lim_{n\to\infty}\ln m_{U_n}(t) \;= \frac{t^2}{2}\,.$$

(5) Let $h := \dfrac{t}{\sigma\sqrt{n}}$. Then $n = \dfrac{t^2}{\sigma^2 h^2}$. Thus, by **(2)**,

$$\ln m_{U_n}(t) \;=\; n\,\ln m(h) \;=\; \frac{t^2}{\sigma^2 h^2}\,\ln m(h) \;=\; \frac{t^2}{\sigma^2}\left(\frac{\ln m(h)}{h^2}\right).$$

Thus,

(6) $$\lim_{n\to\infty}\ln m_{U_n}(t) \;=\; \frac{t^2}{\sigma^2}\,\lim_{h\to 0}\frac{\ln m(h)}{h^2}\,.$$

Since $m(0)=e^0=1$, the right–hand limit is of the form $0/0$. We use L'Hôpital's rule [126] twice in a row and obtain, since $m(t)=m_{\widetilde{Y}_n}(t)$ and hence, $m''(0)=E[\widetilde{Y}_n^2]$,

$$\begin{aligned}\lim_{h\to 0}\frac{\ln m(h)}{h^2} \;&=\; \lim_{h\to 0}\frac{\big[1/m(h)\big]\,m'(h)}{2h} \;=\; \lim_{h\to 0}\frac{m'(h)}{2hm(h)}\\ &=\; \lim_{h\to 0}\frac{m''(h)}{2m(h)+2hm'(h)} \;=\; \frac{m''(0)}{2m(0)+0} \;=\; \frac{m''_{\widetilde{Y}_n}(0)}{2} \;=\; \frac{E[\widetilde{Y}_n^2]}{2}\,.\end{aligned} \tag{7}$$

(8) Since $\widetilde{Y}_n = Y_n-\mu$ and hence, $E[\widetilde{Y}_n^2]=E[(Y_n-\mu)^2]=Var[Y_n]=\sigma^2$, **(7)** implies

$$\lim_{h\to 0}\frac{\ln m(h)}{h^2} \;=\; \frac{\sigma^2}{2}\,.$$

Thus, by **(6)**, $\lim\limits_{n\to\infty}\ln m_{U_n}(t) \;=\; \dfrac{t^2}{\sigma^2}\cdot\dfrac{\sigma^2}{2} \;=\; \dfrac{t^2}{2}$.

We have shown **(4)** and this finishes the proof. ■

Remark 13.7. In statistical applications the CLT often is employed as follows: Carefully designed statistical techniques have resulted in the estimate $\mu=\mu_0$ for μ, the unknown mean of the population of interest. But this has been quite some time ago. Today there is reason to believe that this value is now outdated and one wants to obtain supporting evidence for that claim.

- We make $\mu=\mu_0$ our working hypothesis.

[126] in the form $\lim\limits_{x\to 0}\dfrac{f(x)}{g(x)} \;=\; \lim\limits_{x\to 0}\dfrac{f'(x)}{g'(x)}$

- An SRS $\vec{Y}$ of size n is taken and $c_0 := \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$ is computed from the sample mean realization $\bar{y} = \sum_{j=1}^{n} y_j$ which one obtains from the realization $\vec{y} = \vec{Y}(\omega)$ of the sample.
- If $\bar{Y}(\omega)$ is close to μ_0, then $P\left\{ \left| \dfrac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \right| > c_0 \right\}$ will be very small.

For example, assume that $c_0 = 3$, i.e., $|\bar{y} - \mu_0| = 3 \cdot (\sigma)/\sqrt{n})$. The r.v. $\omega \to \bar{Y}(\omega)$ satisfies

$$E[\bar{Y}] = E[Y] = \mu = \mu_0 \quad \text{and} \quad Var[\bar{Y}] = \frac{Var[Y]}{n} = \frac{\sigma^2}{n}, \quad \text{i.e.,} \quad \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{Y}}\,.$$

Thus, $c_0 = 3$ signifies that $\bar{Y}$ is three SDs away from its mean. According to the CLT, $\dfrac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$ is approximately standard normal and we can employ the the 68%–95%–99.7% rule for the normal distribution (the empirical rule). It tells us that the probability of $\dfrac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$ being within the ± 3 SD range is close to a huge 99.7%. But then we obtain a very small

$$P\left\{ \left| \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \right| > c_0 \right\} \approx 1 - 0.997 = 0.003\,.$$

In other words, the probability that a $\bar{Y}$ belonging to a random sample like ours (with the same sample size) is 3 SDs or more away from μ_0.

- So it was just the luck of the draw that let us obtain a sample that only has a chance of one in 333 of being picked. **Or is there another explanation?**

How about this? $\alpha_0 = 0.05$ was obtained contingent on the hypothesis that μ still equals μ_0. Let us change our point of view and assume that there was nothing unusual about our sample.

- We reject the hypothesis $\mu = \mu_0$, since the data obtained from the sample suggest that $|\bar{Y} - \mu| < |\bar{Y} - \mu_0|$ and that necessitates $\mu \neq \mu_0$.
- In the extreme, we could dispense with any effort to find a well founded estimate of μ. Instead, we act as if our particular sample serves that purpose and replace μ_0 with $\mu_1 := \bar{y}$.

In the extreme, we could dispense with any effort to find a well founded estimate of μ. Instead, we act as if our particular sample serves that purpose and replace μ_0 with $\mu_1 := \bar{y}$. But of course, that generally is not a good idea and one should follow the established process to obtain a new estimate of μ. □

Remark 13.8. This is a continuation of the previous example.

- The procedure outlined there to decide whether or not to reject the hypothesis $\mu = \mu_0$ involved the computation of the expression $c_0 := \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$.
- However, knowledge of the population variance $\sigma^2 = Var[Y_j]$ of a sample pick Y_j from that population is the exception rather than the rule and σ^2 must be estimated from the sample. The obvious way of doing so is use of the sample variance realization $s^2 = S^2(\omega)$.
- We have the following problem. The CLT asserts that, for large enough n, $\omega \mapsto \dfrac{\bar{Y}(\omega) - \mu_0}{\sigma/\sqrt{n}}$ is approximately standard normal. We used that fact to compute $P\left\{ \left| \dfrac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \right| > c_0 \right\}$ and we based the decision to reject or not reject the hypothesis $\mu = \mu_0$ on that number.

- But what happens if we replace σ with $S(\omega)$? If the random variable $\omega \mapsto \dfrac{\bar{Y}(\omega) - \mu_0}{S(\omega)/\sqrt{n}}$ also is approximately standard normal for large n, then our problem is solved. □

To show that the CLT indeed remains in force if σ^2 is replaced by S^2, we must collect some material.

Theorem 13.14 (Student t converges to normal distribution)**.**

Let $T_1, T_2, \dots$ be a sequence of random variables such that $T_j \sim t(df = j)$. Then T_j converges in distribution to a standard normal variable.

PROOF: Omitted. [127] Note though that the graphs of the t–PDFs shown in Remark 13.4 on p.328 visually support the assertion of this theorem.

Lemma 13.1. ★ *Let $\vec{y} := (y_1, \dots, y_n) \in \mathbb{R}^n, (n \in \mathbb{N})$, and $\bar{y} := \sum_{j=1}^{\infty} y_j$ the arithmetic mean of $\vec{y}$. Then,*

(a) $$\sum_{j=1}^{n}(y_j - c)^2 \;=\; \sum_{j=1}^{n}(y_j - \bar{y})^2 \;+\; \sum_{j=1}^{n}(\bar{y} - c)^2\,,$$

(b) *$\bar{y}$ minimizes the expression $\sum_{j=1}^{n}(y_j - c)^2$, where $c \in \mathbb{R}$):*

$$\sum_{j=1}^{n}(y_j - c)^2 \;\geq\; \sum_{j=1}^{n}(y_j - \bar{y})^2 \quad \textit{for all } c \in \mathbb{R}\,,$$

PROOF: To show **(a)**, we observe that

$$\begin{aligned}\sum_{j=1}^{n}(y_j - \bar{y})\,(\bar{y} - c) \;&= \bar{y}\sum_{j=1}^{n} y_j \;+\; \bar{y}\cdot c\sum_{j=1}^{n} 1 \;-\; c\sum_{j=1}^{n} y_j \;-\; \bar{y}\cdot\bar{y}\sum_{j=1}^{n} 1 \\ &= \bar{y}(n\bar{y}) \;+\; (\bar{y}c)n \;-\; c(n\bar{y}) \;-\; (\bar{y}^2)n \;=\; 0\,.\end{aligned} \tag{13.21}$$

Hence,

$$\begin{aligned}\sum_{j=1}^{n}(y_j - c)^2 \;&= \sum_{j=1}^{n}(y_j - \bar{y} + \bar{y} - c)^2 \\ &= \sum_{j=1}^{n}(y_j - \bar{y})^2 \;+\; 2\sum_{j=1}^{n}(y_j - \bar{y})\,(\bar{y} - c) \;+\; \sum_{j=1}^{n}(\bar{y} - c)^2 \\ &\overset{(13.21)}{=} \sum_{j=1}^{n}(y_j - \bar{y})^2 \;+\; \sum_{j=1}^{n}(\bar{y} - c)^2\,.\end{aligned}$$

This proves **(a)**. Clearly, the last expression is minimal when the right–hand summation term vanishes, i.e., when $\bar{y} = c$. This proves **(b)**. ■

[127] A proof can be found at this StackExchange link.

Corollary 13.1. ⋆

The sample variance $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2$ *of any sample* $\vec{Y} := (Y_1, \ldots, Y_n)$, $(n \in \mathbb{N})$, *satisfies*

$$(n-1)S^2 \;=\; \sum_{j=1}^{n} Y_j^2 \;-\; n\,\bar{Y}^2\,.$$

PROOF: We apply formula **(a)** of Lemma 13.1 with $c = 0$ and obtain

$$\sum_{j=1}^{n} Y_j^2 \;=\sum_{j=1}^{n}(Y_j - \bar{Y})^2 \;+\; \sum_{j=1}^{n}\bar{Y}^2\,. \;=\; \sum_{j=1}^{n}(Y_j - \bar{Y})^2 \;+\; n\cdot\bar{Y}^2\,.$$

Thus,

$$(n-1)S^2 \;=\; \sum_{j=1}^{n}(Y_j - \bar{Y})^2 \;=\; \sum_{j=1}^{n} Y_j^2 \;-\; n\,\bar{Y}^2\,. \;\blacksquare$$

Theorem 13.15 (Sample variance converges to population variance)**.**

Let $\vec{Y} := (Y_1, \ldots, Y_n) \in \mathbb{R}^n, (n \in \mathbb{N})$, *be a random sample from the distribution of a random variable* Y *with finite variance* $\sigma^2 < \infty$.

Then the sample variance $S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2$ *converges a.s (hence, also in probability and in distribution) to* σ^2.

PROOF: ⋆ Let $U_n := \frac{n-1}{n}S_n^2$ and $\bar{Y}_n := \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$. By Corollary 13.1,

(A)
$$U_n \;=\frac{1}{n}\sum_{j=1}^{n} Y_j^2 \;-\; \bar{Y}_n^2\,.$$

Since the sample picks Y_j are iid, so are their squares. Note that

$$E[Y_j^2] \;=\; Var[Y_j] + \left(E[Y_j]\right)^2 \;=\; \sigma^2 + \mu^2$$

We apply the Strong Law of Large Numbers to the iid sequences Y_j^2 and Y_j and obtain

(B)
$$\text{a.s.}\!-\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} Y_j^2 \;=\; \sigma^2 + \mu^2\,, \qquad \text{a.s.}\!-\lim_{n\to\infty}\bar{Y}_n \;=\; \mu\,.$$

Next, we apply Theorem 13.3 (Convergence is maintained under continuous transformations) on p.318 to the continuous function $x \mapsto x^2$. It follows from $\text{a.s.}\!-\lim_{n\to\infty}\bar{Y}_n = \mu$ obtained in **(B)**, that

(C)
$$\text{a.s.}\!-\lim_{n\to\infty}\bar{Y}_n^2 \;=\; \mu^2\,.$$

It now follows from the definition of U_n and from **(A)** and **(B)** and **(C)**, that

$$\text{a.s.}\!-\lim_{n\to\infty} S_n^2 = \text{a.s.}\!-\lim_{n\to\infty} \frac{n-1}{n} S_n^2 = \text{a.s.}\!-\lim_{n\to\infty} U_n = (\sigma^2+\mu^2) - \mu^2 = \sigma^2 .$$

It follows from Theorem 13.1 (Relationship between the modes of convergence) on p.317 that convergence $S_n^2 \to \sigma^2$ also takes place in probability and in distribution. ■

We now are able to provide a version of the CLT which allows us to work with $\omega \mapsto \dfrac{\bar{Y}(\omega)-\mu_0}{S(\omega)/\sqrt{n}}$ instead of $\omega \mapsto \dfrac{\bar{Y}(\omega)-\mu_0}{\sigma/\sqrt{n}}$ and solves the issue brought up in Remark 13.8 on p.333.

Theorem 13.16 (CLT – Sample variance version)**.**

Let $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a vector of iid random variables with common expectation $E[Y_j] = \mu$ and finite variance $Var[Y_j] = \sigma^2$. Let Z be a standard normal variable. For $n \in \mathbb{N}$, let

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i , \quad S_n^2 := \frac{1}{n-1}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2 , \quad S_n := \sqrt{S^2} , \quad W_n := \frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} .$$

(Thus, $\bar{Y}_n$ and S_n are sample mean and sample standard deviation of the RSA $\vec{Y}$).

Then W_n converges to Z in distribution as $n \to \infty$.

PROOF: ⋆ [128] Let $U_n := \dfrac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$.

According to the standard version of the CLT (Theorem 13.13 on p.331) $U_n \xrightarrow{\mathbf{D}} Z$ and, according to Theorem 13.15 (Sample variance converges to population variance) on p.335, $S_n^2 \xrightarrow{\mathbf{D}} \sigma^2$.
By Theorem 13.3 (Convergence is maintained under continuous transformations) on p.318,

$$\sigma\, U_n \xrightarrow{\mathbf{D}} \sigma\, Z \qquad \text{and} \qquad S_n = \sqrt{S_n^2} \xrightarrow{\mathbf{D}} \sqrt{\sigma^2} = \sigma .$$

Since the limit σ of S_n is constant, we can apply Slutsky's theorem (Theorem 13.2 on p.318) and obtain

$$W_n = \frac{\sigma U_n}{S} \xrightarrow{\mathbf{D}} \frac{\sigma Y}{\sigma} = Y . \blacksquare$$

Remark 13.9. Note that it follows from Theorem 13.10 on p.329 that, in the special case that the sample picks Y_j are $\mathcal{N}(\mu, \sigma^2)$,

$$W_n = \frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \sim t(\text{df} = n-1) .$$

For that reason, one would rather approximate W_n with a $t(\text{df} = n-1)$ distribution than a standard normal distribution, if the following was true:

(1) The population is known to approximately follow a normal distribution.
(2) The sample size is rather small (rule of thumb: $n < 40$. For such small n, the distribution of W_n may be too far away from $\mathcal{N}(0,1)$, the limit for $n \to \infty$. □

[128] Adapted from stats stackexchange link.

Example 13.9 (WMS Ch.07.3, Example 7.8)**.** ACME Corp. produces X-widgets. When the machines work properly, their weight, in pounds, has a mean of 38 and a variance of 49.

(a) A random sample of $n = 144$ X-widgets was taken yesterday. It had a mean weight of 40 pounds. Does this sample provide sufficient evidence that the manufacturing process is off and the machines need to be recalibrated?

(b) What would be the situation if $n = 100$, $\bar{y} = 39.4$, $\mu = 38$ and $\sigma^2 = 121$?

Solution for (a):

Let $\bar{Y}$ denote the mean of a random sample of $n = 144$ X–widgets from a population with $\mu = 38$ and $\sigma^2 = 49$. According to the CLT (Theorem 13.13 on p.331),

$$U := \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{Y} - 38}{7/12}$$

is approximately $\mathcal{N}(0, 1)$. Thus, if Z denotes a standard normal random variable,

$$P\{\bar{Y} \geq 40\} = P\left\{U \geq \frac{40 - 38}{7/12}\right\} = P\left\{U \geq \frac{2 \cdot 12}{7}\right\} \approx P\left\{Z \geq \frac{2 \cdot 12}{7}\right\} \approx 0.0003\,.$$

Because this probability is so small, it is unlikely that the sampled X–widgets constitute a random sample from machinery that produces them with $\mu = 38$ and $\sigma^2 = 49$. The evidence suggests that the machinery needs to be recalibrated.

Solution for (b):

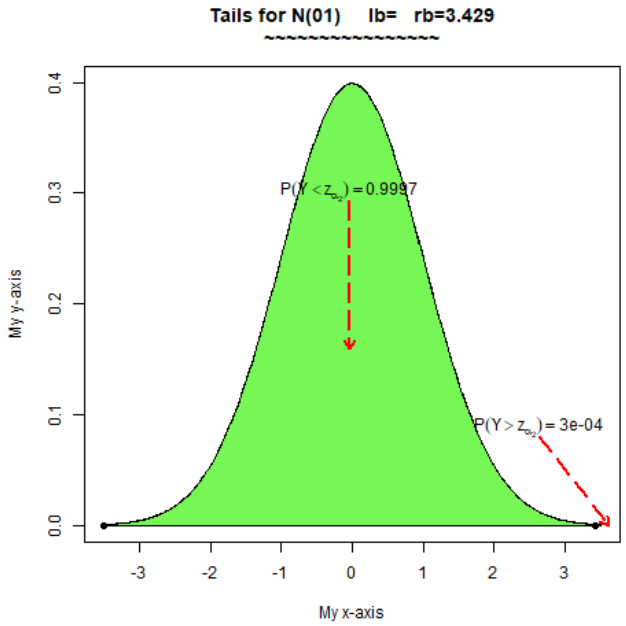
On the other hand, if $n = 100$, $\bar{y} = 39.4$, $\mu = 38$ and $\sigma^2 = 121$, then

$$U = \frac{\bar{Y} - 38}{10/11}$$

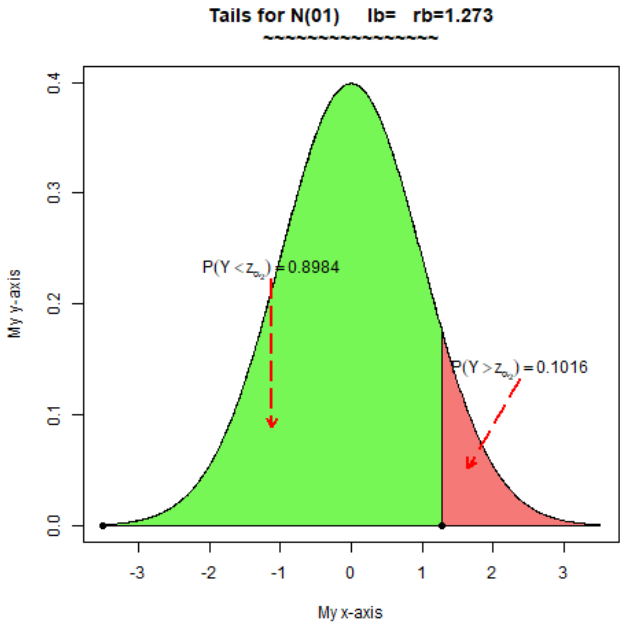
and

$$P\{\bar{Y} \geq 39.4\} = P\left\{U \geq \frac{39.4 - 38}{10/11}\right\} \approx P\{Z \geq 1.273\} \approx 0.1016\,.$$

This means that more than one in ten random samples of size $n = 100$ from a population with $\mu = 38$ and $\sigma^2 = 49$ possess a sample mean above 39.4. That is too big a chance to ignore and one would probably not spend a lot of time and money on adjusting the machines.



13.2 (Figure). **(a) No area in the right tail.**



13.3 (Figure). **(b) right tail area > 10%.**

Because this probability is so small, it is unlikely that the sampled X–widgets constitute a random sample from machinery that produces them with $\mu = 38$ and $\sigma^2 = 49$. The evidence suggests that the machinery needs to be recalibrated.

This example illustrates the use of probability in the process of testing hypotheses, a common technique of statistical inference. □

Example 13.10 (WMS Ch.07.3, Example 7.9)**.** The average life time of an A–widget is documented as $\mu = 4500$ hours, with a standard deviation of $\sigma = 1500$ hours.

A random sample of $n = 81$ A–widgets has been taken and their life times $\vec{y} = y_1, \ldots, Y_n$ have an average of $\bar{Y} = 4250$ hours. Does this deviation of 250 hours from μ indicate that $\mu = 4500$ is outdated and the formal process to determine should be set in motion?

Solution:

```
        round( pnorm(4250, mean=4500, sd=1500/9), 4)
## [1] 0.0668

        round( pnorm((4250 - 4500)*9/1500, mean=0, sd=1), 4)
## [1] 0.0668
```

We do the same steps as in Example 13.9. Because n is rather large,

$$U := \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{Y} - 4500}{1500/9}$$

is approximately $\mathcal{N}(0, 1)$. Thus, if Z denotes a standard normal random variable,

$$P\{\bar{Y} \le 4250\} = P\left\{U \le \frac{4250 - 4500}{1500/9}\right\} \approx P\left\{Z \le \frac{4250 - 4500}{1500/9}\right\} \approx 0.0668\,. \tag{13.22}$$

The probability of obtaining a random sample of 81 A–widgets with a sample mean no higher than 4250 hours under the assumption that the population mean equals 4500, is approximately 0.0668. There is no clear–cut answer for a P–value of this size, even though it is larger than the generally accepted reject/don't reject threshold of 0.05. □

Remark 13.10. When we computed the probabilities of interest in Examples 13.9 and 13.10, we did so by replacing the random variable $\bar{Y}$ which possesses expectation μ and variance $\sigma/\sqrt{n}$, with the random variable $(\bar{Y}-\mu)/(\sigma/\sqrt{n})$ which possesses expectation 0 and variance 1.

Was that necessary? Since $(\bar{Y}-\mu)/(\sigma/\sqrt{n})$ and a standard normal random variable Z have approximately the same distribution, the random variables

$$\bar{Y} \;=\; (\sigma\sqrt{n})\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right) + \mu \quad \text{and} \quad W \;:=\; (\sigma\sqrt{n})\,Z + \mu$$

also have approximately the same distribution. It follows from Theorem 12.5 (Linear combinations of uncorrelated normal variables are normal) on p.309 that $W \sim \mathcal{N}\big(\mu,\, \sigma^2/n\big)$. [129]

One sees that, e.g., (13.22) could have been expressed as follows:

$$P\{\bar{Y} \le 4250\} \;\approx\; P\{W \le 4250\}\,. \tag{13.23}$$

Most statistical software can directly compute probabilities associated with a $\mathcal{N}(\mu,\sigma^2)$ distribution for arbitrary μ an σ^2 For example, the R language handles (13.22) this way:

```
round( pnorm(4250, mean=4500, sd=1500/9), 4)
  ## [1] 0.0668
```

and (13.23) as follows:

```
round( pnorm((4250 - 4500)*9/1500, mean=0, sd=1), 4)
  ## [1] 0.0668
```

Here is an explanation of those calls: pnorm$(x, \text{mean} = \mu, \text{sd} = \sigma^2)$ computes $P\{X \le x\}$, for a $\mathcal{N}(\mu,\sigma^2)$–distributed random variable X. Invoking round$(a, 4)$ rounds a to 4 decimals. □

Example 13.11 (WMS Ch.07.4, Example 7.10)**.** Example 7.10 of the WMS text also discusses an application of the CLT The approximation of a binomial distribution with a normal distribution. You are strongly encouraged to study it. □

Example 13.12 (WMS Ch.07.4, Example 7.11)**.** Example 7.11 of the WMS text also discusses the so called **continuity correction** that should be done whe one approximates a binomial distribution with a normal distribution. You are strongly encouraged to study that example. □

[129] Considering that $\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ and Z have approximately the same distribution and that multiplication by $(\sigma/\sqrt{n})$ followed by addition of μ transforms those random variables into $\bar{Y}$ and W, it should not be a surprise that $\bar{Y}$ and W share the same expectation and variance.

14 Sample Problems for Exams

14.1 Practice Midterm 1 for Math 447 - Chris Haines

Here are some commented excerpts of a practice exam for the first midterm. It was written by Prof. Christopher Haines and forwarded to me by Prof. Adam Weisblat, both at Binghamton University (October 2023).

Exercise 14.1. Practice Midterm 1 (C. Haines) – # 01

SKIPPED □

Answer: N/A ■

Exercise 14.2. Practice Midterm 1 (C. Haines) – # 02

The Lakers and Heat are playing in the NBA Finals. The series is a best–of–seven (first team to win four games clinches the series). The Lakers will win each game with probability $3/4$.

(a) Given that the Heat won game one, what is the probability the Lakers go on to win the series?

(b) Given that the Heat win at least two games in the series, what is the probability the Lakers go on to win the series?

□

Solution:

We denote a sequence of games as $\vec{x} = (x_1, x_2, \ldots, x_n)$, where $n \leq 7$ and $x_j = H$ if the Heat win game j and $x_j = L$ if the Lakers win game j. Note that $n < 7$ is possible, for example, if $\vec{x} = (H, H, H, H)$. (The series is finished.)

Solution to (a):

- Let $A :=$ { The Lakers win the series }
- Let $B :=$ { The Heat win game #1}
- □

Assume that $\vec{x} \in A \cap B$. Then $x_1 = H$ and

- either $x_2 = x_3 = x_4 = x_5 = L \Rightarrow$ one choice
- or one of $x_2, \ldots, x_4$ is H and the other three and x_5 are $L \Rightarrow \binom{4}{1} = 4$ choices
- or two of $x_2, \ldots, x_5$ are H and the other three and x_6 are $L \Rightarrow \binom{5}{2} = 10$ choices
- Thus, $P(A \cap B) = 1 \cdot \frac{1}{4} \cdot \left(\frac{3}{4}\right)^4 + 4 \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^4 + 10 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^4$

We obtain $P(A \mid B) = P(A \cap B)/P(B) = 1701/2048$. ■

Solution to (b): Note that my solution differs from that given in the original (see course materials page!)

- Let $A :=$ { The Lakers win the series },
- $B :=$ { The Heat win at least 2 games },
- $B_2 :=$ { The Heat win precisely 2 games }.
- $B_3 :=$ { The Heat win precisely 3 games },
- Then $A \cap B = A \cap \left(B_2 \uplus B_3\right)$ (Heat cannot win more than 3 if Lakers win the series).

To compute $P(A \cap B) = P(A \cap B_2) + P(B_3 \cap B_3)$, we note that

- either $\vec{x} \in A \cap B_2 \Leftrightarrow$ exactly two of $x_1, \ldots, x_5$ are H and $x_6 = L \Rightarrow \binom{5}{2} = 10$ choices
- or $\vec{x} \in A \cap B_3$, i.e., exactly 3 of $x_1, \ldots, x_6$ are H and $x_7 = L \Rightarrow \binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3!} = 20$ choices
- Thus, $P(A \cap B) = 10 \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^4 + 20 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^4$

Next, we compute $P(B^{\complement})$.

- Let $B_0 := \{$ The Heat win precisely 0 games $\}$. Then $\vec{x} \in B_0 \Leftrightarrow x_1 = x_2 = x_3 = x_4 = L \Rightarrow 1$ choice
- Let $B_1 := \{$ The Heat win precisely 1 game $\}$. Then $\vec{x} \in B_1 \Leftrightarrow$ exactly one of $x_1, \ldots, x_4$ is H and $x_5 = L \Rightarrow 4$ choices
- Further, $P(B^{\complement}) = P(B_0) + P(B_1) = \left(\frac{3}{4}\right)^4 + 4 \cdot \frac{1}{4}\left(\frac{3}{4}\right)^4 = 2\left(\frac{3}{4}\right)^4$.

Thus,

$$P(A \mid B) = \frac{P(A \cap B)}{1 - P(B^{\complement})} = \frac{10 \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^4 + 20 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^4}{1 - 2\left(\frac{3}{4}\right)^4} \quad \blacksquare$$

15 Other Appendices

15.1 Greek Letters

The following section lists all greek letters that are commonly used in mathematical texts. You do not see the entire alphabet here because there are some letters (especially upper case) which look just like our latin alphabet letters. For example: $A =$ Alpha $B =$ Beta. On the other hand there are some lower case letters, namely epsilon, theta, sigma and phi which come in two separate forms. This is not a mistake in the following tables!

α	alpha	θ	theta	ξ	xi	ϕ	phi
β	beta	ϑ	theta	π	pi	φ	phi
γ	gamma	ι	iota	ρ	rho	χ	chi
δ	delta	κ	kappa	ϱ	rho	ψ	psi
ϵ	epsilon	$\varkappa$	kappa	σ	sigma	ω	omega
ε	epsilon	λ	lambda	ς	sigma		
ζ	zeta	μ	mu	τ	tau		
η	eta	ν	nu	υ	upsilon		

Γ	Gamma	Λ	Lambda	Σ	Sigma	Ψ	Psi
Δ	Delta	Ξ	Xi	Υ	Upsilon	Ω	Omega
Θ	Theta	Π	Pi	Φ	Phi		

15.2 Notation

This appendix on notation has been provided because future additions to this document may use notation which has not been covered in class. It only covers a small portion but provides brief explanations for what is covered.

For a complete list check the list of symbols and the index at the end of this document.

Notation 15.1. **a)** If two subsets A and B of a space Ω are disjoint, i.e., $A \cap B = \emptyset$, then we often write $A \uplus B$ rather than $A \cup B$ or $A + B$. The complement $\Omega \setminus A$ of A is denoted $A^{\complement}$. .

b) $\mathbb{R}_{>0}$ or $\mathbb{R}^+$ denotes the interval $]0, +\infty[$, $\mathbb{R}_{\geq 0}$ or $\mathbb{R}_+$ denotes the interval $[0, +\infty[$,

c) The set $\mathbb{N} = \{1, 2, 3, \cdots\}$ of all natural numbers excludes the number zero. We write $\mathbb{N}_0$ or $\mathbb{Z}_+$ or $\mathbb{Z}_{\geq 0}$ for $\mathbb{N} \uplus \{0\}$. $\mathbb{Z}_{\geq 0}$ is the B/G notation. It is very unusual but also very intuitive. $\square$

References

[1] Peter J. Bickel and Kjell A. Doksum. Mathematical Statistics. Holden-Day, San Francisco, 1st edition, 1977.

[2] Thomas Björk. Arbitrage Theory in Continuous Time. Oxford University Press, 2nd edition, 2004.

[3] George Casella and Roger L. Berger. Statistical Inference. Cengage, 2nd edition, 2001.

[4] Saeed Ghahramani. Fundamentals of Probability:With Stochastic Processes. Chapman and Hall, 4th edition, 2018.

[5] Robert Hogg, Joseph McKean, and Allen Craig. Introduction to Mathematical Statistics. Pearson, 8th edition.

[6] Vladislav Kargin. Lecture Notes for the Introduction to Probability Course. May 24, 2022 edition, 2022.

[7] Hossein Pishro-Nik. Introduction to Probability, Statistics, and Random Processes, available at https://www
Kappa Research, LLC, 2014.

[8] Sheldon M. Ross. A First Course in Probability. Macmillan, New York, 3rd edition, 198.

[9] Georgi Evgen'evitch Shilov. Elementary Real and Complex Analysis. Dover, Mineola, 1st edition, 1996.

[10] Steve E. Shreve. Stochastic Calculus for Finance I: The Binomial Asset Pricing Model. Springer, 1st edition, 2003.

[11] Steve E. Shreve. Stochastic Calculus for Finance II: Continuous-Time Models. Springer, 1st edition, 2004.

[12] James Stewart. Single Variable Calculus. Thomson Brooks Cole, 7th edition, 2012.

[13] D. Wackerly, W. Mendenhall, and R.L. Scheaffer. Mathematical Statistics with Applications. Thomson Brooks/Cole, 7th edition, 2008.

[14] Richard E. Williamson and Hale F. Trotter. Multivariable Mathematics. Prentice Hall, 3rd edition, 1995.

List of Symbols

Index