Lecture Notes for Math 447 - Probability

Student edition with proofs

Michael Fochler
Department of Mathematics
Binghamton University

Last update: December 7, 2023

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 [5] Wackerly, D. and Mendenhall, W. and Scheaffer, R.L.: Mathematical Statistics with Applications, 7th edition.

At this point in time (December, 2023) it focuses 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 understand 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 proposiitions or parts thereof.

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

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 and what is the risk that I'll be off by more than 0.5 percent?"

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.

Then we expect that in the long run, i.e., for large k, close to half of $X_1, X_2, \ldots, X_k$ result in an even outcome. In the language of mathematics, if we write P for probability, and n_k for the number of even outcomes during those k rolls, we define

$$P\{\text{ rolling the die yields an even outcome }\} = \lim_{k\to\infty} \frac{n_k}{n}$$

(and we expect this particular limit to be 0.5.) More precisely, this would be our method to determine the **empirical probability** of that event.

We also could have used the concept of a fair die instead, i.e., a die for which each of the outcomes $1, 2, \ldots, 6$ is equally likely, so each outcome must have the same likelihood (probability) of $1/6$, so

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

Note that fair dice do not exist in the real world. Matter of fact, if we had a sample of 100 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 100 different answers, due to imperfections in the manufacturing process.

We model the random action of rolling a fair die as follows.

- We write Ω for the set of all potential outcomes, [1] i.e., $\Omega = \{1,2,3,4,5,6\}$.
- We associate with each element ω of Ω the probability $P(\{\omega\}) = 1/6$.
- Let $A \subseteq \Omega$, i.e., A is a subset of Ω, i.e., each element of A also belongs to Ω. We associate with A the probability $P(A) = \sum\limits_{\omega\in A} P(\{\omega\})$.

  Example: Let $A = \{2,4,6\}$, the set of all even outcomes. Then, no different from above,

  $$P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/6 + 1/6 + 1/6 = 1/2\,.$$

Observe that this assignment $A \mapsto P(A)$ satisfies the following.

- $0 \le P(A) \le 1$. Here $\emptyset$ denotes the empty set which contains no elements.
- $P(\emptyset) = 0$. Here $\emptyset$ denotes the empty set which contains no elements.
- $P(\Omega) = 6(1/6) = 1$.
- If the subsets A, B of Ω have no elements in common (we speak of nutually disjoint sets), then the union $P(A \cup B)$ satisfies

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

[1] Ω denotes the Greek capital letter Omega. For a list of all Greek letters see Section 12.1 (Greek Letters) on page 187.

We are ready for a formal definition of probability. **It is PRELIMINARY and will be amended!**

Definition 1.1 (Sets)**.** [Probability - Preliminary Definition]
A probability P on a set Ω is a function [2] which assigns to each subset A of Ω a real number $P(A)$ between 0 and 1 such that

- $P(\emptyset) = 0$ and $P(\Omega) = 1$. Here $\emptyset$ denotes the empty set which contains no elements.
- If the subsets A, B of Ω have no elements in common, then probability is **additive**:

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

Note the following about this definition.

- It says nothing about how one should interpret the number $P(A)$!
- Empirical probability satisfies those three conditions. This is obvious for the first two. As to #3, let A and B be two events with nothing in common, and for which we want to determine $P(A)$ and $P(B)$ empirically. for $k = 1, 2, \dots$ 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,

If the subsets A, B of Ω have no elements in common, then empirical probability satisfies

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

[2] we'll review functions briefly in Section 2.1 (Sets, Numbers, Sequences and Functions) on page 8.

1.3 Blank Page after Ch.1

This page is intentionally left blank!

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:

$$\begin{aligned}
&\{x : x \text{ is an integer and } 18 \leq x \leq 24\},\\
&\{\alpha : \alpha \text{ is an integer and } 18 \leq \alpha \leq 24\},\\
&\{\mathfrak{z} : \mathfrak{z} \text{ is an integer and } 18 \leq \mathfrak{z} \leq 24\} \quad \square
\end{aligned}$$

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 "$B \supsetneq A$")

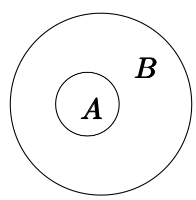


Figure 2.1: Set inclusion: $A \subseteq B,\ B \supseteq A$ $\square$

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, "$\Leftrightarrow$" 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 "$\Leftrightarrow$". $\square$

Definition 2.4 (unions, intersections and disjoint unions)**.** Given are two arbitrary 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 $A \cup B$.

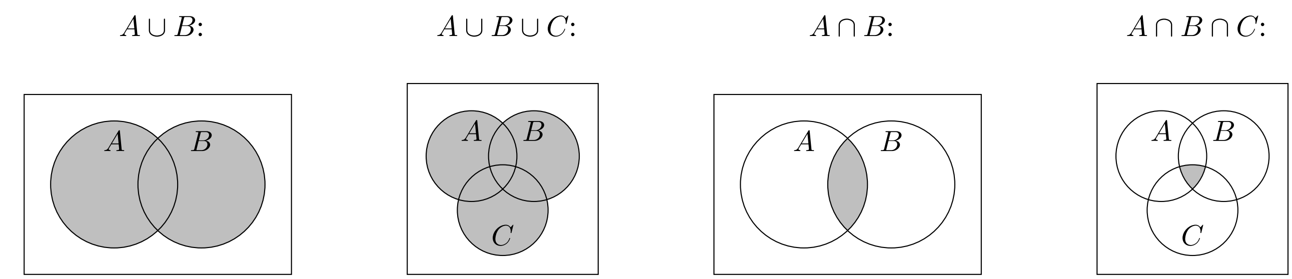


Figure 2.2: Union and intersection of sets

Since $A \cup B = B \cup A$ and $A \cap B = B \cap A$ and $A \uplus B = B \uplus A$, it is obvious how to specify those operations to any finite or infinite collection of sets. Let J be a nonempty, finite or infinite subset of the set $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \dots\}$ of all integers. In particular, $J = \mathbb{Z}$ is allowed. Assume that each $j \in J$ is associated with a set A_j. [3] We say that

- The **union** $\bigcup\limits_{j \in J} A_j$ is defined as the set of all elements which belong to at least one A_j, where $j \in J$.
- The **intersection** $\bigcap\limits_{j \in J} A_j$ 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 $\biguplus\limits_{j \in J} A_j$ rather than $\bigcup\limits_{j \in J} A_j$. $\square$

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

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

[3] You might call this a **collection** of sets A_i which is **indexed by** the elements of J and write $(A_j)_{j \in J}$ for this **indexed collection**.

If $J = \{k_\star, k_\star+1, k_\star+2, \dots, \}$, we also write

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

Examples: If $I = \{-1, 0, 1, 2\}$, then $\bigcap\limits_{i\in I} A_i \;=\; \bigcap\limits_{i=-1}^{2} A_i \;=\; A_{-1} \cap A_0 \cap A_1 \cap A_2$.

If $U = \{5, 6, 7, \dots\}$, then $\bigcup\limits_{j\in U} C_j \;=\; \bigcap\limits_{j=5}^{\infty} C_j \;=\; C_5 \cup C_6 \cup C_7 \cup \cdots$. $\square$

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

(2.2) $$A \cap B \;\subseteq\; A \;\subseteq\; A \cup B,$$

(2.3) $$A \subseteq B \;\Rightarrow\; A \cap B = A \text{ and } A \cup B = B,$$

(2.4) $$A \subseteq B \;\Rightarrow\; A \cap C \;\subseteq\; B \cap C \text{ and } A \cup C \;\subseteq\; B \cup C.$$

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 2.5 (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:

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

- 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:

  (2.6) $$A \triangle B := (A \cup B) \;\setminus\; (A \cap B) \;\square$$

Definition 2.6 (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 2.3.

(a) Often the context are the real numbers and their subsets. An appropriate universal set will then be $\mathbb{R}$. [4]

(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 2.7 (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:

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

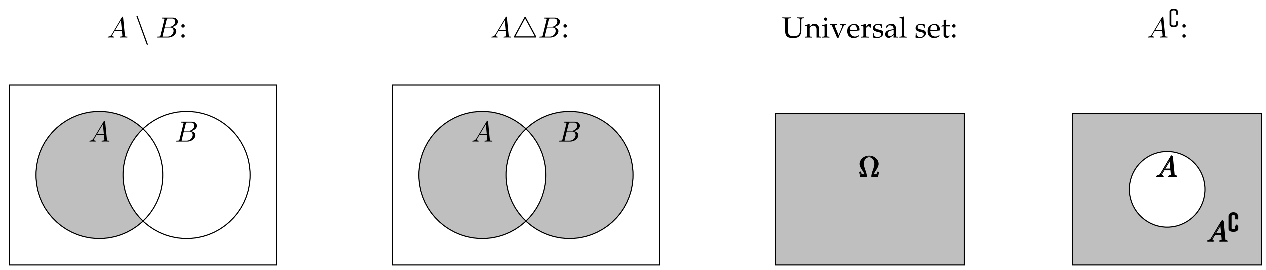


Figure 2.3: Difference, symmetric difference, universal set, complement

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 2.4 (Complement of a set relative to the unit interval)**.** Assume we are exclusively dealing with the unit interval, i.e., $\Omega = [0,1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$. 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" [5] 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

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

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.

[4] $\mathbb{R}$ 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.16).

[5] 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.

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

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

PROOF: The proof is left as exercise 2.2. See p.29. ■

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: $\boxed{\star}$ We only prove (2.11). The proof of (2.12) is left as exercise 2.1.

PROOF of "$\subseteq$": Let $x \in (A \cup B) \cap C$. It follows from (2.2) on p.11 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.11 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:*

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

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 2.8 (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 2.9 (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 $A \in \mathfrak{A}$. □

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

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

Example 2.5.
- **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.2.9 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 $[1, \infty[$. □

Definition 2.10 (Size of a set)**.**

a. Let X be a finite set, i.e., a set which only contains finitely many elements. We write $\left|X\right|$ for the number of its elements, and we call $\left|X\right|$ the **size** of the set X.

b. For infinite, i.e., not finite sets Y, we define $|Y| := \infty$. $\square$

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 $a, b, c, \dots$ 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 $a, b, c, \dots$.

> 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.

• $\forall\, x \dots$	For all x...
• $\exists\, x$ s.t. ...	There exists an x such that ...
• $\exists!\, x$ s.t. ...	There exists a UNIQUE x such that ...
• $P \Rightarrow Q$	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 2.11 (Types of numbers)**.** Here is a definition of the various kinds of numbers in a nutshell.

> $\mathbb{N} := \{1, 2, 3, \dots\}$ denotes the set of **natural numbers**.
> $\mathbb{Z} := \{0, \pm 1, \pm 2, \pm 3, \dots\}$ 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:

$\mathbb{N}_0 := \mathbb{Z}_+ := \mathbb{Z}_{\geq 0} := \{0, 1, 2, 3, \dots\}$ 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 2.12 (Intervals of Numbers)**.** For $a, b \in \mathbb{R}$ we have the following intervals.

- $[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$ 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.
- $[a, b[:= \{x \in \mathbb{R} : a \leq x < b\}$ and $]a, b] := \{x \in \mathbb{R} : a < x \leq b\}$ 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":

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

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 $a \geq b$ $\square$

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

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

$$\begin{aligned} [\,3, n\,]_{\mathbb{Z}} &= [\,3,\ n] \cap \mathbb{Z} \;=\; \{k \in \mathbb{Z} : 3 \leq k \leq n\}, \\ ]-\infty, 7\,]_{\mathbb{Z}} &=]-\infty, 7\,] \cap \mathbb{Z} \;=\; \{k \in \mathbb{Z} : k \leq 7\} \;=\; \mathbb{Z}_{\leq 7}, \\ ]a, b[_{\mathbb{Q}} &=]a, b[\cap \mathbb{Q} \;=\; \{q \in \mathbb{Q} : a < q < b\}. \end{aligned}$$

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 2.13 (Absolute value, positive and negative part)**.** Let $x, y \in \mathbb{R}$. We define the following.

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

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

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

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 2.8. For any real number x we have

$$\sqrt{x^2} \;=\; |x|. \quad \Box \tag{2.17}$$

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:*

$$\textit{Triangle Inequality}: \quad |a_1 + a_2 + \dots + a_n| \;\leq\; |a_1| + |a_2| + \dots + |a_n| \tag{2.18}$$

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

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 <u>superset</u> 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 $0 \le P(A) \le 1$. 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

$$f : C \longrightarrow D$$

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 2.14 (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

(2.19) $$f : X \to Y, \qquad x \mapsto f(x).$$

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

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

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 2.9. 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 2.15 (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 $y \in Y$. $\square$

Remark 2.10. that One can show that a function f has an inverse f^{-1} if and only if

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

Remark 2.11. 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 2.6. If h is a function, we write Dom_h and Cod_h for its domain and codomain. Be sure you understand the following:

(a) $f : \mathbb{R} \to \mathbb{R};\; x \to e^x$ <u>does not have an inverse</u> $f^{-1}(y) = \ln(y)$ 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
$$\begin{aligned} &Dom_{g^{-1}} = Cod_g =]0,\infty[, && Cod_{g^{-1}} = Dom_g = \mathbb{R},\\ &e^{\ln(y)} = y \text{ for } 0 < y < \infty, && \ln(e^x) = x \text{ for all } x \in \mathbb{R}. \;\square \end{aligned}$$

Definition 2.16 (Restriction/Extension of a function). $\boxed{\star}$ 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

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

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 2.17. 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 $j = n_\star, n_\star + 1, n_\star + 2, \dots$.
- **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 $j = n_\star, n_\star + 1, \dots, n^\star$.

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

- In the first case we usually write $x_{n_\star}, x_{n_\star+1}, x_{n_\star+2}, \dots$ 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 $x_{n_\star}, x_{n_\star+1}, \dots, x_{n^\star}$ 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**, $x_{n_\star}, x_{n_\star+1}, x_{n_\star+2}, \dots$. $\square$

Example 2.7.

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

(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 $[-\frac{4}{3}, \frac{4}{3}]$, $[-\frac{5}{4}, \frac{5}{4}]$, $[-\frac{6}{5}, \frac{6}{5}], \dots$. This is a sequence of sets! $\square$

Remark 2.12 (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 2.7**(1)**, we could chose $\mathbb{Z}$ as that codomain. In Example 2.7**(2)** $2^{\mathbb{R}}$, the power set of $\mathbb{R}$ would be an appropriate choice. □

Definition 2.18.

- 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 $(x_{n_j})_{j \leq K}$ 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 2.13. Does it matter whether we look at a sequence $(x_j)_{j \in J}$ 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). □

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 2.19 (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 a nonempty set **uncountable** if it is not countable, i.e., its elements cannot be sequenced.
(c) 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!* $\square$

$\boxed{\star}$ 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 $\mathbb{Q} = \bigcup\limits_{d=1}^{\infty} A_d$ (WHY?)

Example 2.8. $\boxed{\star}$ 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

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

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

There is a name for those "generalized sequences" $\left(x_i\right)_{i\in I}$ 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 2.20 (Families). $\boxed{\star}$

Let I and X be nonempty sets such that each $i \in I$ is associated with some $x_i \in X$. Then
- **a.** $\left(x_i\right)_{i\in I}$ 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** $(x_i)_{i\in I}$. $\square$

Remark 2.14 (Families are functions). that

We saw in example 2.12 on p.23 that sequences $\left(x_n\right)_n$ 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 $\left(x_i\right)_{i\in I}$ in X is written in a way that each member explicitly tracks the index that it is associated with, i.e., we write $\left(i, x_i\right)_{i\in I}$. The set

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

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.10 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 [6]

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

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 2.21 (Arbitrary unions and intersections)**.** 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** $\bigcup\limits_{j \in J} A_j := \{x : \exists\, i_0 \in J \text{ s.t. } x \in A_{i_0}\}$.
- The **intersection** $\bigcap\limits_{j \in J} A_j = \{x : \forall\, i \in J : x \in A_i\}$.
- If the sets A_i are disjoint, we often write $\biguplus\limits_{j \in J} A_j$ rather than $\bigcup\limits_{j \in J} A_j$.
- 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)** $X = \biguplus\limits_{j \in J} B_j$. See Definition 2.9 on p.14. $\square$

Remark 2.15. $\boxed{\star}$ For typographical reasons I sometimes use the following notation.

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

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

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

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

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

Be sure to understand the following example (draw a picture!)

[6] See paragraph 2.2.0.4 (Some Convenient Shorthand Notation) on p.16 about $\forall$ and $\exists$.

Example 2.9. $\boxed{\star}$ For $a, b \in \mathbb{R}$, let $Q_{(a,b)} := \{(x,y) \in \mathbb{R}^2 : |x-a| \leq 3/2,\ |y-b| \leq 3/2\}$. Thus, $Q_{(a,b)}$ is the square in the plane with center (a,b) and side length 3. Compute $\bigcap\limits_{(a,b)\in K} Q_{(a,b)}$ and $\bigcup\limits_{(a,b)\in K} Q_{(a,b)}$.

For $K = \{(a,b) \in \mathbb{R}^2 : -1 \leq a, b \leq 1\}$, compute $\bigcap\limits_{(a,b)\in K} Q_{(a,b)}$ and $\bigcup\limits_{(a,b)\in K} Q_{(a,b)}$.

Solution:

Let $U := \bigcap\limits_{(a,b)\in K} Q_{(a,b)}$ and $V := \bigcup\limits_{(a,b)\in K} Q_{(a,b)}$.

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,

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

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 $-(1/2) \leq y \leq 1/2$. Thus,

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

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,

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

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). [7]

Proposition 2.5 (De Morgan's Law for sequences of sets)**.** *Let $(A_n)_n$ be a finite or infinite sequence 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:*

$$(2.22) \qquad \textbf{(a)} \quad \left(\bigcup_k A_k\right)^{\complement} = \bigcap_k A_k^{\complement}; \qquad \textbf{(b)} \quad \left(\bigcap_k A_k\right)^{\complement} = \bigcup_k A_k^{\complement};$$

PROOF:

Not very complicated, but we skip it ■

Note that the order of the sequencing does not matter for De Morgan and the next proposition.

[7] Matter of fact, both propositions extend to arbitrary families.

Proposition 2.6 (Distributivity of unions and intersections)**.** *Let $(A_n)_n$ be a finite or infinite sequence of sets and let B be a set. Then*

$$\bigcup_j (B \cap A_j) = B \cap \bigcup_j A_j, \tag{2.23}$$

$$\bigcap_{j \in I} (B \cup A_j) = B \cup \bigcap_j A_j. \tag{2.24}$$

PROOF: ■

2.5 Cartesian Products

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

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

and quads

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

which you certainly have encountered in multivariable calculus.

Definition 2.22 (Cartesian Product)**.** Let X and Y be two sets The set

$$X \times Y := \{(x, y) : x \in X, y \in Y\} \tag{2.25}$$

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

$$X_1 \times X_2 \cdots \times X_n := \{(x_1, x_2, \ldots, x_n) : x_j \in X_j \text{ for each } j = 1, 2, \ldots n\} \tag{2.26}$$

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

$$(x, y) \in \mathbb{R}^2, \quad (a, b) \in \mathbb{R}^2, \quad (x, y, z) \in \mathbb{R}^3, \quad (a, b, c) \in \mathbb{R}^3. \tag{2.27}$$

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 $\big(\beta_1, \beta_2, \dots, \beta_m\big)$ is an element of $\mathbb{R}^m$ which we call an m–dimensional **vector** of real numbers.

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

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 (2.27) does not scale for higher dimensional vectors, in particular, if the dimension is arbitrary. On the other hand, $\big(\beta_1, \beta_2, \dots, \beta_m\big)$ is very suitable. But this is very lengthy notation, so we use the symbol for the subscripted components (that's β) and write an arrow on top to indicate that we are dealing with a vector. [8]

We will use as much as possible this arrow notation for vectors. Here are some examples.

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

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

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),

$\vec{x} \in X^n$ is shorthand for $\vec{x} \;=\; \big(x_1, x_2, \dots, x_n\big) \in X^n$ (i.e., $x_j \in X$ for $j = 1, 2, \dots, n$.)

We also extend this convention to the case $X_1 \times \cdots \times X_n$ with potentially different sets X_j. This is best explained by example. Having pairs of numbers $a_i < b_i$ for $i = 1, 2, \dots, d$,

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

Example 2.11. 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 mopdeled 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 $\square$

[8] We borrow that notation from physics.

Here is an abstract example.

Example 2.12. The graph Γ_f of a function with domain X and codomain Y (see def.2.20) is a subset of the cartesian product $X \times Y$. $\square$

Proposition 2.7. *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.,

$$\left| X_1 \times X_2 \times \cdots \times X_n \right| = |X_1| \cdot |X_2| \cdot |X_3| \cdots |X_n| . \tag{2.28}$$

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

$$\left|X_1 \times X_2 \times X_3\right| = \left|(X_1 \times X_2) \times X_3\right| = \left|X_1 \times X_2\right| \cdot \left|X_3\right| . = \left|X_1\right| \cdot \left|X_2\right| \cdot \left|X_3\right| .$$

We repeat this procedure for $n = 3, 4, 5, \dots$ sets.
Case n: We arrange the elements of $X_1 \times X_2 \times \times X_{n-1}$ into a single row and

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

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 $\left|X_1 \times \cdots \times X_{n-1}\right| = \left|X_1\right| \cdots \left|X_{n-1}\right|$. Hence,

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

2.6 Exercises for Ch.2

2.6.1 Exercises for Sets

Exercise 2.1. Prove (2.12) of prop.2.2 on p.13.

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**). $\blacksquare$

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.** $\{y\} \subseteq X$ $\square$

For the subsequent exercises refer to Definition 2.10 on p.15 of the size $|A|$ of a set A and to Definition 2.22 on p.27 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.** $\{y\} \in X \Delta Y$ □

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

a. What is $X \times Y$? **c.** What is $|X \times Y|$? **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 $|X \times Y|$? **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 $x \in A$ then $A = (A \setminus \{a\}) \uplus \{a\}$.
b. If $x \notin A$ then $A = (A \uplus \{a\}) \setminus \{a\}$.
□

2.7 Addenda to Ch.2

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 $j < n \Rightarrow x_j \leqq x_n$.
- We call x_n a **strictly increasing** sequence, if $j < n \Rightarrow x_j < x_n$.
- We call x_n a **nonincreasing** or **decreasing** sequence, if $j < n \Rightarrow x_j \geqq x_n$.
- We call x_n a **strictly decreasing** sequence, if $j < n \Rightarrow x_j > x_n$.
- We write $x_n \uparrow$ for nondecreasing x_n, and $x_n \uparrow x$ to indicate that $\sup_n x_n = x$,
- We write $x_n \downarrow$ for nonincreasing x_n, $x_n \downarrow x$ to indicate that $\inf_n x_n = x$. □

Example 2.13.

(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.
□

2.8 Blank Page after Ch.2

This page is intentionally left blank!

3 The Probability Model

3.1 Probability Spaces

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 and let this motivate us to give a preliminary definition of probability as a function

$$P : 2^{\Omega} \longrightarrow [0,1]$$

which assigns to each element A in the power set of a given set Ω a number $P(A)$ between zero and one, such that

(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 are disjoint, then probability is **additive**:

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

Note that additivity holds for three disjoint sets $A, B, C \in 2^{\Omega}$ since,

$$(\star) \qquad P(A \biguplus B \biguplus C) \;= P[(A \biguplus B) \biguplus C] \;=\; P(A \biguplus B) + P(C) \;=\; P(A) + P(B) + P(C)\,.$$

From $(\star)$ you get additivity for four disjoint $A, B, C, D \in 2^{\Omega}$ since,

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

Now that you have additivity for four disjoint sets, you get it by the same method for five, and then for six, ... and thus, for any finite number of disjoint subsets $A_1, \dots, A_n$ of Ω.

But we are not satisfied since it has proven extremely fruitful to replace **(b)** with the stronger condition

(b′) If $\big(A_n\big)_{n \in \mathbb{N}}$ is a sequence of disjoint subsets of Ω, then probability is "σ–**additive**": [9]

$$P\left(\biguplus_{j=1}^{\infty} A_j \right) \;=\; \sum_{j=1}^{\infty} P(A_j)\,.$$

Unfortunately, this comes with a trade-off. Consider the following example.

Example 3.1. A point located somewhere at $]-\infty, 0[$ starts moving to the right at a constant velocity and is stopped completely 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 $[\frac{k-1}{n}, \frac{k}{n}]$ $(k = 1, 2, \dots, n)$.

It should be obvious that the only reasonable probability function on $\Omega := [0,1]$ is

$$P : [0,1] \to [0,1]\,; \qquad [\alpha, \beta] \;\mapsto\; \beta - \alpha\,,$$

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

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 $P(\Omega) = 1$.

Unfortunately, it has been proven [10] that no σ–additive function that satisfies those properties exists on the entire power set of $[0,1]$.

The only way out of this dilemma without sacrificing σ–additivity is to relax the condition that $P(A)$ must exist for all $A \subseteq \Omega$. $\square$

It follows from this example that we must define probability as a function

$$P : \mathfrak{F} \longrightarrow [0,1]\,, \qquad \text{where } \mathfrak{F} \text{ is a suitable subset of } 2^{\Omega},$$

which satisfies $P(\emptyset) = 0$ and $P(\Omega) = 1$ and

$$P\left(\biguplus_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \quad \text{for disjoint } A_1, A_2, \dots \in \mathfrak{F}.$$

To answer the question what conditions a useful domain $\mathfrak{F}$ for a probability function P should satisfy, it helps to remember De Morgan's Law for finite or infinite sequences of sets. See Proposition 2.5 on p.26. Also, the following proposition which shows how to rewrite any countable union (finite or infinite) as a DISJOINT union will be relevant.

Proposition 3.1 (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*

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

Then

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

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

(c) *The sets C_j are mutually disjoint and $\bigcup\limits_{j=1}^{n} A_j = \biguplus\limits_{j=1}^{n} C_j$.*

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

PROOF: $\boxed{\star}$ **(a)** and **(b)** are trivial. For the proof of **(c)** and **(d)**, convince yourself that

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

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

We return to the question what the domain $\mathfrak{F}$ of a probability should satisfy.

[10] such a proof is outside the scope of these notes.

If A has a probability $P(A)$, then $A^{\complement}$ should have probability $1 - P(A)$. Since probabilities can only be assigned to elements of $\mathfrak{F}$, we want

(A) $$A \in \mathfrak{F} \;\Rightarrow\; A^{\complement} \in \mathfrak{F}\,.$$

If $A_n \in \mathfrak{F}$ are pairwise disjoint, then $\biguplus_{j=1}^{\infty} A_j$ should have probability $\sum_{j=1}^{\infty} P(A_j)$. Since probabilities can only be assigned to elements of $\mathfrak{F}$, we want

$$A_n \in \mathfrak{F} \text{ disjoint } \Rightarrow \biguplus_{j=1}^{\infty} A_j \in \mathfrak{F}\,.$$

Since we have seen that any union of a sequence of events can be written as a disjoint union, we need more than the above. We really want

(B) $$A_n \in \mathfrak{F} \text{ arbitrary } \Rightarrow \bigcup_{j=1}^{\infty} A_j \in \mathfrak{F}\,.$$

Also, it is very reasonable to demand that $P(\emptyset) = 0$ for the impossible event which contains no potential outcomes, i.e., the empty set. it is just as reasonable to ask that $P(\Omega) = 1$ for the sure event, Ω, since it contains all potential outcomes. Thus, we ask that

(C) $$\emptyset \in \mathfrak{F} \qquad \text{and} \qquad \Omega \in \mathfrak{F}\,.$$

All this leads to the definition of a σ–algebra.

Definition 3.1 (σ–algebra)**.** Let Ω be a nonempty set and $\mathfrak{F} \subseteq 2^{\Omega}$ a collection of subsets of Ω, 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**.

$\mathfrak{F}$ is also called a σ**–field**, but this is considered old–fashioned terminology. $\square$

Proposition 3.2. *σ–algebras $\mathfrak{F}$ satisfy the following.*

(a) $\Omega \in \mathfrak{F}$.

(b) $A_1, A_2, \dots, A_n \in \mathfrak{F} \;\Rightarrow\; A_1 \cup A_2 \cup \dots \cup A_n \in \mathfrak{F}$.

(c) *Let $n \in \mathbb{N}$ and $A_1, A_2, \dots \in \mathfrak{F}$. Let $A = \bigcap_{k=1}^{n}$ and $B = \bigcap_{k=1}^{\infty}$. Then $A \in \mathfrak{F}$ and $B \in \mathfrak{F}$.* $\square$

PROOF: $\boxed{\star}$

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 $A_1, \dots, A_n$ 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.3.1**(b)** that $\bigcup\limits_{j=1}^{\infty} B_j \in \mathfrak{F}$. 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 $\bigcup\limits_{j=1}^{n} A_j \in \mathfrak{F}$. This proves **(b)**.

PROOF of **(c)**: According to De Morgan's laws, any countable intersection can be written as the union of its complements. Thus we automatically get from **(A)** and **(B)** that countable intersections of a sequence in $\mathfrak{F}$ will again belong to $\mathfrak{F}$.

Here is a detailed argument. For each j let $C_j := A_j^{\complement}$. Further, let $C := \bigcup\limits_{j=1}^{n} C_j$ and $D := \bigcup\limits_{j=1}^{\infty} C_j$.

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)**. ■

Definition 3.2 (Probability measures and probability spaces)**.**

Given are a nonempty set Ω with a σ–algebra $\mathfrak{F} \subseteq 2^{\Omega}$ and a function

$$P : \mathfrak{F} \longrightarrow [0,1]; \quad A \mapsto P(A) \qquad \text{as follows.}$$

(3.1) $\quad P(\emptyset) \ = \ 0\,,$ (3.2) $\quad P(\Omega) \ = \ 1\,,$

(3.3) $\quad (A_n)_{n \in \mathbb{N}} \in \mathfrak{F}$ disjoint $\Rightarrow \ P\Big(\biguplus_{n \in \mathbb{N}} A_n\Big) \ = \ \sum_{n=1}^{\infty} P(A_n) \ = \ \sum_{n \in \mathbb{N}} P(A_n).$ (σ–additivity)

- We call P a **probability measure** or simply a **probability**
- The triplet $(\Omega, \mathfrak{F}, P)$ is called a **probability space**.
- If Ω is countable, we call $(\Omega, \mathfrak{F}, P)$ a **discrete probability space**.
- We call the elements of Ω **outcomes** and the subsets of Ω **events**. □

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 3.1. The WMS text uses different notation

- What we call a probability space, WMS calls a **sample space**.
- WMS uses the letter S rather than Ω for the "carrier set" and completely ignores σ–algebras. Thus, WMS refers to a sample space (S, P) rather than to a probability space $(\Omega, \mathfrak{F}, P)$
- WMS typically writes $x, y, \dots$ rather than ω for the outcomes.

I prefer to use the term "probability space" since we 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. However I give you a choice in this matter.

You may refer in your quizzes, exams and homework to sample spaces and the symbol S, but you must write $(S, \mathfrak{F}, P)$ rather than (S, P) if the role of a σ–algebra $\mathfrak{F}$ matters.

And more good news: We have introduced σ–algebras to properly deal with the issue that was raised in Example 3.1 on p.32 It won't be long and we will on only few occasions deal with σ–algebras and usually refer to a probability space (Ω, P) $\square$

Remark 3.2. How do we interpret formula (3.3) for σ–additivity in the definition of a probability measure, $P\Big(\biguplus\limits_{n\in\mathbb{N}} A_n\Big) = \sum\limits_{n=1}^{\infty} P(A_n) = \sum\limits_{n\in\mathbb{N}} P(A_n)$? What is the meaning of $\biguplus\limits_{n\in\mathbb{N}} A_n$ as opposed to $\biguplus\limits_{n=1}^{\infty} A_n$; and what is the meaning of $\sum\limits_{n\in\mathbb{N}} P(A_n)$, as opposed to $\sum\limits_{n=1}^{\infty} P(A_n)$?

(a) Unions are defined without any reference to an order "first A_1, then A_2, then $A_3, \dots$", since the definition of $a \in \biguplus\limits_{n\in\mathbb{N}} A_n$ 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 $\biguplus\limits_{n=1}^{\infty} A_n$ is that it looks more familiar. By the way, what was said here about disjoint unions also applies to arbitrary unions and to intersections.

(b) But what about the summation $\sum\limits_{n\in\mathbb{N}} P(A_n)$? Does it really not matter in which order we add the terms of an inifinite series? The answer is that this depends. If you are curious, look at this optional footnote. [11] You will find there the following. Because $P(A_j) \geq 0$ for all j, the value of the infinite series $\sum\limits_{n=1}^{\infty} P(A_n)$ does not depend on the order in which the terms $P(A_j)$ are arranged. $\square$

In Section 1.2 (A First Look at Probability) we used throws of a die to illustrate the concepts of

[11] $\boxed{\star}$ Let $\sum a_j$ be an infinite series. Then **(1)** If $\sum_{j=1}^{\infty} |a_j| = |a_1| + |a_2| + \cdots < \infty$, (we call such a series **absolutely convergent**), then the original series converges (to a finite limit), and any rearrangement $\sum_{j=1}^{\infty} a_{n_j} = a_{n_1} + a_{n_2} + \cdots$ converges to the same limit. **(2)** If $\sum_{j=1}^{\infty} b_j$ converges to a real number ($\neq \pm\infty$), but $\sum_{j=1}^{\infty} |b_j| = \infty$, then the following is true: Pick any $-\infty \leq x \leq \infty$. Then you can rearrange the terms b_j in such a way that the rearranged sequence, call it $\sum_{j=1}^{\infty} b_{n_j}$, converges to x. In other words, you can jumble the terms such that the limit is π. A different permutation of the indices has limit 0, yet another converges to $-\sqrt{e^{30}}$, ... **(3)** For brevity, let $p_j := P(A_j)$. Then $0 \leq p_j \leq 1 \Rightarrow p_j = |p_j|$ and hence,

$$\sum_{j=1}^{\infty} |p_j| = \sum_{j=1}^{\infty} p_j = P\big(\bigcup_j A_j\big) = \leq P(\Omega) = 1 < \infty.$$

Thus, $\sum p_j$ converges absolutely and the order of the p_j is immaterial. But then the order of the A_j is immaterial, and this allows us to write $\sum\limits_{n\in\mathbb{N}} P(A_n)$ for $\sum\limits_{n=1}^{\infty} P(A_n)$.

random actions and their potential outomes and let this motivate us to give a preliminary definition of probability as a function

Example 3.2. **(a)** We model k rolls of a fair die ($k \in \mathbb{N}$) as follows. Let

$$\Omega := \{1,2,3,4,5,6\}^k = \{(a_1, a_2, \dots, a_k) : a_j = 1,2,..,6 \text{ for each } j = 1,2,\dots,k\}.$$

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.,

$$\omega_1 = (1,1,\dots,1,1),\ \omega_2 = (1,1,\dots,1,2),\ \dots,\ \omega_{6^k-1} = (6,6,\dots,6,5),\ \omega_{6^k} = (6,6,\dots,6,6).$$

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

$$\Omega = \biguplus \left[\{\omega\} : \omega \in \Omega\right] = \biguplus_{j=1}^{\infty} \{\omega_j\}.$$

is a union of a sequence of disjoint set, 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.

$$P(A) := \frac{|A|}{|\Omega|} = \frac{|A|}{6^k}. \tag{3.4}$$

Observe the following.

(1) $A \subseteq \Omega \ \Rightarrow\ 0 \leq |A| \leq |\Omega| = 6^k \ \Rightarrow\ 0 \leq P(A) \leq 1$.
(2) The empty set has size $|\emptyset| = 0$ and $|\Omega| = 6^k$ Thus, $P(\emptyset) = 0$ and $P(\Omega) = 1$.
(3) Assume that $A_1, A_2, \dots$ are disjoint subsets of Ω. Since Ω is finite, only finitely many A_j are not empty (THINK!),
(4) We rearrange the sequence such that the nonempty members will be $A_1, A_2, \dots, A_m$ for some suitable m.
(5) Then, $A = A_1 \uplus A_2 \uplus \cdots \uplus A_m$ is a finite union
and disjointness of the $A_j \ \Rightarrow\ |A| = |A_1| + |A_2| + \cdots + |A_m|$
(6) Thus, σ–additivity: $P(A) = |A|/6^k = \sum\limits_{j=1}^{m} (|A_j|/6^k) = \sum\limits_{j=1}^{m} P(A_j) = \sum\limits_{\text{all } j} P(A_j)$
Last equation: The omitted sets $A_{m+1}, A_{m+2}, \dots$ 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.

(b) One easily sees the generalization 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

$$P(A) := \frac{|A|}{|\Omega|} = \frac{|A|}{N}. \tag{3.5}$$

Then everything stated in **(1) – (6)** of **(a)** remains valid if we replace 6^k with N, and this shows that P is a probability measure on 2^Ω.

(c) The finiteness of Ω was crucial: 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, $P(\Omega) = \sum_j P(\omega_j) = p + p + \cdots \le 0$, but we require $P(\Omega) = 1$.

(2) Thus, $p > 0$. Thus, $P(\Omega) = \sum_j P(\omega_j) = p + p + \cdots = \infty$, but we require $P(\Omega) = 1$.

(d) We will see that the most important probability measures on the uncountable set $\mathbb{R}$ [12] satisfy $P(x) = 0$ for all $x \in \mathbb{R}$. That is no contradiction to σ–additivity and $P(\mathbb{R}) = 1$, since one cannot write the real numbers as a countable union $\mathbb{R} = x_1 \cup x_2 \cup \cdots$ Rather, P often is characterized by integrals $P([a,b]) = \int\limits_a^b \varphi(t)dt$. (This explains $P(x) = \int\limits_x^x \varphi(t)dt = 0$.) $\square$

Example 3.3. For the most general example of a countable probability space, let

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\} \quad \text{for some finite or infinite sequence } \big(\omega_j\big)_j .$$

Let us write I_Ω for the corresponding index set $\{1, 2, \dots\}$, so that $\Omega = \{\omega_j : j \in I_\Omega\}$.

(a) Assume that $P : 2^\Omega \to [0,1]$ is a probability measure. We abbreviate $p_j := P\big(\omega_j\big)$. Any nonempty subset A of Ω is of the form $A = \{\omega_{n_1}, \omega_{n_2}, \dots\}$ for a suitable, finite or infinite, subsequence of the ω_j. We write I_A for the corresponding set of indices $\{n_1, n_2, \dots\}$. With that notation, P satisfies

$$A = \biguplus_{j \in I_A} \{\omega_j\} \;\Rightarrow\; P(A) = \sum_{j \in I_A} P\big(\omega_{n_j}\big) = \sum_{j \in I_A} p_{n_j}, \qquad (\sigma\text{–additivity}) \tag{3.6}$$

$$\text{In particular,} \quad P(\Omega) = \sum_{j \in I_\Omega} P\big(\omega_j\big) = \sum_{j \in I_\Omega} p_j = 1, \quad (\text{full sequence } \omega_1, \omega_2, \dots) \tag{3.7}$$

$$0 \le p_j \le 1. \quad (\text{since } 0 \le P(B) \le 1 \text{ for all arguments } B) \tag{3.8}$$

(b) In reverse, associate with each ω_j a number p_j such that the sequence $\big(p_j\big)_j$ satisfies

$$0 \le p_j \le 1 \text{ for all } j, \qquad \text{and} \qquad \sum_j p_j = 1. \tag{3.9}$$

Recall that each nonempty subset U of Ω is of the form $U = \{\omega_{n_1}, \omega_{n_2}, \dots\}$ for a suitable, finite or infinite, index set $I_U = \{n_1, n_2, \dots\}$. Note that $k \in I_U$ then means that k must be either n_1 or n_2 or ... We use that notation to define the function

$$P : 2^\Omega \longrightarrow \mathbb{R}; \qquad U \mapsto \begin{cases} \sum_{k \in I_U} p_k & \text{for } U \neq \emptyset, \\ 0 & \text{for } U = \emptyset. \end{cases}$$

[12] the so-called distributions of continuous random variables

Clearly, $p_j \geq 0 \ \forall j$ implies $P(U) \geq 0$. Also $I_U \subseteq I_\Omega$ yields $P(U) \leq P(\Omega) = \sum\limits_{\text{all } j} p_j = 1$. Since $P(\emptyset)$ by definition, we have seen that P satisfies all properties of a probability measure, except for σ–additivity. That one follows from general calculus rules for sequences of series with nonnegative terms.

(c) We summarize **(a)** and **(b)** as follows.

If Ω is countable and nonempty, i.e., $\Omega = \{\omega_j : j \in I_\Omega\}$ for some finite or infinite index set $I_\Omega = \{1, 2, \dots\}$, then each probability measure P on 2^Ω determines a sequence of numbers $\big(p_j\big)_{j \in I_\Omega}$ such that $p_j \geq 0$ for all j and $\sum_j p_j = 1$, by means of the equations

$$p_j \ = \ P\big(\{\omega_j\}\big) \quad (j \in I_\Omega)\,.$$

In reverse, each sequence $\big(p_j\big)_{j \in I_\Omega}$ that satisfies $p_j \geq 0$ for all j and $\sum_j p_j = 1$, determines a probability measure on 2^Ω by means of

$$P(A) \ := \ \begin{cases} \sum\limits_{k \in I_A} p_k & \text{for } A \neq \emptyset, \\ 0 & \text{for } A = \emptyset. \end{cases}$$

Note that this yields, for $A = \{\omega_j\}$, that $P\big(\{\omega_j\}\big) \ = \ p_j$. $\square$

Remark 3.3. The probability spaces $(\Omega, \mathfrak{F}, P)$ we will be faced with are in one of the following categories:

- **(a)** Ω is countable, i.e., $\Omega = \{\omega_1, \omega_2, \dots\}$. for some suitable, finite or infinite sequence $(\omega_n)_n$. Then $P(A)$ is defined for all sets $A \in \Omega$ (and thus, $\mathfrak{F} = 2^\Omega$).
- **(b)** Ω is uncountable, but there is a finite or infinite sequence $\omega_n \in \Omega$ such that $\sum_n P(\omega_n) = 1$. In other words, P is concentrated on the countable set $U := \{\omega_1, \omega_2, \dots\}$ in the sense that $P(U) = 1$ and thus $P(U^\complement) = 0$. again, $P(A)$ is defined for all sets $A \in \Omega$ (and thus, $\mathfrak{F} = 2^\Omega$).
- **(c)** $\Omega = \mathbb{R}$ and $P(A)$ is known (at a minimum) for intervals such as $[a, b]$ or $]a, b]$ or $[a, b[$ or $]a, b[$.
- **(d)** $\Omega = \mathbb{R}^n$ 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.

- **(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
$$P\big(]a, b]\big) \ = \ \int_a^b f(u)\, du\,.$$
This makes it 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

$$P\big( ]a_1, b_1] \times]a_2, b_2] \times \cdots \times]a_n, b_n]\big) = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_1}^{b_1} f(\vec{u})\, d\vec{u}$$
$$= \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_1}^{b_1} f(u_1, \dots, u_n)\, du_1\, du_2 \cdots du_{n-1}\, du_n\,.$$

Thus, the σ–algebra $\mathfrak{B}^n$ for such P should contain all rectangles $]a_1, b_1] \times]a_2, b_2] \times \cdots \times]a_n, b_n]$

You may have Nnticed 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 $\int_a^a \dots da$ is always zero. Nevertheless, 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 $\mathbb{R}^n$. □

Note that the next definition is marked as optional.

Definition 3.3. $\boxed{\star}$ One can show that that there are such things as

- the smallest σ–algebra of subsets of $\mathbb{R}$ which contains all intervals of real numbers. It is denoted $\mathfrak{B}$.
- the smallest σ–algebra of subsets of $\mathbb{R}^n$ which contains all n–dimensional rectangles. It is denoted $\mathfrak{B}^n$.

We call $\mathfrak{B}$ and $\mathfrak{B}^n$ the Borel σ–algebras of $\mathbb{R}$ and of $\mathbb{R}^n$ **Borel σ–algebra** and we call their members Borel sets. **Borel set**

It is sufficient for this course that you just remember that

- The Borel sets are the sufficiently well behaved sets of $\mathbb{R}$ and $\mathbb{R}^n$
- The intervals and n–dimensional rectangles are amon those sets.
- Only completely weird and useless sets are not Borel. □

Remark 3.4 (σ–algebras will be ignored)**.** Consider this a continuation of Remark 3.3. We can summarize it as follows.

There are only two kinds of probability spaces $(\Omega, \mathfrak{F}, P)$.

- There is a countable subset $C = \{\omega_1, \omega_2, \dots\}$ of Ω such that $\sum\limits_{\omega \in C} P(\{\omega\}) = 1$. Then $\boxed{\mathfrak{F} = 2^\Omega}$, since the above allows us to define $P(A)$ for arbitrary $A \subseteq \Omega$ as

$$P(A) = \sum_{\omega \in C \cap A} P(\{\omega\})\,.$$

- $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$. Then $\boxed{\mathfrak{F} = \text{the Borel sets.}}$

Now that we understand the structure of the domain $\mathfrak{F}$ of the probability measures P we will be dealing with, there is no more need to keep carrying this baggage with us.

Henceforth, we will, with very few exceptions, do the following.

> We will ignore that probability measures cannot always be given on the entire power set 2^Ω (true only we deal with $(\mathbb{R}, \mathfrak{B}, P)$ or $(\mathbb{R}^n, \mathfrak{B}^n, P)$) and that this necessitated us to introduce a σ–algebra $\mathfrak{F}$ as the domain of that probability measure. Accordingly, we will ignore the σ–algebra and talk about
>
> $$\text{probability spaces } (\Omega, P), \qquad \text{rather than } (\Omega, \mathfrak{F}, P).$$

$\square$

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 $P\{\omega\}$.

Remark 3.5. The following is preliminary and will be expanded.

MF terminology	**WMS terminology**
• σ–algebra $\mathfrak{F}$	concept DNE $\square$
• probability space $(\Omega, \mathfrak{F}, P)$ with "underlying set" Ω, σ–algebra $\mathfrak{F}$, probability measure $P(A)$ defined for $A \in \mathfrak{F}$	sample space S with a probability $P(A)$ defined for all sets of interest $A \subseteq S$
• xx	yy $\square$

Remark 3.6. This remark is preliminary.

(A) 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 (can have big k like $k = 25$. For such a high precision we can model the potential outcomes Ω as the continuum $[0, 1[$.

(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: $|\Omega| = 2^{10^9} = 2^{10 \cdot 10^8} \;=\; \big(2^{10}\big)^{10^8} \approx \big(10^3\big)^{10^8} \;=\; 10^{3 \cdot 10^8}$

(6) A selection of n items from a population is a sample of size n.

(B) A supreme being decides to pick "this" ω. This pick seems random to us since we do not know what choice this being will make. $\square$

3.2 Preimages and Indicator Functions

Introduction 3.1. The major part of this course will be about functions $\omega \mapsto f(\omega)$ which assign the outcomes (= elements) ω of a probability space to items $f(\omega)$ which are usually numbers or vectors of numbers. In other words, the codomain will usually be (a subset of) $\mathbb{R}$ or $\mathbb{R}^n$. We illustrate this with the following example.

Let the probability space (Ω, P) [13] represent the outcomes of two rolls of a fair die:

- $\Omega = \{1, 2, \dots, 6\}^2$. Interpret $\omega = (\omega_1, \omega_2)$ as die$_1$ yields ω_1, die$_2$ yields ω_2. [14]
  $\boxdot$ Thus, $\omega = (5, 2)$ represents the outcome of die$_1$ giving a 5 and die$_2$ giving a 2.
- Probability measure P is determined by $P(\omega_1, \omega_2) = 1/|\Omega| = 1/36$. See Example 3.3 on p.38.

Consider the function which associates with each outcome (ω_1, ω_2) the sum of the throws, i.e.,

- $Y : \Omega \to \{2, 3, 4, \dots, 11, 12\}; \quad (\omega_1, \omega_2) \mapsto Y\big((\omega_1, \omega_2)\big) := \omega_1 + \omega_2.$
  Get used to the notation! WMS loves to use the letters (X, Y, Z) for function names.

We will create a probability measure P' on $\Omega' := \{2, 3, 4, \dots, 11, 12\}$, the codomain of the function Y.

- Since Ω' is countable, it suffices to specify $P'(\{2\}), P'(\{3\}), \dots, P'(\{12\})$. (Again, Example 3.3.)
- Define $P'(\{10\}) := P\big(\{(\omega_1, \omega_2) \in \Omega : Y(\omega_1, \omega_2) = 10\}\big) = P\big(\{(4, 6), (5, 5), (6, 4)\}\big) = 1/12$. This is the probability that the sum of the throws is 10!
- In general, for $\omega' \in \Omega'$, define $P'(\{(\omega')\}) := P\big(\{(\omega_1, \omega_2) \in \Omega : Y(\omega_1, \omega_2) = \omega'\}\big)$. This is the probability that the sum of the throws is ω'!
- One can show quite easily [15] that, if $B \subseteq \Omega'$, then

  $$B \subseteq \Omega' \;\Rightarrow\; P'(B) = P\big(\{\omega \in \Omega : Y(\omega) \in B\}\big). \quad \text{(We wrote } \omega \text{ for } (\omega_1, \omega_2).) \tag{3.10}$$

  This is the probability that the sum of the throws is in B!
- We have created a probability measure $P'(B)$ on the codomain of Y by assigning P, the original probability on the domain Ω, to the set

  $$\{\, \omega \in \Omega \,:\, Y(\omega) \in B \,\}$$

  of all those arguments $\omega \in \Omega$ which are mapped by Y into B.

That makes those sets so important that they warrant their own definition. $\square$

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

Definition 3.4.

[13] As promised, no more σ–algebra unless absolutely necessary!
[14] We often prefer to write ω rather than $\vec{\omega}$ if the the symbol Ω is involved, even if the elements are vectors.
[15] with the help of Proposition 3.4 (f^{-1} is compatible with all basic set ops) further down, on p.45

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

$$f^{-1}(B) := \{x \in X : f(x) \in B\} \tag{3.11}$$

is a subset of X which we call the **preimage** of B under f. $\square$

Remark 3.7. $\boxed{\star}$

- 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 $f^{-1}(B) \in 2^X$).
- The symbol f^{-1} is the same as that for the ordinary inverse function $f^{-1}(y) = x$, **if this inverse function exists!**
- $f^{-1}(B)$ exists for any choice of $X, Y, f : X \to Y$, and $B \subseteq Y$, even if the inverse function does not exist!

As an example, let

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

If there was an inverse function, then 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?

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 $g(x) = y$?

For an even more extreme example, consider

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

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 $h(x) = y$? $\square$

Notational conveniences 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 3.4 (Preimages). Let $f:\mathbb{R}\to\mathbb{R};\quad f(x)=x^2$.

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. $f^{-1}(]-4,-2[\cup[1,2]\,\cup\,[5,6]) \;=\; \{\, x\in\mathbb{R} : x^2\in\,]-4,-2[$ 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 3.5 (Preimages). Let $f:\mathbb{R}\to\mathbb{R};\quad f(x)=x^2$.

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

Proposition 3.3. *Some simple properties:*

(3.12) $\quad f^{-1}(\emptyset) \;=\; \emptyset$
(3.13) $\quad B_1 \;\subseteq\; B_2 \;\subseteq\; Y \Rightarrow f^{-1}(B_1) \;\subseteq\; f^{-1}(B_2)$ $\quad$ (***monotonicity of*** $f^{-1}\{\dots\}$)
(3.14) $\quad f^{-1}(Y) \;=\; X$ $\quad$ *always!*

PROOF of 3.13:
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 3.12 and 3.13 is left as an exercise. ■

Remark 3.8 (Notational conveniences II:).

In probability theory the following notation is also very common:
$\{f\in B\} := f^{-1}(B),\ \{f=y\} := f^{-1}\{y\}$.
Let $\mathscr{N}$ be either of $\mathbb{Z},\mathbb{Q},\mathbb{R}$. Let $a,b\in\mathscr{N}$ such that $a<b$. We write $\{a\le f\le b\} := f^{-1}([a,b]_{\mathscr{N}}),\ \{a<f<b\} := f^{-1}(]a,b[_{\mathscr{N}})$,
$\{a\le f<b\} := f^{-1}([a,b[_{\mathscr{N}}),\ \{a<f\le b\} := f^{-1}(]a,b]_{\mathscr{N}}),\ \{f\le b\} := f^{-1}(]-\infty,b]_{\mathscr{N}})$, etc.
□

Example 3.6. In the introduction we were examining

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

It is very important that you **remember the first three** of the five formulas of the next proposition.

Proposition 3.4 (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.* [16] *Further assume that $B \subseteq Y$ and that $B_j \subseteq Y$ for all j. Then*

$$f^{-1}\big(\bigcap_{j \in J} B_j\big) = \bigcap_{j \in J} f^{-1}(B_j) \tag{3.15}$$

$$f^{-1}\big(\bigcup_{j \in J} B_j\big) = \bigcup_{j \in J} f^{-1}(B_j) \tag{3.16}$$

$$f^{-1}(B^{\complement}) = \big(f^{-1}(B)\big)^{\complement} \tag{3.17}$$

$$B_1 \cap B_2 = \emptyset \Rightarrow f^{-1}(B_1) \cap f^{-1}(B_2) = \emptyset. \tag{3.18}$$

$$f^{-1}(B_1 \setminus B_2) = f^{-1}(B_1) \setminus f^{-1}(B_2) \tag{3.19}$$

$$f^{-1}(B_1 \Delta B_2) = f^{-1}(B_1) \Delta f^{-1}(B_2) \tag{3.20}$$

Note that (3.18) implies that the preimages of a disjoint family form a disjoint family.

PROOF: $\boxed{\star}$ MF330 notes, ch.8 ■

Proposition 3.5 (Preimages of function composition)**.** *Let X, Y, Z be arbitrary, nonempty 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

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}, \text{ i.e., } (g \circ f)^{-1}(W) = f^{-1}\big(g^{-1}(W)\big) \text{ for all } W \subseteq Z. \tag{3.21}$$

PROOF: $\boxed{\star}$ MF330 notes, ch.8 ■

Try to understand the sbove 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).

3.3 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 3.5 (indicator function for a set)**.** Let Ω be a nonempty set and $A \subseteq \Omega$. Let $1_A : \Omega \to \{0, 1\}$ be the function defined as

$$1_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{3.22}$$

[16] 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

1_A is called the **indicator function** of the set A. [17] $\square$

Example 3.7. The so-called density function for the exponential distribution with parameter $\beta > 0$ is

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

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

Proposition 3.6. *Let A, B, C be subsets of Ω. Then*

(3.23) $$A \subseteq B \;\Rightarrow\; 1_A \le 1_B,$$
(3.24) $$1_{A\cup B} \;=\; \max(1_A, 1_B),$$
(3.25) $$1_{A\cap B} \;=\; \min(1_A, 1_B),$$
(3.26) $$1_{A^\complement} \;=\; 1 - 1_A,$$
(3.27) $$1_{A \uplus B} \;=\; 1_A + 1_B \quad (A, B \textit{ disjoint})$$

PROOF: The proof is an easy exercise.

3.4 Random Variables and their Probability Distributions

Introduction 3.2. We continue with an observation we made in the introduction 3.1 to Section 3.2 (Preimages and Indicator Functions, p.42). There,

- $\Omega \;=\; \{1, 2, \dots, 6\}^2$ 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 $Y : \Omega \to \Omega' := \{2, 3, 4, \dots, 11, 12\};\; \vec{\omega} \mapsto Y\big(\vec{\omega}\big) := \omega_1 + \omega_2$, 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 (3.10):
$$B \subseteq \Omega' \;\Rightarrow\; P'(B) \;=\; P\{\vec{\omega} \in \Omega : Y\big(\vec{\omega}\big) \in B\}\,.$$
Observe that the set Ω' has been transformed into a probability space, (Ω', P')).
- With preimage notation and the notational shortcuts of Remark 3.8 on p.44, 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 is not entirely true, but it will be true for the situations we will deal with in this class. If you are curious, read this optional footnote. [18] $\square$

[17] In abstract algebra this is often called the **characteristic function** of A. Some authors write χ_A or $\mathbb{1}_A$ instead of 1_A.

[18] $\boxed{\star}$ We have to recall that there really is a σ–algebra $\mathfrak{F}$ on Ω and that $P(A)$ only exists if $A \in \mathfrak{F}$. What if $B \subseteq \Omega'$ does not have a nice preimage, i.e., $\{Y \in B\} \notin \mathfrak{F}$? The only way out is not to allow arbitrary $B \in 2^{\Omega'}$, but **(a)** to also require a σ–algebra $\mathfrak{F}'$ on the codomain Ω', which **(b)** is so "small" that $B \in \mathfrak{F}' \Rightarrow Y^{-1}(B) \in \mathfrak{F}$; or, if you prefer, $\mathfrak{F}'$ must be so "big" that $B \in \mathfrak{F}' \Rightarrow Y^{-1}(B) \in \mathfrak{F}$. There is a name for triplets $[Y, \mathfrak{F}, \mathfrak{F}']$ which satisfy this reltionship. The function Y is called **measurable** with respect to $\mathfrak{F}$ and $\mathfrak{F}'$ or $(\mathfrak{F}, \mathfrak{F}')$–**measurable** None of this will be an issue in this course!

The next theorem and the subsequent definitions are very important.

Theorem 3.1.

Let (Ω, P) be a probability space, Ω' a nonempty set, and $Y : \Omega \to \Omega'$ a function. Then the formula

$$P_Y(B) := P\{Y \in B\} \quad (B \subseteq \Omega') \tag{3.28}$$

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 (3.17) 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 (3.16) to the index set $\mathbb{N}$ of a sequence of disjoint subsets $B_1, B_2, \dots$ of Ω'. Let $B := B_1 \biguplus B_2 \biguplus B_3 \biguplus \cdot$. Then

$$P_Y(B) = P\Big(Y^{-1}\Big(\biguplus_{j\in\mathbb{N}} B_j\Big)\Big) = P\Big(\bigcup_{j\in\mathbb{N}} Y^{-1}(B_j)\Big)$$

By (3.18), the sets $Y^{-1}(B_j)$ are disjoint. Thus,

$$P_Y(B) = P\Big(\biguplus_{j\in\mathbb{N}} Y^{-1}(B_j)\Big) = \sum_{j\in\mathbb{N}} P\big(Y^{-1}(B_j)\big) = \sum_{j\in\mathbb{N}} P_Y(B_j) .$$

This proves σ–additivity. ■

Definition 3.6 (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' which is given by

$$P_Y(B) := P\{Y \in B\} \quad (B \subseteq \Omega') \tag{3.29}$$

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 3.7 (Random Variables and Random Vectors)**.** Let (Ω, P) be a probability space and let $n \in \mathbb{N}$.

Let $U \subseteq \mathbb{R}$. A function
$$Y : \Omega \longrightarrow U ; \quad \omega \mapsto Y(\omega)$$
is called a **random variable** on $(\Omega, \mathfrak{F}, P)$. Let $V \subseteq \mathbb{R}^n$. A function
$$\vec{X} = \big(X_1, X_2, \dots, X_n\big) : \Omega \longrightarrow V ; \quad \omega \mapsto \vec{X}(\omega) = \big(X_1\omega), \dots, X_n(\omega)\big)$$
is called a **random vector** on $(\Omega, \mathfrak{F}, P)$.
If there is a countable subset $U^* = \{y_1, y_2, \dots\}$ of U such that $\sum_j P_Y\{y_j\} = 1$ (i.e., $P\{Y \notin U^*\} = 0$), we call Y a **discrete random variable**. Likewise, if there is a countable subset V^* of V such that $P\{\vec{X} \notin V^*\} = 0$, we call $\vec{X}$ a **discrete random vector**. $\square$

Note that random variables and vectors which have a countable range are discrete. Also, if you found the footnote at the end of the introduction interesting, have a look at this (optional) one, [19]

Remark 3.9. In many instances the exact nature of the codomain U of a random variable Y is unimportant. Of course it must be a set of numbers, i.e., $U \subseteq \mathbb{R}$, and it must be big enough to accommodate all function values $Y(\omega)$, i.e., $Y(\omega) \subseteq U$. [20] 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. $\square$

Not all interesting functions on a probability space take values in $\mathbb{R}$ or $\mathbb{R}^n$. Here is an example.

Example 3.8. 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 $0.d_1 d_2 d_3$ with inifinitely many decimal digits. If necessary, we append infinitely many zeroes to the right. For example, we write $0,25000\dots$ 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. $\square$

Considering that last example, it seems awkward $\boxed{\text{not}}$ 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.

The next definition is non–standard and you will not be quizzed on it. Note though that I will use the term "'random item' in these lecture notes and in my lectures,

[19] $\boxed{\star}$ Technically speaking, Y must be $(\mathfrak{F}, \mathfrak{B})$–**measurable** and $\vec{X}$ must be $(\mathfrak{F}, \mathfrak{B}^n)$–**measurable**. In other words, you must be able to assign probabilities to all preimages of Borel sets. Again, none of this will be an issue in this course!

[20] 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 $y \in U$ and that requires that U is minimal: $U = Y(\Omega)$. The same thought also applies to random vectors.

Definition 3.8 (Random item). $\boxed{\star}$ Let $(\Omega, \mathfrak{F}, P)$ be a probability space and Ω' a nonempty set. We call a function $X : \Omega \to \Omega'$ a **random item** on Ω. $\square$

Remark 3.10. We can phrase Theorem 3.1 and the subsequent Definition 3.6 as follows.
All random items X on a probability space $(\Omega, \mathfrak{F}, P)$ have a distribution

$$P_X(B) \ = \ P\{X \in B\} \ = \ P\big(X^{-1}(B)\big) \ \ (B \subseteq \Omega'). \ \square$$

Remark 3.11. 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 $\mathscr{S}$, decides to pick "this" particular $\omega_0 \in \Omega$. As a result, all random items $X, Y, Z, \dots$ are invoked with ω_0 as argument, resulting in the outcomes $X(\omega_0), Y(\omega_0), Z(\omega_0), \dots$. With this interpretation it makes a lot of sense to talk about functions on (Ω, P) as **random** items since, when we interpret $\omega \in \Omega$ as "randomness",

$$x \ = \ X(\omega) \ \text{ simply means that } \ x \text{ is a function of randomness.}$$

Only $\mathscr{S}$ knows what ω_0 will be picked. But if we know, say, the distribution P_X of a random variable X, then we can at least quantify the likelihood that $\mathscr{S}$ chose an ω such that $17.8 \leq X(\omega) \leq 21.3$ It will be $P_X([17.8, 21.3]) = P\{17.8 \leq X \leq 21.3\}$. $\square$

Often it only is the distribution of a random item with values in a set Ω' that matters and there may be many different choices of probability space plus random item which result in that same 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 3.8. We fix $n = 3$ since this example illustrates all essential points.

Example 3.9. (a) Let $\Omega_1 := \{0, 1\}^3$ with the probability measure $P\{(a, b, c)\} = 1/|\Omega_1| = 1/8$.
Let $Y_1 : \Omega_1 \to \{H, T\}^3$ the random item that changes each H into a 1 and each T into a 0. For example, $Y_1(1, 0, 1) = (H, T, H)$ and $Y_1(0, 0, 1) = (T, T, H)$.
Then P_{Y_1} is the same probability measure as $P_{\vec{X}}$ of Example 3.8, 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 $Y_2 : \Omega_2 \to \{H, T\}^3$ 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., $Y_2(\omega) = \omega$ for all $\omega \in \Omega_2$.
Clearly, P_{Y_2} also assigns probability $P_{Y_2}(\{\omega\}) = 1/8$ to each element of $\{H, T\}^3$.
(c) Let $\Omega_3 := \{H, T\}^3 \times \{1, 2, 3, 4\}$ with the probability measure $P\{(a, b, c, d)\} = 1/|\Omega_3| = 1/32$. (Same as in **(a)**, except that now a, b, c represent either of H or T rather than 0 or 1.)
Let $Y_3 : \Omega_3 \to \{H, T\}^3$ be the function defined as $Y_3(a, b, c, d) := (a, b, c)$. We compute the distribution P_{Y_3} for the outcomes (a, b, c) of the probability space $(\{H, T\}^3, P_{Y_3})$.

$$\begin{aligned}(a, b, c) \in Y_3 \ \Rightarrow P_{Y_4}\{(a, b, c, d)\} \ &= \ P\{Y_4 = (a, b, c, d)\} \\ &= P\{(a, b, c, 1), (a, b, c, 2), (a, b, c, 3), (a, b, c, 4)\} \ = \ 4(1/32) \ = \ 1/8\,.\end{aligned}$$

We have have obtained in this example and Example 3.9 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 items $\vec{X}, Y_1, Y_2, Y_3$ which were defined on four different probability spaces. Thus, you have multiple choices of probability spaces and random itens to model a distribution. you will hopefully agree that Y_1 and Y_2 are much better choices than $\vec{X}$ and Y_3. $\square$

3.5 Conditional Probability and Independent Events

A
@@Author

This section should be moved directly after Section 3.1 (Probability Spaces).

Definition 3.9 (Conditional probability)**.**

Given are a probability space $(\Omega, \mathscr{F}, P)$ and two events $A, B \in \mathscr{F}$. We call

$$P(A \mid B) \ := \begin{cases} \dfrac{P(A \cap B)}{P(B)}\,, & \text{if } P(B) > 0\,, \\ \textbf{undefined}\,, & \text{if } P(B) = 0\,, \end{cases} \tag{3.30}$$

(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. $\square$

Theorem 3.2.

Given are a probability space $(\Omega, \mathscr{F}, P)$ and an event $B \in \mathscr{F}$ such that $P(B) > 0$. Then

$$P(\cdot \mid B) : \mathfrak{F} \longrightarrow [0,1]\,; \qquad A \mapsto P(A \mid B) \tag{3.31}$$

is another probability measure on $(\Omega, \mathscr{F})$.

In other words, $P(\cdot \mid B)$ satisfies (3.1) – (3.3) of Definition 3.2 (Probability measures and probability spaces) on p.35.

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 (3.1): Since $P(\emptyset \cap B) = 0, \quad P(\emptyset \mid B) = 0/P(B) = 0$.
PROOF of (3.2): Since $\Omega \cap B = B, \quad P(\Omega \mid B) = P(\Omega \cap B)/P(B) = P(B)/P(B) = 1$.
PROOF of (3.3): 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 (2.23) on p.27,

$$\biguplus_{n \in \mathbb{N}} (B \cap A_n) \ = \ B \cap \biguplus_{n \in \mathbb{N}} A_n\,.$$

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 (3.3). ■

Proposition 3.7. *If $(\Omega, \mathfrak{F}, P)$ is a probability space and $A, B, C \in \mathfrak{F}$, then*

(3.32) $$P(A\cap B\cap C) \;=\; P(A \mid B\cap C)\cdot P(B \mid C)\cdot P(C)\,.$$

PROOF:

$$P(A\cap B\cap C) \;=\; P(A \mid B\cap C)\cdot P(B\cap C)\,. \;=\; P(A \mid B\cap C)\cdot P(B \mid C)\cdot P(C)\,. \blacksquare$$

This generalizes to arbitrarily many sets as follows.

Proposition 3.8 (Iterative conditioning formula)**.**

If $(\Omega, \mathfrak{F}, P)$ is a probability space, $n \in \mathbb{N}$ and $A_1, \dots, A_n \in \mathfrak{F}$, then

(3.33) $$\begin{aligned} P(A_1\cap A_2\cap\dots\cap A_n) &= P(A_1 \mid A_2\cap\dots\cap A_n)\cdot P(A_2 \mid A_3\cdots\cap A_n)\cdots \\ &\qquad \cdots P(A_{n-2} \mid A_{n-1}\cap A_n)\, P(A_{n-1} \mid A_1)\, P(A_n). \end{aligned}$$

PROOF:

It is easier to work with the reverse sequence $A_n\cap A_{n-1}\cap\dots\cap A_1$ instead of $A_1\cap A_2\cap\dots\cap A_n$. Repeated use of $P(U\cap V) = P(U \mid V)P(V)$ with $U = A_j$ and $V = A_{j-1}\cap\dots\cap A_1$ yields

$$\begin{aligned} &P(A_n\cap A_{n-1}\cap\dots\cap A_1) \\ &= P(A_n \mid A_{n-1}\cap\dots\cap A_1)\, P(A_{n-1}\cap\dots\cap A_1) \\ &= P(A_n \mid A_{n-1}\cap\dots\cap A_1)\, P(A_{n-1} \mid A_{n-2}\cdots\cap A_1)\, P(A_{n-2}\cdots\cap A_1) \\ &= \dots\dots\dots\dots\dots \\ &= P(A_n \mid A_{n-1}\cap\dots\cap A_1)\, P(A_{n-1} \mid A_{n-2}\cdots\cap A_1)\cdots P(A_3 \mid A_2\cap A_1)\, P(A_2 \mid A_1)\, P(A_1). \blacksquare \end{aligned}$$

Definition 3.10 (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

(3.34) $$P(A\cap B) \;=\; P(A)\cdot P(B)\,. \;\square$$

Independence of three events is not defined as you may have guessed from that last definition.

Definition 3.11 (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

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B) \cdot P(C)\,, \\ P(A \cap B) &= P(A) \cdot P(B)\,, \\ P(A \cap C) &= P(A) \cdot P(C)\,, \\ P(B \cap C) &= P(B) \cdot P(C)\,. \;\square \end{aligned} \tag{3.35}$$

We can state (3.35) as follows. It must be true for any subsequence of events that the probability of the intersection equals the product of the probabilities of the individual events.

Remark 3.12. 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 3.11 shows us how to generalize independence to any number of events.

Definition 3.12 (Finitely many independent events).

Given are a probability space $(\Omega, \mathscr{F}, P)$, $n \in \mathbb{N}$ and events $A_1, A_2, \dots, A_n \in \mathscr{F}$. We say that $A_1, A_2, \dots, A_n$ are **independent** if, for ANY subselection of indices

$$1 \leq j_1 < j_2 < \cdots < j_k \leq n\,,$$

it is true that

$$P(A_{j_1} \cap A_{j_1} \cap A_{j_k}) = P(A_{j_1}) \cdot P(A_{j_2}) \cdot P(A_{j_k})\,. \;\square \tag{3.36}$$

Finally, we define independence for infinitely many events.

Definition 3.13 (Independent events – the general case).

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a sequence of events $A_1, A_2, \dots \in \mathscr{F}$ We say that this sequence is **independent** if, for ANY FINITE subselection of distinct indices $j_1, j_2, \dots, j_k \in \mathbb{N}$, it is true that

$$P(A_{j_1} \cap A_{j_1} \cap A_{j_k}) = P(A_{j_1}) \cdot P(A_{j_2}) \cdot P(A_{j_k})\,. \;\square \tag{3.37}$$

Remark 3.13. Note that the number k in Definition 3.12 and Definition 3.13 is not fixed. $\square$

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 2.20 on p.24) of events

Definition 3.14 (Independence of uncountably many events). $\boxed{\star}$

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\left(A_i\right)_{i \in I}$ 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 $i_1, i_2, \dots, i_k \in I$, it is true that

$$(3.38) \qquad P(A_{i_1} \cap A_{i_1} \cap A_{i_k}) \ = \ P(A_{i_1}) \cdot P(A_{i_2}) \cdot P(A_{i_k}) \,. \ \square$$

Next, we examine connections between conditional probabilities and independence.

Theorem 3.3.

Given are a probability space $(\Omega, \mathscr{F}, P)$ and two events $A, B \in \mathscr{F}$ such that $P(B) > 0$. Then

$$(3.39) \qquad A \text{ and } B \text{ are independent} \quad \Leftrightarrow \quad P(A \mid B) \ = \ P(A) \,.$$

PROOF of "⇒":
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 "⇐":
Since $P(A \mid B) = P(A)$ and $P(B) > 0$,

$$P(A) \cdot P(B) \ = P(A \mid B) \cdot P(B) \ = \ \frac{P(A \cap B)}{P(B)} \cdot P(B) \ = \ P(A \cap B) \,. \ \blacksquare$$

Corollary 3.1.

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

$$(3.40) \qquad A \text{ and } B \text{ are independent} \quad \Leftrightarrow \quad P(A \mid B) \ = \ P(A) \quad \Leftrightarrow \quad P(B \mid A) \ = \ P(B) \,.$$

PROOF: Obious ■

The next theorem is marked optional, but it is just as easy to remember as the corollary that follows it.

@@Author

Theorem 3.4 and Corollary 3.2 below belong directly after Definition 3.14 (Independence of uncountably many events), but they were moved here so the numbering of existing theorem, formulas, ... is not affected.

Theorem 3.4. ⋆

Given are a probability space $(\Omega, \mathscr{F}, P)$ and a family $\left(A_i\right)_{i \in I}$ 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 $\left(A_i\right)_{i \in I}$ implies that of $\left(B_i\right)_{i \in I}$.

PROOF: Utilizes advanced probabilistic methods that are outside the scope of this course ■

Note that the following corollary is NOT marked as optional!

Corollary 3.2.

Given are a $(\Omega, \mathfrak{F}, P)$ is a probability space, $n \in \mathbb{N}$ and independent events $A_1, \dots, A_n \in \mathfrak{F}$. 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 $i = 1, 2, \dots, n$, let B_i be either A_i or $A_i^{\complement}$. Then independence of $A_1, \dots, A_n$ implies that of $B_1, \dots, B_n$.

PROOF: $\boxed{\star}$

(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 $A_1, \dots, A_n$ be independent. For convenience, let $B := A_1 \cap \dots \cap A_{n-1}$.

Since $P(B \cap A_n) = P(A_1 \cap \dots \cap A_n) = P(A_1) \cdots P(A_n) = P(B) \cdot P(A_n)$, 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 $B_{j_1} := A_{j_0}^{\complement}$, that $B_1, \dots, B_n$ 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 $C_1, \dots, C_n$ 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 $D_1, \dots, D_n$ 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. ■

4 Combinatorial Analysis

In many important cases we find ourselves in the situation of Example 3.2**(a)** on p.37, 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.

4.1 The Multiplication Rule

The first result is known under names such as the basic principle of counting ([3] Ross, Sheldon M.: A First Course in Probability, 3rd edition) and the mn rule (WMS text).

Theorem 4.1 (Multiplication rule)**.**

(A) *Assume that two actions A and B are performed such that*

- *the first one has m outcomes, $\{a_1, a_2, \dots, a_m\}$,*
- *the second one has n outcomes $\{b_1, b_2, \dots, b_n\}$ 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 $A_1, \dots, A_k$ are performed such that*

- *action A_1 has n_1 outcomes, $\{a_1^{(1)}, a_2^{(1)}, \dots, a_{n_1}^{(1)}\}$,*
- *action A_2 has n_2 outcomes, $\{a_1^{(2)}, a_2^{(2)}, \dots, a_{n_2}^{(2)}\}$ for each outcome of A_1,*
- *action A_3 has n_3 outcomes, $\{a_1^{(3)}, a_2^{(3)}, \dots, a_{n_3}^{(3)}\}$ 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, $\{a_1^{(k)}, a_2^{(k)}, \dots, a_{n_k}^{(k)}\}$ 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 $a_1^{(j)}, \dots, a_{n_j}^{(j)}$.*
- *Then there are $n_1 \cdot n_2 \cdots n_k$ combined outcomes $(x_1, x_2, \dots, x_k)$.*
  Here, each x_j is one of the n_j outcomes $a_1^{(j)}, \dots, a_{n_j}^{(j)}$ of A_j.

PROOF: We identify the actions with their outcomes, i.e., we define

$$A_j \;=\; \{a_1^{(j)}, \dots, a_{n_j}^{(j)}\}, \quad \text{for } \; j = 1, 2, \dots, k.$$

Now, the multiplication rule merely states that $\left|\, A_1 \times A_2 \times \cdots \times A_n \,\right| \;=\; |A_1| \cdot |A_2| \,\cdots\, |A_n|$, and this is true according to (2.28) on p.29. ■

Example 4.1 (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: $26^3 \cdot 10^4 = 175,760,000$ $\square$

Example 4.2 (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: $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$ $\square$

Example 4.3. How many 7–digit license plates can be created if the first three are letters (CAPS) and the lst four are digits and none of the letters can be repeated?

Answer: $26 \cdot 25 \cdot 24 \cdot 10^4 \;=\; 26 \cdot 600 \cdot 10^4 \;=\; 15,600 \cdot 10^4 \;=\; 15,600,000.$ $\square$

Example 4.4 (Ross-prob-thy-3ed Example 2d)**.** If $|\Omega| = n$, 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 $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, then

- we have 2 choices for the $\psi(\omega_1)$ selection.
- For each of those there are 2 choices for the $\psi(\omega_2)$ selection.
- For each of those $\psi(\omega_1), \psi(\omega_2)$ selections there are 2 choices for the $\psi(\omega_3)$ selection.
- -
- For each of those $\psi(\omega_1), \dots, \psi(\omega_{n-1})$ selections there are 2 choices for the $\psi(\omega_n)$ selection.

So we have $2 \cdot 2 \cdots 2 = 2^n$ selections. $\square$

Example 4.5. If $|\Omega| = n$, how many subsets of Ω, including $\emptyset$ and Ω, do exist?

Answer: If $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ and $A \subseteq \Omega$, then

- we have 2 choices: either $\omega_1 \in A$ or $\omega_1 \notin A$.
- For each of those, either $\omega_2 \in A$ or $\omega_2 \notin A$.
- -
- For each of those $n-1$ choices $\omega_j \in A$ or $\omega_j \notin A$ (j = 1, 2, ..., n-1), either $\omega_n \in A$ or $\omega_n \notin A$.

So we have $2 \cdot 2 \cdots 2 = 2^n$ choices. $\square$

4.2 Permutations

Definition 4.1 (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^n_r. $\square$

Theorem 4.2 (WMS Ch.02.6, Theorem 2.2)**.**

(4.1)
$$P_r^n \;=\; n(n-1)(n-2)\cdot(n-r+1) \;=\; \frac{n!}{(n-r)!}\,.$$

Here, $n!$ ("n factorial") is defined as follows.

(4.2)
$$n! \;=\; \begin{cases} n(n-1)\cdots 2\cdot 1\,, & \text{if } n\in\mathbb{N}\,,\\ 1\,, & \text{if } n=0\,.\end{cases}$$

PROOF: We can consider each permutation as the result of the following actions $A_1,\dots,A_r$.
- 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. ■

Example 4.6. 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 consider 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. □

4.3 Combinations, Binomial and Multinomial Coefficients

A simple application of the multiplication rule showed us that for a set Ω of finite size, its powerset 2^Ω has size $|\Omega| = 2^{|\Omega|}$. (See example 4.5 on p.56.)

A related question would be how many elements of 2^Ω have a given size k, i.e., how many subsets of Ω have size k?

Examining how many permutations of size k can be obtained from the elements $\omega_1, \omega_2, \dots, \omega_n$ 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 $\omega_{n_1}, \omega_{n_2}, \dots, \omega_{n_k}$ of such a subset lead to different permutations although the subset remains the same.

Thus, we must divide P^n_k, 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, we have obtained the following result.

Theorem 4.3.

Let $0 \leq k \leq n$. A set of size n has

$$\frac{n!}{k!(n-k)!}\,.$$

of size k.

PROOF: We saw in the discussion before the theorem that the number we are looking for is P^n_k / P^k_k. But

$$\frac{P^n_k}{P^k_k} = \frac{n(n-1)\cdots(n+k-1)}{k!} = \frac{n(n-1)\cdots\big(n-(k-1)\big)}{k!} \cdot \frac{(n-k)!}{(n-k)!} = \frac{n!}{k!(n-k)!}\,.$$

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 $n \geq k$,) 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 4.2 (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 $\binom{n}{r}$ for this number. □

Remark 4.1.

(a) Some texts also use the symbol C^n_k 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. $\square$

Most of the remainder of this subsection will be about multiple selections from a collection of items.

Theorem 4.4.

Given are n items of which n_1 are alike, n_2 are alike, $\dots, n_r$ are alike ($n_1 + \cdots + n_r = n$). Then the number of distinguishable arrangements of those n items is

$$\binom{n}{n_1, n_2, \dots n_r} = \frac{n!}{n_1!\, n_2! \cdots n_r!}\,.$$

PROOF:

- We tag the group 1 items as $x_1^{(1)}, x_2^{(1)}, \dots, x_{n_1}^{(1)}$,
- the group 2 items as $x_1^{(2)}, x_2^{(2)}, \dots, x_{n_2}^{(2)}$,
- -
- the group r items as $x_1^{(r)}, x_2^{(r)}, \dots, x_{n_r}^{(r)}$,

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 $x_1^{(1)}, x_2^{(1)}, \dots, x_{n_1}^{(1)}$.

$$\text{Then only} \quad \frac{n!}{n_2!\, n_1!} \quad \text{arrangements remain,}$$

but for those the artificial distinction which was introduced by the subscipts is gone in groups 1 and 2.
We keep going with the remainin 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 4.7. 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

$$\frac{11!}{5! \cdot 4! \cdot 3!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{4 \cdot 3 \cdot 2 \cdot 3 \cdot 2} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{(3 \cdot 3)\,(4 \cdot 2) \cdot 2} = 11 \cdot 10 \cdot 7 \cdot 3 = 770 \cdot 3 = 2{,}310$$

distinguishable arrangements of the word SHANANANANA. □

Definition 4.3 (Multinomial coefficients)**.**

The numbers

$$\binom{n}{n_1\, n_2 \cdots n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}. \tag{4.3}$$

that appear in Theorem 4.4 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

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \qquad \text{for} \qquad \binom{n}{k,\, n-k} \tag{4.4}$$

and speak of **binomial coefficients**. Convention: We define $\binom{n}{k} := 0$ for $k > n$. □

The next theorem explains the appropriateness of the previous definition.

Theorem 4.5.

Let $r, n \in \mathbb{N}$ such $r \leq n$ and $x_1, x_2, \dots x_r \in \mathbb{R}$. Then

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{\substack{n_1,\dots,n_r \geq 0 \\ n_1+\cdots+n_r=n}} \binom{n}{n_1, n_2, \dots n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}. \tag{4.5}$$

In particular, if $n = 2$, we obtain the ***binomial theorem****:*

$$(x_1 + x_2)^n = \sum_{j=0}^{n} \binom{n}{j} x_1^j x_2^{n-j}.$$

PROOF:

First, we show that the case $n = 2$ follows from 4.5.

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

$$(x_1 + x_2 + \cdots + x_r)^n = (x_1 + x_2 + \cdots + x_r)(x_1 + x_2 + \cdots + x_r) \cdots (x_1 + x_2 + \cdots + x_r)$$

and obtain in the resulting expansion terms of the form

$a_1 \cdot a_2 \cdots a_n$ such that each factor a_j is either x_1 or x_2 ... or x_r.

In the following we consider the sizes $n_1, n_2, \dots, n_r$ as fixed

Note that it is not possible to obtain two selections

$\vec{a} = (a_1, a_2, \dots, a_n)$ and $\vec{b} = (b_1, b_2, \dots, b_n)$ such that $a_j = b_j$ 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 $a_1 a_2 \dots a_n$ 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 $(x_1 + x_2 + \cdots + x_r)^n$ 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 4.4, there are

$$\binom{n}{n_1, n_2, \dots n_r} = \frac{n!}{n_1!\, n_2! \cdots n_r!}.$$

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 $(x_1 + x_2 + \cdots + x_r)^n$, it follows that

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{\substack{n_1,\dots,n_r \geq 0 \\ n_1+\cdots+n_r = n}} \binom{n}{n_1, n_2, \dots n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r} . \ \blacksquare$$

Theorem 4.6.

Given are n distinct items and r distinct bins of fixed sizes $n_1, n_2, \dots, n_r$ 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

$$\binom{n}{n_1, n_2, \dots n_r} = \frac{n!}{n_1!\, n_2! \cdots n_r!} .$$

The proof is given after the following example which will help clarify how to interpret Theorem 4.6.

Example 4.8. 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. $\square$

PROOF of Theorem 4.6:

The proof is very similar to that of Theorem 4.4, 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 $x_1^{(1)}, x_2^{(1)}, \dots, x_{n_1}^{(1)}$.

$$\text{Then only}\quad \frac{n!}{n_2!\, n_1!}\quad \text{arrangements remain,}$$

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 4.1.

(A) *There are* $\binom{n-1}{r-1}$ *distinct integer–valued vectors* $\vec{x} = \big(x_1, x_2, \dots, x_r\big)$ *such that*

$$x_1 + x_2 + \cdots + x_r = n \quad \text{and} \quad x_i > 0,\ i = 1, \dots, r\,.$$

(B) *There are* $\binom{n+r-1}{r-1}$ *distinct integer–valued vectors* $\vec{y} = \big(y_1, y_2, \dots, y_r\big)$ *such that*

$$y_1 + y_2 + \cdots + y_r = n \quad \text{and} \quad y_i \geq 0,\ i = 1, \dots, r\,.$$

PROOF of **(A):**
Each such equation corresponds to an arrangement of n symbols $\otimes$ which denote the numbers $1, 2, \dots, n$ 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

$$\bullet\,\bullet \mid \bullet\,\bullet\,\bullet\bullet \mid \bullet\,\bullet\,\bullet$$

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**) $$\bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \otimes \bullet \cdots \otimes \bullet \otimes \bullet$$

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 $\sum\limits_{j=1}^{r} y_j = n; y_j \geq 0$ of part **(B)** becomes an equation $\sum\limits_{j=1}^{r} x_j = n + r; x_j > 0$ of part **(A)**, by setting $x_j := y_j + 1$.

In reverse, equation $\sum\limits_{j=1}^{r} x_j = n + r; x_j > 0$ of part **(A)** becomes an equation $\sum\limits_{j=1}^{r} y_j = n; y_j \geq 0$ 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 $\sum\limits_{j=1}^{r} x_j = n + r; x_j > 0$. Thus, there also that many of the form $\sum\limits_{j=1}^{r} y_j = n; y_j \geq 0$. 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 distinct 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. (But of course, $n_1 + \cdots + n_r = n$.)

Then each such selection corresponds to an integer vector $\vec{n} = (n_1, \dots, n_r)$ 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 $\vec{n} = (n_1, \dots, n_r)$ which is a solution of the equation $\sum_{j=1}^{r} n_j = n; n_j > 0$.

We obtain from Proposition 4.1 the following.

Proposition 4.2.

(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 4.1. ■

Example 4.9. 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 4.10 (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 $n - m$ 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.

$$\text{Thus, there are } \binom{n-d+1}{d} \text{ functioning arrangements.} \quad \square$$

We summarize the results of Theorem 4.4, Theorem 4.6, Proposition 4.1, and Proposition 4.2.

Remark 4.2. The multinomial coefficients

$$\binom{n}{n_1\, n_2 \cdots n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

of Definition 4.3 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 $\left(x_1 + x_2 + \cdots x_k\right)^n$.
- Distinct selections of n items into k distinct bins of fixed sizes $n_1, \ldots, n_k$. That is the WMS definition in their Theorem 2.3 of Ch.02.6.
- Subdividing n indistinguishable items into k partitions, where the sizes $n_1, \ldots, n_k$ of those partitions are allowed to vary for different subdivisions. Example: number of integer valued vectors $(n_1, \ldots, n_k)$ such that $n_k \geq 0$ and $\sum\limits_j n_j = n$. $\square$

5 More on Probability

This chapter corresponds to material found in WMS ch.2

5.1 Total Probability and Bayes Formula

Theorem 5.1 (Total Probability and Bayes Rule)**.**

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*

$$P(A) = \sum_{j=1}^{\infty} P(A \mid B_j)\, P(B_j)\,. \tag{5.1}$$

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum\limits_{i=1}^{k} P(A \mid B_i)\, P(B_i)}\,. \tag{5.2}$$

PROOF: See WMS ch.2. ■

5.2 Random Sampling and Urn Models With and Without Replacement

The following definition is **PRELIMINARY** and will be amended in Definition 5.2 (Sampling as a Random item) below (see p.68).

Definition 5.1.

(a) We call the action of picking n items $x_1, x_2, \dots, x_n$ 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 $x_1, x_2, \dots, x_n$) a **sample of size** n. □

Example 5.1. Sampling actions are each of the following.

(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 $n = 5$ times in a row recording each time the outcome and then replacing the ball (putting it back).

(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 **(f)**, 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 $[0, 1]$. (Computers can do that!) See Example 3.2**(b)** on p.37. Since there are infinitely many such numbers and the computer can generate any one of them, [21] $N = \infty$.

(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 $1, 2, \dot{;} 10$. This creates a very boring sample as far as a course called "Probability Theory" is concerned because no randomness is involved. □

Remark 5.1.

(a) We only are interested in sampling actions that involve randomness. In other words, if there is a set U such that $x_j \in U$ for all j, we have, for fixed n, a random item $\vec{X} : (\Omega, P) \to U^n$. Since deterministic actions also are (constant) random items, deterministic sampling actions are also covered.

(b) Since the "population" from which each item is sampled is the set U from **(a)**, it is possible to choose $\Omega = U^N$ 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 a formula or even a specific number for 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 6 (Discrete Random Variables and Random Items) where going into specific settings would hinder rather than help the understanding. □

Here is the promised amended version of Definition 5.1.

Definition 5.2 (Sampling as a Random item)**.** 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 $n \in \mathbb{N}, n \leq N$ (so $n < \infty$).

(a) Let $\vec{X} : (\Omega, P) \longrightarrow U^n$ be a random item with codomain U^n. If we interpret $\vec{X}$ as the action of picking n items

$$\vec{x} = x_1, x_2, \ldots, x_n = X(\vec{\omega}) = X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$$

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 the specific outcome (the list $\vec{x} = (x_1, x_2, \ldots, x_n)$) a **sample of size** n.

(c) In yet another instance of notational abuse, some people will refer to both the sampling action and an outcome of this action as a sample if the context makes it clear what is being considered. □

[21] in theory, since there is no such thing as "infinitely many") in our physical reality

Remark 5.2.

(a) You may wonder about the difference between a U^n–valued random item 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 the practice of WMS to call any probability space a sample space, the author likes to think not of (Ω, P), but only of $(U^n, P_{\vec{X}})$ as a sample space, since 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 sample picks happen with or without replacement? In other words, can the same $x \in U$ be picked more than once or are all sample picks 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 5.1. 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 5.3. 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.
So let us assume that n is fixed. What will make the sample representative of the population, i.e., what will be the best guarantee that the composition of the sample mirrors that of the population? It certainly would not help if the sample has, e.g., 90% students if the population of interest only has 20%. So, we can fix that by establishing quota and restrict the proportion of students to 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 atrributes for which we want to maintain close to identical proportions reasonably well.

(b) Clearly, a practical limit to the number of ways a (hopefully small) 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 a, say "is male" of attribute B (here: gender), when compared to the proportion in the population. And we need this for all important a and B.
The solution is to make the sample selection as random as possible. If we pick the first item at random, i.e., with the same chance $\frac{1}{N}$, then pick #2 at random from the remaining $N-1$, then pick #3 at random from the remaining $N-2$, ... and finally pick #n at random from the remaining $N-n+1$ items, then this degree of randomness should prevent any kind of gross distortion (bias) in the sample.

(c) So then, that means that every item has equal chance of being selected, doesn't it? **The answer is NO.** Rather, any collection $\vec{x} = x_1, \dots, x_n$ should have the same chance of being selected as any other collection $\vec{x}' = x_1', \dots, x_n'$. 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 must be $1/\binom{N}{n}$.
• If order does matter and we deal with permutations, then the answer is $1/P_n^N$.

(d) Would the above requirement be the same as simply asking that each item in the population has the same probability, $1/N$, of being selected? Next comes a counterexample. □

Example 5.2. 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 the ideal kind of samples (with respect to randomness of the selection). Note that the following definition is tied to sampling without replacement!

Definition 5.3 (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) We call the specific outcome of an SRS action a **simple random sample of size** n and also, in brief, an **SRS**.
(c) As in Definition 5.2 (Sampling as a Random item), as long as the context makes it clear what is being considered, some people will call both the SRS action and an outcome of this action as an SRS. ☐

Definition 5.4 (Urn models)**.** SRS requires that a single item is selected with equal probability $|U| = 1/N$. 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. Some authors also use the scenario of tickets in a box rather than balls in an urn.

(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 sample will contain no duplicates.
(b) An **urn model with replacement** describes a mechanism by which ablindfolded 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) The process is repeated until all n balls have been selected. ☐

More material may be added to this section at a later time.

6 Discrete Random Variables and Random Items

This chapter corresponds to material found in WMS ch.3

6.1 Probability Mass Function and Expectation

We start with a trivial observation.

Proposition 6.1. *A real–valued function of a random item is a random variable.*

PROOF: Let $X : (\Omega, P) \to \Omega'$ be a random item 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 6.1 (Probability mass function)**.**

For a discrete random item X on (Ω, P), define

$$p(x) := p_X(x) := P_X\{x\} = P\{X = x\}\,. \tag{6.1}$$

We call p_X the **probability mass function** (WMS: **probability function**) for X. We also write **PMF** for probability mass function. □

Theorem 6.1.

If p_X is the probability mass function of a discrete random item X, then

$$0 \le p_X(x) \le 1; \quad \textit{for all } x \tag{6.2}$$

$$\sum_{x \text{ s.t. } p_X(x)>0} p_X(x) = 1 \tag{6.3}$$

Proof: See WMS ch.3. ■

Next, we elaborate on the meaning of $\sum\limits_{x \text{ s.t. } p_X(x)>0} \dots$.

Definition 6.2 (Absolute Convergence)**.** $\boxed{\star}$

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

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

Remark 6.1. We mentioned in a footnote of Remark 3.2**(b)** on p.36 that the following is true for an absolutely convergent series $\sum a_j$:

Any rearrangement $\sum\limits_{j=1}^{\infty} a_{n_j} = a_{n_1} + a_{n_2} + \cdots$ converges to the same limit. $\square$

We make the following blanket assumption.

Assumption 6.1 (All series are absolutely convergent)**.**

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 item X which takes values in a set Ω', $g : \Omega' \to \mathbb{R}$ is a real–valued function and x_n is a sequence in Ω', then the series $\sum g(x_j)p_X(x_j)$ is absolutely convergent. $\square$

Remark 6.2. Assume that $p_X(x)$ is the probability mass function of a discrete random item 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. All that follows is trivial in the finite case, so let us 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 6.1 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

$$\sum_{x \in \Omega^*} g(x)p_X(x) \qquad \text{rather than} \qquad \sum_{j=1}^{\infty} g(x_j)p_X(x_j)\,.$$

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:

$$\begin{aligned} \sum_{x} g(x)p_X(x) \;&=\; \sum_{x \in \Omega'} g(x)p_X(x) \;=\; \sum_{x \in \Omega^*} g(x)p_X(x) \\ &=\; \sum_{p_X(x)>0} g(x)p_X(x) \;=\; \sum_{j=1}^{\infty} g(x_j)p_X(x_j)\,. \end{aligned} \tag{6.4}$$

Choosing $g(x) = 1$, we can express probabilities involving X as follows. If $B \subseteq \Omega'$, then

$$P\{X \in B\} \;= P_X(B) \;=\; \sum_{x \in B} p_X(x) \;=\; \sum_{x \in \Omega^* \cap B} p_X(x) \;=\; \sum_{x \in B, p_X(x)>0} p_X(x)\,. \; \square \tag{6.5}$$

Whereas a PMF is defined for any discrete random item Y, the next definition needs that the values of Y are numbers.

Definition 6.3 (WMS Ch.03.2, Definition 3.4)**.**

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 6.3.

A strict definition of $E[Y]$ would explicitly require that the sum $\sum\limits_y y \cdot p_Y(y)$ 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 6.1, we will quietly asssume that absolute convergence is satisfied for all random variables for which the expected value is used. $\square$

Theorem 6.2.

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:

$$E[g(Y)] \;=\; \sum_{\text{all } y} g(y)\, p_Y(y) \;=\; \sum_{\text{all } y} g(y)\, P\{Y=y\}\,. \tag{6.6}$$

PROOF: ■

The following corresponds to WMS Theorem 4.5.

Theorem 6.3.

Let $c \in \mathbb{R}$, Y be a continuous random variable and $g_1, g_2, g_n : \mathbb{R} \to \mathbb{R}$ be a list of n real-valued functions. Then

$$E[c] \;= c\,, \tag{6.7}$$

$$E[cg_j(Y)] \;= cE[g_j(Y)]\,. \tag{6.8}$$

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:

$$E\left[\sum_{j=1}^{n} g_j \circ Y\right] = \sum_{j=1}^{n} E[g_j \circ Y]\,. \tag{6.9}$$

PROOF: ■

The following cannot be found in the WMS text.

Theorem 6.4.

Let $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$ 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:

$$E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j]\,. \tag{6.10}$$

In other words, the expectation of the sum is the sum of the expectations.

PROOF: Not given here. ■

Remark 6.4.

(1) The last theorem encompasses all variants of Theorem 6.3. For example, (6.9) 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 item $\vec{Y} = (Y_1, \dots, Y_n)$, 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 6.4 (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,

$$Var[Y] := \sigma_Y^2 := E\big[(Y - E[Y])^2\big]. \tag{6.11}$$

We call $\sigma_Y := \sqrt{Var[Y]}$ The **standard deviation** of Y. □

Theorem 6.5.

If Y is a discrete random variable, then

$$Var[Y] = E\big[Y^2\big] - \big(E[Y]\big)^2.$$

PROOF:

$$\begin{aligned} Var[Y] &= E[(Y - E[Y])^2] = E(Y^2 - (2E[Y])Y + (E[Y])^2 \\ &= E(Y^2) - 2E[Y]E[Y] + (E[Y])^2 = E(Y^2) - (E[Y])^2. \ \blacksquare \end{aligned}$$

Theorem 6.6.

Let Y be a discrete random variable and $a, b \in \mathbb{R}$. Then

$$Var[aY + b] = a^2 Var[Y]. \tag{6.12}$$

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: Later. ■

Remark 6.5. Since $\sqrt{a^2} = -a$ for negative numbers a,

$$\sigma(aY) = |a|\sigma(Y). \ \square \tag{6.13}$$

The following cannot be found in the WMS text.

Theorem 6.7 (Bienaymé formula)**.**

Let $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$ 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

$$Var\left[\sum_{j=1}^n Y_j\right] = \sum_{j=1}^n Var[Y_j]. \tag{6.14}$$

In other words, for independent random variables, the variance of the sum is the sum of the variances.

PROOF: Not given here. ■

Remark 6.6. The independence is necessary, otherwise there are counterexamples:
If $Y_1 = Y_2 = Y$ for some random variable Y, then

$$Var[Y+Y] \;=\; Var[2Y] \;=\; 4Var[Y] \;\neq\; Var[Y] + Var[Y]\,. \;\square$$

6.2 Bernoulli Variables and the Binomial Distribution

Definition 6.5 (iid sequences)**.**

Let $X_1, X_2, \dots (\Omega, P) \to \Omega'$ be a sequence of random items. We speak of an **independent and identically distributed sequence**, in short, an We speak of an **iid sequence** of random items, if

(1) the X_j are independent. Here we take the naive definition of independence: The outcomes of any X_k are not influenced by the outcomes of the other X_j. We will give a formulation of independence in terms of probabilities in a later chapter.

(2) All random items have the same distribution:
$P_{X_1}(B) \;=\; P_{X_2}(B) \;=\; P_{X_3}(B) \;=\; \cdots$ for all j and all $B \subseteq \Omega'$.

- Note that this can also be written
$P\{X_1 \in B\} \;=\; P\{X_2 \in B\} \;=\; P\{X_3 \in B\} \;=\; \cdots$ for all j and all $B \subseteq \Omega'$.
- If the X_j are discrete random items, identical distribution translates to identical PMFs $p_{X_1}(x) \;=\; p_{X_2}(x) \;=\; p_{X_3}(x) \;=\; \cdots$ for all j and all $x \in \Omega'$. $\square$

Definition 6.6 (Bernoulli items and variables)**.**

Let X be a binary random item on a probability space (Ω, P), i.e., a random item which only assumes two outcomes, such as
• S (success) or F (failure) • T (true) or F (false) • Y (Yes) or N (No) • 1 or 0
We call X a **Bernoulli random item**. or a **Bernoulli trial**.

- We call $p := P\{X = \text{success}\}$ the **success probability** and $q := 1-p$, i.e., $q = P\{X = \text{failure}\}$, 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**.
- We call an iid sequence of Benoulli trials a **Bernoulli sequence**. $\square$

Remark 6.7.

(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. $\square$

Theorem 6.8 (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

$$E[X] \,=\, p \qquad \textit{and} \qquad Var[X] \,=\, pq\,. \tag{6.15}$$

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, $Var[X] \,=\, p - p^2 \,=\, p(1-p) \,=\, pq$. $\blacksquare$

Definition 6.7 (Binomial Distribution)**.**

Let $n \in \mathbb{N}$ and $0 \leq p \leq 1$. Let Y be a random variable with probability mass function

$$p_Y(y) \,=\, \binom{n}{y} p^y\, q^{n-y}\,. \tag{6.16}$$

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). $\square$

Remark 6.8. How does one see that p_Y of (6.16) satisfies $p_Y(y) \geq 0$ for all y and $\sum_y p_Y(y) = 1$, i.e., it really is a probability mass function?

- $p_Y(y) \geq 0$ is true, since $p, q, \binom{n}{y} \geq 0$.
- We apply the binomial theorem (see Theorem 4.5) 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 6.9. *Let $X_1, X_2, X_n, \dots$ 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: See the deliberations in WMS before their Definition 3.7. $\blacksquare$

Theorem 6.10 (Expected value and variance of a binom(n,p) variable)**.**

Let Y be a binom(n,p) variable. Then

$$E[Y] \,=\, n\,p \qquad \textit{and} \qquad Var[Y] \,=\, n\,pq\,. \tag{6.17}$$

PROOF: Let $X_1, \dots, X_n$ 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 6.8, Theorem 6.4 on p.74, and, since the X_j are independent, Theorem 6.7 (Bienaymé formula) on p.75,

$E[Y'] = \sum_{j=1}^{n} E[X_j] = n\,p$ and $Var[Y'] = \sum_{j=1}^{n} Var[X_j] = n\,p\,q$.

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 6.9 on p.77, Y' has a binom(n, p) distribution.

Since expected value and variance of a discrete random variable are determined by its PMF,

$E[Y] = E[Y'] = n\,p$ and $Var[Y] = Var[Y] = n\,p\,q$. ■

6.3 Geometric + Negative Binomial + Hypergeometric Distributions

Definition 6.8 (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:

$$p_Y(y) = q^{y-1}\,p\,, \qquad \text{for } y = 1, 2, 3, \dots\,. \quad \square \tag{6.18}$$

Theorem 6.11. *Let $X_1, X_2, \dots : (\Omega, P) \to \{S, F\}$ be an infinite Bernoulli sequence with success probability $0 \le p \le 1$.*

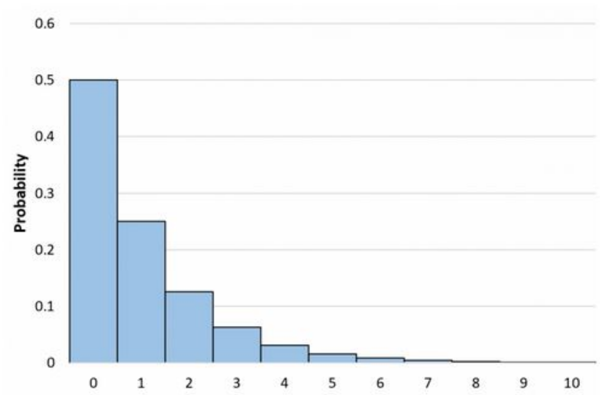
Let $T(\Omega, P) \to \mathbb{N}$ be the random variable

$$T(\omega) := \begin{cases} \textit{smallest integer } k > 0 \textit{ such that } X_k(\omega) = S & \textit{if such a } k \textit{ exists}, \\ \infty\,, & \textit{else}. \end{cases}$$

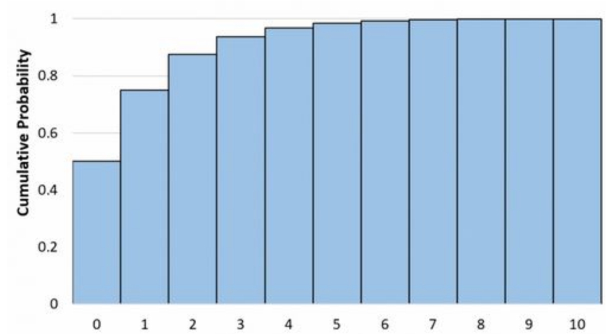
- *Then T is geom(p).*

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

$$\begin{aligned} P\{X_1 = F,\, X_2 = F, X_{n-1} = F, X_n = S\} &= P\{X_1 = F\} \cap \cdots \{X_{n-1} = F\} \cap \{X_n = S\} \\ &= P\{X_1 = F\} \cdot P\{X_2 = F\} \cdots P\{X_{n-1} = F\} \cdot P\{X_n = S\} = q^{n-1}\,p\,. \quad ■ \end{aligned}$$



6.1 (Figure). **PMF for geom(0.5).**



6.2 (Figure). **CDF for geom(0.5).**

Remark 6.9. 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. $\square$

Theorem 6.12 (WMS Ch.03.5, Theorem 3.8).

If Y is a geom(p) random variable, then

$$E[Y] \;=\; \frac{1}{p}\,, \qquad \textit{and} \qquad Var[Y] \;=\; \frac{1-p}{p^2}\,.$$

PROOF: See the WMS text. ■

Definition 6.9 (Negative binomial distribution). $\boxed{\star}$

A random variable Y has a **negative binomial distribution** with parameters p and r if

$$p_Y(y) \;=\; \binom{y-1}{r-1} p^r q^{y-r}\,, \quad \text{where } \; r \in \mathbb{N}, \;\; y = r, r+1, r+2, \dots, \;\; 0 \le p \le 1\,. \quad \square \tag{6.19}$$

This last definition has been marked as $\boxed{\star}$, so you are not expected to recall p_Y from memory. In contrast, the next theorem is NOT optional.

Theorem 6.13. *Let $X_1, X_2, \dots : (\Omega, P) \to \{S, F\}$ 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,

$$X_n(\omega) \;=\; \begin{cases} S \;=\; \textit{success} & \textit{if } n \textit{ is one of } t_1, t_2, \dots, \\ F \;=\; \textit{failure}\,, & \textit{else.} \end{cases}$$

Two points to note:

- *There will be different subsequences $t_1, t_2, \dots$ for different arguments $\omega \in \Omega$. In other words, we are dealing with a sequence of random variables(!)*

$$t_1 = T_1(\Omega),\ t_2 = T_2(\Omega),\ t_3 = T_3(\Omega),\ \dots$$

- *It is possible that we are dealing with an ω for which there are only 18 successes in the entire (infinite) sequence $X_1(\omega), X_2(\omega), \dots$. In this case, we define $T_{19}(\omega) = T_{20}(\omega) = \cdots = \infty$. More generally, if $r \in \mathbb{N}$ and the sequence $X_1(\omega), X_2(\omega), \dots$ 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: See the introductory remarks of WMS Chapter 3.6 before Definition 3.9. ■

Remark 6.10. If we think of the indices n of the sequence X_n as points in time, we can interpret the random variables $T_1, T_2, \dots$ as follows.

- T_r is the time of the rth success in the underlying Bernoulli sequence X_n. □

Theorem 6.14. $\boxed{\star}$

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 6.10 (Hypergeometric distribution)**.**

A random variable Y has a **hypergeometric distribution** with parameters N, R and n if its PMF is

$$p_Y(y) \ =\ \frac{\binom{R}{y}\binom{N-R}{n-y}}{\binom{N}{n}}\,, \tag{6.20}$$

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 6.11. For the following you should review Section 5.2 (Random 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 $X_1, X_2, \dots, X_n$ 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 5.4 (Urn models) on p.70. 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. $\square$

Theorem 6.15.

- *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 5.4**(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 4.1 (multiplication rule of combinatorial analysis) on p.55, 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 6.16 (WMS Ch.03.7, Theorem 3.10)**.**

Let Y be a hypergeometric random variable with parameters N, R and n. Then

$$E[Y] \;=\; \frac{nr}{N} \qquad \textit{and} \qquad Var[Y] \;=\; n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right). \tag{6.21}$$

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:

$$E[Y] \;=\; n \cdot p \qquad \text{and} \qquad Var[Y] \;=\; n \cdot p \cdot q \left(\frac{N-n}{N-1}\right).$$

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. ■

6.4 The Poisson Distribution

We start out with the simple observation that $e^x = \sum\limits_{j=0}^{\infty} \frac{x^j}{j!}$ for any $x \in \mathbb{R}$.

Proposition 6.2. *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 $e^x \;=\; \sum\limits_{j=0}^{\infty} \frac{x^j}{j!}$, 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 6.11 (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

$$p_Y(y) \;=\; \frac{\lambda^y}{y!}\,e^{-\lambda}, \qquad \text{for } y = 0, 1, 2, \dots, \quad \square$$

We follow WMS Chapter 3.8 to show what phenomena can be modeled by a Poisson variables

Proposition 6.3. *Given is some event of interest, E.*

(1) *We define a random variable Y which counts 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 (twodimensional unit).*

(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,*
  $\boxdot$ *the $X_j^{(n)}$ are independent*
  $\boxdot$ *for each j, $P\{X_j^{(n)} = 0$ or $1\} = 1$: E (i.e., the event of interest) happens at most once in such a small subunit*
  $\boxdot$ $p_n := P\{X_j^{(n)} = 1\}$ *is constant in j ($j = 1, 2, \dots, n$)*
  $\boxdot$ $\lambda := n \cdot p_n$ *is constant in n*

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:*

$$\lim_{n\to\infty} \binom{n}{p} p^y (1-p)^{n-y} \;=\; e^{-\lambda} \cdot \frac{\lambda^y}{y!} \qquad \text{for } y = 0, 1, 2, \dots, \tag{6.22}$$

PROOF: We follow WMS:

Recall that $\lambda = np$. Thus,

$$\begin{aligned}
\binom{n}{p} p^y (1-p)^{n-y} &= \frac{n(n-1)\cdots(n-y+1)}{y!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\
&= \frac{\lambda^y}{y!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-y+1)}{n^y} \left(1 - \frac{\lambda}{n}\right)^{-y} \\
&= \left(\frac{\lambda^y}{y!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-y} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{y-1}{n}\right).
\end{aligned} \tag{$\star$}$$

From calculus we obtain $\lim\limits_{n\to\infty} \left(1 - \dfrac{\lambda}{n}\right)^n = e^{-\lambda}$. 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

$$\lim_{n\to\infty}\binom{n}{p}p^y(1-p)^{n-y} \;=\; \left(\frac{\lambda^y}{y!}\right)e^{-\lambda}\,. \;\blacksquare$$

Theorem 6.17 (WMS Ch.03.8, Theorem 3.11)**.**

A poisson(λ) random variable has expectation and variance λ. In other words,

$$E[Y] \;=\; Var[Y] \;=\; \lambda\,. \tag{6.23}$$

PROOF: We only show that $E[Y]=\lambda$.

$$E(Y) \;=\; \sum_y y p_Y(y) \;=\; \sum_{y=0}^{\infty} y\,\frac{\lambda^y e^{-\lambda}}{y!}\,. \;=\; \sum_{y=1}^{\infty} y\,\frac{\lambda^y e^{-\lambda}}{y!} \;=\; \lambda\sum_{y=1}^{\infty}\frac{\lambda^{y-1}e^{-\lambda}}{(y-1)!}\,.$$

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 $E[Y]=\lambda$. $\blacksquare$

We refer to the WMS text for examples of random variables with a Poisson distribution.

6.5 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 6.12 (kth Moment)**.**

If Y is a random variable and $k\in\mathbb{N}$,

$$\mu'_k \;:=\; E[Y^k] \tag{6.24}$$

is called the kth **moment** of Y. μ'_k also is referred to as the kth **moment of Y about the origin**. $\square$

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 6.13 (kth Central Moment)**.**

If Y is a random variable and $k \in \mathbb{N}$,

$$\mu_k \;:=\; E[(Y - E[Y])^k] \;=\; E[(Y-\mu)^k] \tag{6.25}$$

is called the kth **central moment** of Y. μ_k also is referred to as the kth **moment of Y about its mean**. $\square$

Proposition 6.4. $\boxed{\star}$ *Under fairly slight assumptions the following is true for two random variables Y_1 and Y_2.*

$$\textit{If}\;\; E[Y_1^k] \;=\; E[Y_2^k] \;\; \textit{for } k = 1,2,3,\dots, \quad \textit{then} \;\; P_{Y_1} \;=\; P_{Y_2}\,.$$

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, [22] $m_Y(t)$ uniquely determines the entire distribution of Y.

Definition 6.14 (Moment–generating function)**.**

Let Y be a random variable for which one can find $\delta > 0$ (no matter how small), such that

$$m(t) \;:= m_Y(t) \;:=\; E\big[e^{tY}\big] \qquad \text{is finite for } \; |t| < \delta\,. \tag{6.26}$$

Then we say that Y has **moment–generating function**, in short, **MGF**, $m_Y(t)$. $\square$

Theorem 6.18. *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 $t=0$:

$$\mu_k' \;= m^{(k)}(0) \;=\; \frac{d^k m(t)}{dt^k}\Big|_{t=0}\,. \tag{6.27}$$

PROOF: We write $m(t)$ for $m_Y(t)$. From the series expansion $e^x \;=\; \sum\limits_{k=0}^{\infty} \frac{x^k}{k!}$, we obtain

$$\begin{aligned} m(t) \;&= E\big[e^{tY}\big] \;=\; E\left[\sum_{k=0}^{\infty} \frac{t^k Y^k}{k!}\right] \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!} E\big[Y^k\big] \\ &= 1 \,+\, t\mu_1' \,+\, \frac{t^2}{2!}\mu_2' \,+\, \frac{t^3}{3!}\mu_3' \,+\, \cdots\,. \end{aligned}$$

[22] See Proposition 6.4

Taking derivatives repeatedly,

$$\begin{aligned} m^{(1)}(t) &= \mu_1' + \frac{2t}{2!}\mu_2' + \frac{3t^2}{3!}\mu_3' + \cdots, \\ m^{(2)}(t) &= \mu_2' + \frac{2t}{2!}\mu_3' + \frac{3t^2}{3!}\mu_4' + \cdots, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ m^{(k)}(t) &= \mu_k' + \frac{2t}{2!}\mu_{k+1}' + \frac{3t^2}{3!}\mu_{k+2}' + \cdots. \end{aligned}$$

Thus, for $t = 0$,

$$m^{(1)}(0) = \mu_1', \quad m^{(2)}(0) = \mu_2', \quad \ldots, \quad m^{(k)}(0) = \mu_k'. \ \blacksquare$$

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 6.5. $\boxed{\star}$ *If Y is a poisson(λ) random variable ($\lambda > 0$), its MGF is*

(6.28) $$m_Y(t) = e^{\lambda(e^t - 1)}. \ \square$$

PROOF: For this proof, we abbreviate $\tilde{\lambda} := \lambda e^t$. We obtain

$$\begin{aligned} m_Y(t) &= E\left(e^{tY}\right) = \sum_{y=0}^{\infty} e^{ty} p(y) = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{(-\lambda)}}{y!} = \sum_{y=0}^{\infty} \lambda^y (e^t)^y \frac{e^{-\lambda}}{y!} \\ &= \sum_{y=0}^{\infty} \frac{\left(\lambda e^t\right)^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\tilde{\lambda}^y}{y!} = e^{-\lambda} e^{\tilde{\lambda}} e^{-\tilde{\lambda}} \sum_{y=0}^{\infty} \frac{\tilde{\lambda}^y}{y!} \\ &= \left(e^{-\lambda} e^{\tilde{\lambda}}\right) \sum_{y=0}^{\infty} e^{-\tilde{\lambda}} \frac{\tilde{\lambda}^y}{y!} = \left(e^{-\lambda} e^{\tilde{\lambda}}\right) \sum_{y=0}^{\infty} p_{\tilde{Y}}(y), \end{aligned}$$

where $p_{\tilde{Y}}$ is the PMF of a poisson($\tilde{\lambda}$) random variable $\tilde{Y}$. It follows that $\sum\limits_{y=0}^{\infty} p_{\tilde{Y}}(y) = 1$, hence,

$$m_Y(t) = e^{-\lambda} e^{\tilde{\lambda}} = e^{(-1)\lambda} e^{\lambda e^t} = e^{\lambda}(e^t - 1). \ \blacksquare$$

- The subsection titled "The Tchebysheff Inequality" which was at this location has been integrated into subsection 7.8 (Inequalities for Probabililities)

7 Continuous Random Variables

7.1 Cumulative Distribution Function of a Random Variable

The material found in this section does not make any references to continuous random variables.

Definition 7.1 (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.

(7.1) $$F(y) \;:=\; F_Y(y) \;:=\; P\{Y \le y\} \qquad \text{for } y \in \mathbb{R}\,.\ \square$$

Example 7.1. 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,

$$p_Y(0) \;=\; \frac{1}{16}\,, \qquad p_Y(1) \;=\; 2\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) \;=\; \frac{6}{16}. \qquad p_Y(2) \;=\; \frac{9}{16}\,.$$

It follows that

- $y < 0 \;\Rightarrow\; F_Y(y) \;=\; P_Y(\emptyset) \;=\; 0.$
- $0 \le y < 1 \;\Rightarrow\; F_Y(y) \;=\; p_Y(0) \;=\; 1/16.$
- $1 \le y < 2 \;\Rightarrow\; F_Y(y) \;=\; p_Y(0) + p_Y(1) \;=\; 7/16.$
- $y \ge 2 \;\Rightarrow\; F_Y(y) \;=\; p_Y(0) + p_Y(1) + p_Y(2) \;=\; 1.$

Note that F_Y is constant on intervals A of $\mathbb{R}$ if $p_Y(a) = 0$ for all $a \in A$. $\square$

Theorem 7.1 (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.30.*

PROOF: Obvious. ■

Remark 7.1. $\boxed{\star}$ There is a fourth property that is satisfied by all CDFs:
$y \mapsto F_Y(y)$ is **right continuous** at all arguments y.

This 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

$$\lim_{n\to\infty} F(y_n) = F(y) . \quad \square$$

7.2 Continuous Random Variables and Probability Density Functions

Definition 7.2 (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 7.1. *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,*

$$a \in \mathbb{R} \Rightarrow P\{Y = a\} = P_Y\{a\} = 0 , \tag{7.2}$$

$$\begin{aligned} -\infty < a < b < \infty \Rightarrow P\{a < Y < b\} &= P\{a \leq Y < b\} \\ = P\{a < Y \leq b\} &= P\{a \leq Y \leq b\} . \end{aligned} \tag{7.3}$$

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),

$$P\{Y = a\} \leq P\{a - \frac{1}{n} < Y \leq a\} = P\{Y \leq a\} - P\{Y \leq a - \frac{1}{n}\} = F_Y(a) - F_Y\left(a - \frac{1}{n}\right) .$$

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 (7.2).
This result, plus additivity of probability measures, plus

$$[a, b] =]a, b[\biguplus \{a\} \biguplus \{b\} , \quad [a, b] = [a, b[\biguplus \{b\} , \quad [a, b] =]a, b] \biguplus \{a\} ,$$

show that (7.3) 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 7.1 (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. $\square$

This last assumption allows us to make the following definition.

Definition 7.3 (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. $\square$

Theorem 7.2.

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*

$$P\{a < Y \leq b\} \;=\; F_Y(b) - F_Y(a) \;=\; \int_a^b f(y)dy\,. \tag{7.4}$$

(2) $f_Y(y) \;\geq\; 0$ *for* $-\infty < y < \infty$.

(3) $\int\limits_{\infty}^{\infty} f(y)dy \;=\; 1$.

PROOF: **(1)** is the fundamental theorem of calculus. Of course, we interpret $\int\limits_a^b f(y)dy$ as follows. Assume that some of the points y at which $f'_Y(y)$ does not exist fall with the interval $[a, b]$. Our assumption guarantee that there are only finitely many, say,

$$a \;\leq y_1 \;<\; y_2 \;< \cdots y_k \leq\; b\,.$$

Then, by the definition of integrals,

$$\int\limits_a^b f(y)dy \;=\; \int\limits_a^{y_1} f(y)dy \;+\; \int\limits_{y_1}^{y_2} f(y)dy \;+ \cdots +\; \int\limits_{y_k}^{b} f(y)dy\,.$$

(2) and **(3)** are obvious. $\blacksquare$

Remark 7.2. We combine (7.3) and (7.4) 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

$$\begin{aligned} P\{a < Y < b\} \;&= P\{a \leq Y \leq b\} \;=\; P\{a \leq Y < b\} \\ &= P\{a < Y \leq b\} \;=\; \int_a^b f(y)dy\,. \;\square \end{aligned} \tag{7.5}$$

The next definition applies to any random variable, be it continuous or discrete or neither. It is based on the following elementary observation.

Remark 7.3. $\boxed{\star}$ 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 $\big[\,\alpha < \alpha'$ and $F_Y(\alpha) \geq p\,\big] \ \Rightarrow \ F_Y(\alpha') \geq p$.
- In other words, $\big[\,\alpha < \alpha'$ 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 $A_p = [\beta, \infty[$.

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, [23] $F_Y(\beta) = \lim\limits_{n\to\infty} F_Y(\beta_n)$.
- 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. $\square$

Definition 7.4 (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.,

$$\phi_p \ = \ \min\{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\} \tag{7.6}$$

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. $\square$

Example 7.2. 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 CDF $F_Y(y) = 0$ for $y < 0$, $F_Y(y) = 0.5$ for $0 \leq y < 1$, $F_Y(y) = 1$ for $y \geq 1$.
We now easily compute ϕ_p for any $0 < p < 1$ by separately considering the cases

$$\begin{aligned}
0 < p < \tfrac{1}{2}: &\quad F_Y(\alpha) \geq p \ \Leftrightarrow \ \alpha \geq 0. \text{ Thus, } \phi_p = 0.\\
p = \tfrac{1}{2}: &\quad F_Y(\alpha) \geq \tfrac{1}{2} \ \Leftrightarrow \ \alpha \geq 0. \text{ Thus, } \phi_{1/2} = 0.\\
\tfrac{1}{2} < p < 1: &\quad F_Y(\alpha) \geq p \ \Leftrightarrow \ \alpha \geq 1. \text{ Thus, } \phi_p = 1.
\end{aligned}$$

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

$$\min\{\alpha \in \mathbb{R} : F_Y(\alpha) \geq p\}$$

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! $\square$

[23] See Remark 7.1 on p.87.

The issues encountered in that last example do not occur if $F_Y(y)$ is a continuous function of y.

Proposition 7.2. *Let Y be a continuous random variable with CDF $F_Y(y)$. Then*

$$\phi_p \;=\; \min\{\alpha \in \mathbb{R} : F_Y(\alpha) = p\}\,. \tag{7.7}$$

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 7.4. 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. □

7.3 Expected Value, Variance and MGF of a Continuous Random Variable

Assumption 7.2 (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 $\int\limits_{-\infty}^{\infty} y f(y) dy$ 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 7.5 (Expected value of a continuous random variable)**.**

Let Y be a continuous random variable with PDF $f_Y(y)$. We call

$$E(Y) \;:=\; \int_{-\infty}^{\infty} y f_Y(y)\,dy \tag{7.8}$$

the **expected value**, also **expectation** or **mean** of Y. □

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 6.2 on p.73. Note that the continuous random variable $\omega \mapsto g\big(Y(\omega)\big)$ of that theorem is covered by Assumption 7.2 on p.91, i.e., $E\big[g \circ Y\big]$ exists.

Theorem 7.3.

Let Y be a discrete or 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

$$E[g(Y)] \;=\; \int_{\infty}^{\infty} g(y) f_Y(y)\, dy\,. \tag{7.9}$$

PROOF: ■

The following corresponds to WMS Theorem 4.5.

Theorem 7.4.

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

$$E[c] \;= c\,, \tag{7.10}$$

$$E[cg_j(Y)] \;= cE[g_j(Y)]\,. \tag{7.11}$$

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:

$$E\left[\sum_{j=1}^{n} g_j \circ Y\right] \;=\; \sum_{j=1}^{n} E[g_j \circ Y]\,. \tag{7.12}$$

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 7.5 and 7.6 below.

Definition 7.6. $\boxed{\star}$

If $Y_1, Y_2, \dots, Y_m$ is a list of discrete random variables and $Y_1', Y_2', \dots, Y_n'$ is a list of continuous random variables, all of which are defined on the same probability space (Ω, P), then we define

$$E\left[\sum_{i=1}^{m} Y_i + \sum_{j=1}^{n} Y_i'\right] := \sum_{i=1}^{m} E[Y_i] + \sum_{j=1}^{n} E[Y_i']\,p\,. \quad \square \tag{7.13}$$

The following is the continuous random variables version of Theorem 6.4 on p.74.

Theorem 7.5.

*Let $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$ be random variables. (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$ by Assumption 7.2.**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:

$$E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j]\,. \tag{7.14}$$

In other words, the expectation of the sum is the sum of the expectations.

PROOF: Not given here. ■

We extend Definition 6.4 on p.74 of the variance and standard deviation of a discrete random variable to the continuous case without modification, i.e.,

$$Var[Y] := \sigma_Y^2 := E\big[(Y - E[Y])^2\big]\,, \tag{7.15}$$

$$\sigma_Y := \sqrt{Var[Y]}\,. \tag{7.16}$$

Theorems 6.5, 6.6 6.7 about the variances of discrete random variables have the following counterpart.

Theorem 7.6. *Let Y be a discrete or continuous random variable. Let $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$ be independent random variables (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$ by Assumption 7.2.**B**). Some may be continuous, others may be discrete. Further, let $a, b \in \mathbb{R}$. Then*

$$Var[Y] = E\big[Y^2\big] - \big(E[Y]\big)^2\,, \tag{7.17}$$

$$Var\,[aY + b] = a^2 Var[Y]\,, \tag{7.18}$$

$$Var\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var[Y_j]\,. \tag{7.19}$$

PROOF: The proof of (7.17) is the same as for Theorem 6.5 on p.75. The proof of the other formulas is not given here. ■

Remark 7.5. Note that independence of $Y_1, \dots, Y_n$ is required for the validity of (7.19)! $\square$

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 7.7. For $k \in \mathbb{N}$, we define

(7.20) $$\mu'_k := E[Y^k] \qquad (k\text{th } \textbf{moment of } Y \textbf{ about the origin})$$

(7.21) $$\mu_k := E[(Y - E[Y])^k] = E[(Y-\mu)^k] \qquad (k\text{th } \textbf{central moment of } Y)$$

(7.22) $$m(t) := m_Y(t) := E\left[e^{tY}\right] \qquad (\textbf{moment–generating function of } Y)$$

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 $|t| < \delta$. $\square$

Theorem 6.18 on p.85 remains valid for continuous random variables:

Theorem 7.7.

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 $t = 0$:

(7.23) $$\mu'_k = m^{(k)}(0) = \frac{d^k m(t)}{dt^k}\Big|_{t=0}.$$

PROOF: The proof of Theorem 6.18 can be used without any alterations. $\blacksquare$

Proposition 7.3.

Let Y be a random variable with MGF $m_Y(t)$. Let $a, b \in \mathbb{R}, Y' := Y + a, Y'' := bY$. Then

(7.24) $$m_{Y'}(t) = e^{ta} m_Y(t),$$

(7.25) $$m_{Y''}(t) = m_Y(bt).$$

PROOF: To prove (7.24), 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 (7.25) follows from

$$m_{Y''}(t) = E[e^{t(bY)}] = E[e^{(tb)Y}] = m_Y(tb). \blacksquare$$

7.4 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 7.8 (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

$$f_Y(y) \;=\; \begin{cases} \dfrac{1}{\theta_2 - \theta_1}\,, & \text{if } \theta_1 \le y \le \theta_2, \\ 0\,, & \text{else.} \end{cases} \tag{7.26}$$ □

Theorem 7.8 (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

$$E[Y] \;=\; \frac{\theta_1 + \theta_2}{2} \qquad \textit{and} \qquad Var[Y] \;=\; \frac{(\theta_2 - \theta_1)^2}{12}\,.$$

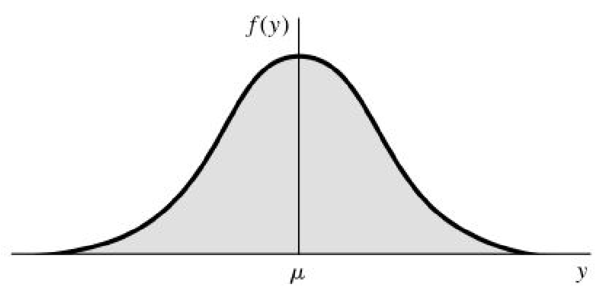
PROOF: A simple exercise in integrating $\int\limits_{\theta_1}^{\theta_2} y\,dy$ and $\int\limits_{\theta_1}^{\theta_2} y^2\,dy$. ■

7.5 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

Definition 7.9 (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

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, \qquad (y \in \mathbb{R}) . \quad \square \tag{7.27}$$

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 7.1.

$$(y-\mu)^2 - 2yt\sigma^2 = \left[y - (\mu + t\sigma^2)\right]^2 - 2\mu t\sigma^2 - t^2\sigma^4 . \tag{7.28}$$

PROOF: We multiply out the right–hand expression and obtain

$$\begin{aligned}
\text{R.S.} &= \left[y - (\mu + t\sigma^2)\right]^2 - 2\mu t\sigma^2 - t^2\sigma^4 \\
&= y^2 - 2y(\mu + t\sigma^2) + (\mu^2 + 2\mu t\sigma^2 + t^2\sigma^4) - 2\mu t\sigma^2 - t^2\sigma^4 \\
&= y^2 - 2\mu y - 2yt\sigma^2 + \mu^2 \\
&= (y-\mu)^2 - 2yt\sigma^2 = \text{L.S.} \ \blacksquare
\end{aligned}$$

Proposition 7.4.

Let the random variable Y be $\mathcal{N}(\mu, \sigma^2)$. Then

$$m_Y(t) = e^{\mu t + (\sigma^2 t^2)/2} . \tag{7.29}$$

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 7.1 and obtain for the exponent the following.

$$\begin{aligned}
-\frac{1}{2\sigma^2}\left[(y-\mu)^2 - 2yt\sigma^2\right] &= -\frac{1}{2\sigma^2}\left\{\left[y - (\mu + t\sigma^2)\right]^2 - 2\mu t\sigma^2 - t^2\sigma^4\right\} \\
&= -\frac{\left[y - (\mu + t\sigma^2)\right]^2}{2\sigma^2} + \frac{1}{2\sigma^2}\left[2\mu t\sigma^2 + t^2\sigma^4\right] \\
&= \mu t + \frac{t^2\sigma^2}{2} - \frac{1}{2}\left[\frac{y - (\mu + t\sigma^2)}{\sigma}\right]^2
\end{aligned}$$

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 $\int\limits_{-\infty}^{\infty} \varphi(y)dy$, 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

$$m_Y(t) = e^{\mu t + \frac{t^2\sigma^2}{2}}. \quad \blacksquare$$

Theorem 7.9 (WMS Ch.04.5, Theorem 4.7)**.**

If Y is a normally distributed random variable with parameters μ and σ, then

$$E(Y) = \mu \quad \textit{and} \quad V(Y) = \sigma^2.$$

PROOF: We differentiate $m_Y(t) = \exp\{\mu t + \frac{t^2\sigma^2}{2}\}$ twice and obtain

$$\begin{aligned} m_Y'(t) &= \left(\mu + t\sigma^2\right) \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}, \\ m_Y''(t) &= \left(\mu + t\sigma^2\right)^2 \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\} + \sigma^2 \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}. \end{aligned}$$

Thus,the first and second moment about the origin are

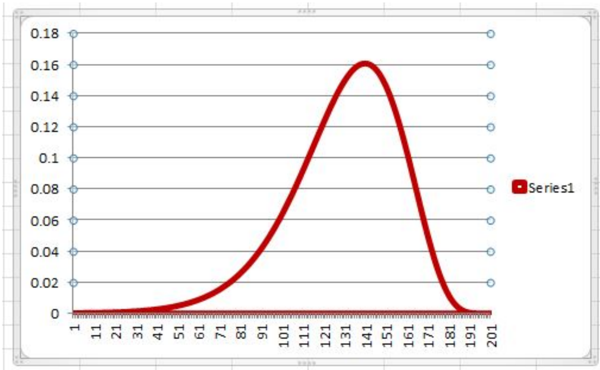
$$\begin{aligned} E[Y] &= \mu_1' = m_Y'(0) = (\mu+0)e^0 = \mu, \\ E[Y^2] &= \mu_2' = m_Y''(0) = (\mu+0)^2e^0 + \sigma^2e^0 = \mu^2 + \sigma^2. \end{aligned}$$

Finally,

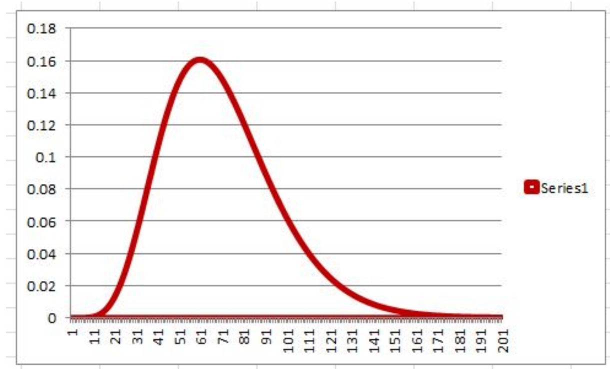
$$Var[Y] = E[Y^2] - \left(E[Y]\right)^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2. \quad \blacksquare$$

7.6 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 7.10 (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

$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,, & \text{if } 0 \leq y < \infty\,, \\ 0\,, & \text{else}\,, \end{cases} \tag{7.30}$$

where $\Gamma(\alpha)$ is the **gamma function**

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy\,. \tag{7.31}$$

We also express that by saying that Y is gamma(α, β). $\square$

Proposition 7.5. *The gamma function satisfies the following:*

$$\Gamma(1) = 1\,, \tag{7.32}$$

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \qquad \textit{for all } \alpha > 1\,, \tag{7.33}$$

$$\Gamma(n) = (n-1)! \qquad \textit{for all } n \in \mathbb{N}\,. \tag{7.34}$$

PROOF: (7.32) is immediate from $\int\limits_0^\infty e^{-y}dy = -e^{-y}\Big|_0^\infty = 0-(-1) = 1$.

We obtain (7.33) 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 (7.34) we observe that repeated application of (7.33) 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 (7.32), it follows that

$$\Gamma(n) \;=\; (n-1)(n-2)(n-3)\cdots 2\cdot 1 \;=\; (n-1)!\,.$$

Proposition 7.6.

If the random variable Y is gamma(α, β) it has MGF

(7.35) $$m_Y(t) \;= \frac{1}{(1-\beta t)^\alpha} \qquad \textit{for } t \;<\; \frac{1}{\beta}\,.$$

PROOF: $\boxed{\star}$ 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,

$$\begin{aligned}m_Y(t) &= E(e^{tY}) \;=\; \int_0^\infty e^{ty}\left[\frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\right]dy\\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}\exp\left[ty \;-\; \frac{y}{\beta}\right]dy\\ &\overset{\textbf{(B)}}{=} \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}\,e^{-y/\tilde{\beta}}\,dy\,.\end{aligned}$$

It follows from **(A)** that $\beta^\alpha \;=\; (1-t\beta)^\alpha\cdot\tilde{\beta}^\alpha$. Hence,

$$m_Y(t) \;=\; \left(\frac{1}{(1-t\beta)^\alpha}\right)\frac{1}{\tilde{\beta}^\alpha\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}\,e^{-y/\tilde{\beta}}\,dy\,. \;=\; \left(\frac{1}{(1-t\beta)^\alpha}\right)\int_0^\infty \varphi(y)\,dy\,,$$

where the function $\varphi(y)$ is the PDF of a gamma$(\alpha, \tilde{\beta})$ random variable. Thus $\int\limits_0^\infty \varphi(y)\,dy = 1$ and we conclude that $m_Y(t) = 1/(1-t\beta)^\alpha$. ■

Theorem 7.10 (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 7.11 (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 7.11 (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 7.10 with $\alpha = \nu/2$ and $\beta = 2$. $\blacksquare$

Definition 7.12 (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

$$f_Y(y) \;=\; \begin{cases} \dfrac{1}{\beta}\,e^{-y/\beta}\,, & \text{for } 0 \le y < \infty\,, \\ 0\,, & \text{else}\,. \;\square \end{cases} \tag{7.36}$$

Remark 7.6. In many textbooks exponential random variables are expressed in terms of $\lambda = 1/\beta$. Then its PDF is

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & \text{for } 0 \le y < \infty\,, \\ 0\,, & \text{else}\,. \quad \square \end{cases} \tag{7.37}$$

Theorem 7.12.

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 7.10 with $\alpha = 1$. ■

Proposition 7.7 (Memorylessness of the exponential distribution)**.** *Let Y be an exponential random variable. Let $t > 0$ and $h > 0$. Then*

$$P\{Y > t+h \mid Y > t\} = P\{Y > h\}\,. \tag{7.38}$$

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 7.7. The property (7.38) of an exponential distribution is referred to as the **memoryless property** of the exponential distribution. It also occurs in the geometric distribution. □

7.7 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 7.13 (Beta distribution). $\boxed{\star}$

A random variable Y has a **beta probability distribution** with parameters $\alpha > 0$ and $\beta > 0$ if it has density function

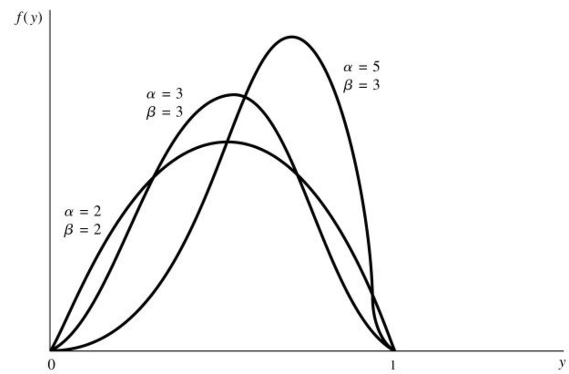
(7.39)
$$f_Y(y) \;=\; \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\,, & \text{if } 0 \le y \le 1\,, \\ 0\,, & \text{else}\,, \end{cases}$$

where

(7.40)
$$B(\alpha,\beta) \;=\; \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy \;=\; \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}\,.$$

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 7.13. $\boxed{\star}$

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 ■

7.8 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 7.14. $\boxed{\star}$

Let Y, Z be continuous or discrete random variables and $a > 0$. Assume further that $Y \geq 0$. Then

$$P\{Y \geq a\}) \leq \frac{E[Y]}{a}, \tag{7.41}$$

$$P\{|Z| \geq a\}) \leq \frac{E[|Z|^n]}{a^n}. \tag{7.42}$$

(7.41) *is known as the **Markov inequality***

PROOF of (7.41): [24] 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 (7.41).
PROOF of (7.42): Since $|Z|^n \geq 0$ and $a^n > 0$, we can apply (7.41) 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 (7.42). ■

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 7.15 (Tchebysheff Inequalities)**.**

Let Y be a random variable with mean $\mu = E[Y]$ and standard deviation $\sigma = \sqrt{Var[Y]}$. Let $k > 0$. Then

$$P\{|Y - \mu| \geq k\sigma\} \leq \frac{1}{k^2}, \tag{7.43}$$

$$P\{|Y - \mu| < k\sigma\} \geq 1 - \frac{1}{k^2}. \tag{7.44}$$

Both (7.43) *and* (7.44) *are known as the **Tchebysheff inequalities***

[24] Source: https://en.wikipedia.org/wiki/Markov%27s_inequality

PROOF: We apply (7.42) 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 (7.43). Since the event $\{|Y - \mu| \,<\, k\sigma\}$ is the complement of the event $\{|Y - \mu| \,\geq\, k\sigma\}$, (7.44) follows. ■

Remark 7.8. 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 7.15.

(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

$$P\{|Y - \mu| \,<\, a\} \,\geq\, 1 - \frac{Var[Y]}{a^2} \qquad \text{and} \qquad P\{|Y - \mu| \,\geq\, a\} \,\leq\, \frac{Var[Y]}{a^2}\,. \;\square$$

7.9 More on the Uniform Probability Distribution

Theorem 7.16.

Assume that Y is a continuous random variable with CDF $F_Y(y)$. Let $U := F_Y(Y)$. Then $U \sim$ uniform$(0,1)$.

PROOF: $\boxed{\star}$ The proof given here follows that of Theorem 2.1.10 in Casella, Berger [1], but it gives additional detail.

Let $0 < p < 1$ and let

(A) $$G(p) \;:=\; \min\{y \in \mathbb{R} : F_Y(y) \geq p\}\,.$$

In other words, $G(p)$ is the pth quantile ϕ_p for the random variable Y. Since G is nondecreasing,

(B) $$F_U(p) \;= P\{U \leq p\} \;=\; P\{F_Y(Y) \leq p\} \;=\; P\{G\big(F_Y(Y)\big) \leq G(p)\}\,.$$

The most difficult part of the proof is to show that

(C) $$P\{G\big(F_Y(Y)\big) \leq G(p)\} \;=\; P\{Y \leq G(p)\}\,.$$

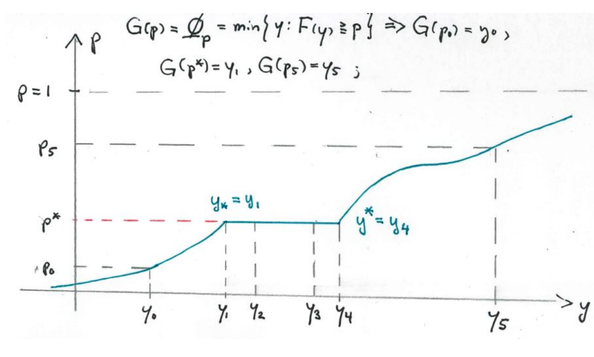
We consider two different cases.

- **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 $F(y) = p$.



7.1 (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 7.17.

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 7.1 (Properties of a Cumulative Distribution Function) on p.87:
• *F is nondecreasing* • $F(-\infty) := \lim\limits_{y\to-\infty} F(y) = 0$ • $F(\infty) := \lim\limits_{y\to\infty} F(y) = 1$

$$\text{(7.45)} \qquad \textit{Let} \quad G:[0,1]\to\mathbb{R}; \quad p\mapsto G(p) := \min\{y\in\mathbb{R}: F(y)\geq p\}\,.$$

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 $y\in\mathbb{R}$.

PROOF: $\boxed{\star}$

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 7.4 on p.90.)
- 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 7.1 on p.104 (non–injective, continuous CDF) in the proof of Theorem 7.16.

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

$$G\big(U(\omega)\big) \leq y \Leftrightarrow \big[G(p')\leq G(p)\big] \overset{\textbf{(d)}}{\Leftrightarrow} \big[p'\leq p\big] \Leftrightarrow \big[U(\omega)\leq F(y)\big]\,.$$

We take probabilities and obtain, since $U\sim$ uniform$(0,1)$ implies $P\{U\leq\tilde p\} = \tilde p$ for $0\leq\tilde p\leq 1$,

$$F_Z(y) = P\{G(U)\leq y\} = P\{U\leq F(y)\} = F(y)\,.$$

To summarize, we have shown that $F_Z(y)=F(y)$ for all $y\in\mathbb{R}$. ■

Remark 7.9. A special case of Theorem **??**an 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. □

8 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).

8.1 Multivariate CDFs, PMFs and PDFs

Assumption 8.1 (Comma separation denotes intersection)**.** We will follow the following convention for the notation of events that are generated by random variables or random items $X, Y, Z \dots$

Separating commas are to be interpreted as "**and**" and not as "**or**". Thus, for example,

$$\begin{aligned}\{X \in B,\, Y = \alpha,\, 5 \leq Z < 8\} &= \{X \in B \textbf{ and } Y = \alpha \textbf{ and } 5 \leq Z < 8\} \\ &= \{X \in B\} \cap \{Y = \alpha\} \cap \{5 \leq Z < 8\}\,. \ \square\end{aligned}$$

Definition 8.1 (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

$$F(y_1, y_2) := F_{Y_1,Y_2}(y_1, y_2) := P(Y_1 \leq y_1, Y_2 \leq y_2)\,, \qquad \text{where } y_1, y_2 \in \mathbb{R}\,, \tag{8.1}$$

the **joint cumulative distribution function** or **bivariate cumulative distribution function** or **joint CDF** or **joint distribution function** of Y_1 and Y_2. $\square$

The following theorem has been copied verbatim from the WMS text.

Theorem 8.1.

Let Y_1 and Y_2 random variables with joint CDF $F_{Y_1,Y_2}(y_1, y_2)$. then

(1) $F_{Y_1,Y_2}(-\infty, -\infty) = F_{Y_1,Y_2}(-\infty, y_2) = F_{Y_1,Y_2}(y_1, -\infty) = 0.$

(2) $F_{Y_1,Y_2}(\infty, \infty) = 1.$

(3) *If $y_1^* \geq y_1$ and $y_2^* \geq y_2$, then*

$$F_{Y_1,Y_2}(y_1^*, y_2^*) - F_{Y_1,Y_2}(y_1^*, y_2) - F_{Y_1,Y_2}(y_1, y_2^*) + F_{Y_1,Y_2}(y_1, y_2) \geq 0\,.$$

PROOF: Part **(3)** follows because

$$\begin{aligned}&F_{Y_1,Y_2}(y_1^*, y_2^*) - F_{Y_1,Y_2}(y_1^*, y_2) - F_{Y_1,Y_2}(y_1, y_2^*) + F_{Y_1,Y_2}(y_1, y_2) \\ &= P(y_1 \leq Y_1 \leq y_1^*, y_2 \leq Y_2 \leq y_2^*) \geq 0\,. \ \blacksquare\end{aligned}$$

Definition 8.2 (Joint probability mass function)**.**

Let Y_1 and Y_2 be discrete random variables. We call

(8.2) $$p(y_1, y_2) := p_{Y_1,Y_2}(y_1, y_2) := P\{Y_1 = y_1, Y_2 = y_2\}, \qquad \text{where } y_1, y_2 \in \mathbb{R},$$

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 finite or countably many pairs of values (y_1, y_2). Again we have 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 8.1 (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)*** $p_{Y_1,Y_2}(y_1, y_2) \geq 0$ *for all* $y_1, y_2 \in \mathbb{R}$,
- ***(2)*** $\sum_{y_1,y_2} p_{Y_1,Y_2}(y_1, y_2) = 1$.
- ***(3)*** $F_{Y_1,Y_2}(y_1, y_2) = \sum_{u_1\leq y_1,\, u_2\leq y_2} p_{Y_1,Y_2}(u_1, u_2) = \sum_{u_1\leq y_1} \sum_{u_2\leq y_2} p_{Y_1,Y_2}(u_1, u_2)$.

PROOF: Obvious. ■

Definition 8.3 (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 8.2 (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 items $X, Y, Z \ldots$

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

$$
\begin{aligned}
F_{Y_1,Y_2}(y_1,y_2) &= \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f_{Y_1,Y_2}(u_1,u_2)\,du_2\,du_1 \\
&= \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f_{Y_1,Y_2}(u_1,u_2)\,du_1\,du_2\,. \\
&= \iint_{]-\infty,y_1\times]-\infty,y_2]} f_{Y_1,Y_2}(u_1,u_2)\,du_1\,du_2\,. \quad \square
\end{aligned}
\tag{8.3}
$$

Definition 8.4 (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. $\square$

Theorem 8.2.

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) $\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1,y_2)dy_1dy_2 = 1.$

PROOF: An easy consequence of Theorem 8.1 on p.107. ■

8.2 Marginal and Conditional Probability Distributions

Given two random variables $Y_1, Y_2 : (\Omega, P) \to \mathbb{R}$, one obtains the marginal distribution of Y_1 by considering, for $B_1, B_2 \subseteq \mathbb{R}$,

$$
\begin{aligned}
P_{Y_1}(B_1) &= P\{Y_1 \in B_1\} \;=\; P\{Y_1 \in B_1,\ Y_2 \in \Omega\} \\
\text{instead of} \quad P_{Y_1,Y_2}(B_1,B_2) &= P\{Y_1 \in B_1,\ Y_2 \in B_2\}
\end{aligned}
$$

and the marginal distribution of Y_2 by considering

$$
\begin{aligned}
P_{Y_2}(B_2) &= P\{Y_2 \in B_2\} \;=\; P\{Y_1 \in \Omega,\ Y_2 \in B_2\} \\
\text{instead of} \quad P_{Y_1,Y_2}(B_1,B_2) &= P\{Y_1 \in B_1,\ Y_2 \in B_2\}\,.
\end{aligned}
$$

In other words, the marginal distribution of Y_1 simply is the distribution of Y_1 and the marginal distribution of Y_2 simply is the distribution of Y_2. The word "marginal" is only used to emphasize that those distributions are considered in the context of the joint distribution of Y_1 and Y_2.

The above 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 8.5 (Marginal distributions)**.**

(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

$$p_{Y_1}(y_1) \;=\; \sum_{\text{all } y_2} p_{Y_1,Y_2}(y_1, y_2) \quad \text{and} \quad p_{Y_2}(y_2) \;=\; \sum_{\text{all } y_1} p_{Y_1,Y_2}(y_1, y_2) \tag{8.4}$$

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

$$f_{Y_1}(y_1) \;=\; \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \quad \text{and} \quad f_{Y_2}(y_2) \;=\; \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1\,. \tag{8.5}$$

the **marginal density functions** or **marginal PDFs** of Y_1 and Y_2. $\square$

Remark 8.1. We recall Definition 3.9 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 3.2 on p.50. We replace the general events A and B with events $\{Y_1 = y_1\}$ and $\{Y_2 = y_2\}$ and obtain, if $P\{Y_2 = y_2\} > 0$,

$$P\{Y_1 = y_1 \mid Y_2 = y_2\} \;=\; \frac{P\{Y_1 = y_1,\, Y_2 = y_2\}}{P\{Y_2 = y_2\}}\,. \tag{8.6}$$

As we always do for conditional probabilities, we interpret (8.6) 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 (8.6) 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. $\square$

Definition 8.6 (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

$$p_{Y_1|Y_2}(y_1 \mid y_2) \;:=\; \begin{cases} P\{Y_1 = y_1 \mid Y_2 = y_2\}\,, & \text{if } P\{Y_2 = y_2\} > 0\,, \\ \textbf{undefined}\,, & \text{if } P\{Y_2 = y_2\} = 0\,, \end{cases} \tag{8.7}$$

the **conditional probability mass function** or the **conditional PMF** of Y_1 given Y_2.

Likewise, we call

$$p_{Y_2|Y_1}(y_2 \mid y_1) \;:=\; \begin{cases} P\{Y_2 = y_2 \mid Y_1 = y_1\}\,, & \text{if } P\{Y_1 = y_1\} > 0\,, \\ \textbf{undefined}\,, & \text{if } P\{Y_1 = y_1\} = 0\,, \end{cases} \tag{8.8}$$

the **conditional PMF** of Y_2 given Y_1. ☐

Remark 8.2. Note that conditional PMFs can be expressed in terms of joint PMF and marginal PMFs:

$$p_{Y_1|Y_2}(y_1 \mid y_2) \;=\; \frac{p_{Y_1,Y_2}(y_1,y_2)}{p_{Y_2}(y_2)} \quad \text{if } p_{Y_2}(y_2) > 0\,, \tag{8.9}$$

$$p_{Y_2|Y_1}(y_2 \mid y_1) \;=\; \frac{p_{Y_1,Y_2}(y_1,y_2)}{p_{Y_1}(y_1)} \quad \text{if } p_{Y_1}(y_1) > 0\,. \;\square \tag{8.10}$$

The author does not think that there is much use for the next definition (WMS Ch.05.3, Definition 5.6) because all jointly continuous random variables come with PDF

$$f_{Y_1,Y_2}(y_1,y_2) \;=\; \frac{\partial^2 F_{Y_1,Y_2}}{\partial y_1 \partial y_2}(y_1,y_2)\,.$$

It is included only for the sake of completeness.

Definition 8.7. ⋆ Let Y_1 and Y_2 be two jointly continuous random variables. Then

$$F_{Y_1|Y_2}(y_1 \mid y_2) \;:=\; P(Y_1 \le y_1 \mid Y_2 = y_2) \tag{8.11}$$

defines the **conditional distribution function** or **conditional CDF** of Y_1 given $Y_2 = y_2$. ☐

Definition 8.8 (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

$$f_{Y_1|Y_2}(y_1 \mid y_2) \;:=\; \begin{cases} \dfrac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)}\,, & \text{if } f_{Y_2}(y_2) > 0\,, \\ \textbf{undefined}\,, & \text{if } f_{Y_2}(y_2) = 0\,, \end{cases} \tag{8.12}$$

the **conditional probability density function** or the **conditional PDF** of Y_1 given Y_2.

Likewise, we call

$$f_{Y_2|Y_1}(y_2 \mid y_1) \;:=\; \begin{cases} \dfrac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)}\,, & \text{if } f_{Y_1}(y_1) > 0\,, \\ \textbf{undefined}\,, & \text{if } f_{Y_1}(y_1) = 0\,, \end{cases} \tag{8.13}$$

the **conditional PDF** of Y_2 given Y_1. ☐

8.3 Independence of Random Variables and Discrete Random Items

Introduction 8.1. Let $X_1, X_2 : (\Omega, P) \to \Omega'$ be two random items (recall that they are random variables if $\Omega' = \mathbb{R}$). 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

$$(8.14)\qquad P\{X_1 \in B_1,\ X_2 \in B_2,\} \;=\; P\{X_1 \in B_1\} \cdot P\{\, X_2 \in B_2,\} \quad \text{for all relevant } B_1, B_2 \subseteq \Omega'.$$

What are the relevant sets B_j? We answer that question for discrete random items (hence, also for discrete random variables) and for continuous random variables.

(a) Assume that $X : (\Omega, P) \to \Omega'$ is a discrete random item with PMF $p_X(x)$. In other words, there is a countable $\Omega^* \subseteq \Omega'$ such that, for any $B \subseteq \Omega'$,

$$P\{X \in B\} \;=\; P_X(B) \;=\; \sum_{x \in \Omega^* \cap B} p_X(x) \;=\; \sum_{x \in B} p_X(x) \;=\; \sum_{x \in B} P\{X = x\}$$

(See (6.5) in Remark 6.1 on p.72.) 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 $\{a < Y \leq b\}$ where $a < b$, are

$$P\{a < Y \leq b\} \;=\; \int_a^b f_Y(y)dy\,.$$

(See (7.4) in heorem 7.2 on p.89.) 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 $B =]a, b]$. [25]

In summary, we could define independence of discrete random items 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

$$(8.15)\qquad p_{X_1,X_2}(x_1, x_2) \;=\; p_{X_1}(x_1) \cdot p_{X_2}(x_2) \qquad \text{for all } x_1, x_2 \in \Omega'.$$

Moreover, independence of continuous random variables Y_1 and Y_2 could be defined as

$$P\{a < X_1 \leq b,\, c < X_2 \leq d\} \;=\; P\{a < X_1 \leq b\} \cdot P\{c < X_2 \leq d\} \quad \text{for all } a < b \text{ and } c < d.$$

Equivalently, this can be expressed as

$$(8.16)\qquad \int_a^b \int_c^d f_{Y_1,Y_2}(y_1, y_2)dy_2\, dy_1 \;=\; \int_a^b f_{Y_1}(y_1)dy_1 \cdot \int_c^d f_{Y_2}(y_2)dy_2 \quad \text{for all } a < b \text{ and } c < d.$$

[25] 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 7.1 on p.88.

The CDF (cumulative distribution function) $F_Y(y)$ gives us for both discrete and continuous random variables (but we must exclude discrete random items) 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 $F_Y(y-) = \lim\limits_{a<y, a\to y} F_Y(a)$ 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

(8.17) $$F_{Y_1,Y_2}(y_1, y_2) \;=\; F_{Y_1}(y_1) \cdot F_{Y_2}(y_2) \quad \text{for all } y_1, y_2 \in \mathbb{R}. \;\square$$

We make (8.17) the basis for the definition of independence of random variables.

Definition 8.9 (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

(8.18) $$F_{Y_1,Y_2}(y_1, y_2) \;=\; F_{Y_1}(y_1) \cdot F_{Y_2}(y_2) \quad \text{for all } y_1, y_2 \in \mathbb{R}.$$

If Y_1 and Y_2 are not independent, we call them **dependent**.

We must treat discrete random items separately since there are no CDFs.

Let X_1 and X_2 be discrete random items 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

(8.19) $$p_{X_1,X_2}(x_1, x_2) \;=\; p_{X_1}(x_1) \cdot p_{X_2}(x_2) \quad \text{for all } x_1, x_2 \in \mathbb{R}.$$

If X_1 and X_2 are not independent, we call them **dependent**. $\square$

Theorem 8.3 (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

(8.20) $$Y_1, Y_2 \textit{ are independent} \quad \Leftrightarrow \quad p_{Y_1,Y_2}(y_1, y_2) \;=\; p_{Y_1}(y_1) \cdot p_{Y_2}(y_2) \;\; \textit{for all } y_1, y_2 \in \mathbb{R}.$$

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

(8.21) $$Y_1, Y_2 \textit{ are independent} \quad \Leftrightarrow \quad f_{Y_1,Y_2}(y_1, y_2) \;=\; f_{Y_1}(y_1) \cdot f_{Y_2}(y_2) \;\; \textit{for all } y_1, y_2 \in \mathbb{R}.$$

PROOF: Omitted. ■

The next theorem will be generalized in Theorem 8.9 on p.117. There Y_1 and Y_2 will be replaced with functions $g(Y_1)$ and (Y_2).

Theorem 8.4.

If Y_1 and Y_2 are independent random variables, then

$$E[Y_1 \cdot Y_2] \;=\; E[Y_1] \cdot E[Y_2]\,. \tag{8.22}$$

PROOF: We show the proof for continuous Y_1 and Y_2. Since $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2)$,

$$\begin{aligned}
E[Y_1Y_2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2 f_{Y_1,Y_2}(y_1,y_2)\,dy_1\,dy_2 \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1y_2 f_{Y_1}(y_1)\,f_{Y_2}(y_2)\,dy_1\,dy_2 \\
&= \int_{-\infty}^{\infty} y_2\left[\int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1)\,dy_1\right] f_{Y_2}(y_2)\,dy_2 \;=\; \int_{-\infty}^{\infty} y_2 E[Y_1] f_{Y_2}(y_2)\,dy_2 \\
&= E[Y_1]\int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2)\,dy_2 \;=\; E[Y_1]\,E[Y_2]\,.
\end{aligned}$$

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 8.5 (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

$$f_{Y_1,Y_2}(y_1,y_2) > 0 \quad \Leftrightarrow \quad a \le y_1 \le b \;\; \text{and} \;\; c \le y_2 \le d\,.$$

$$\text{Then} \quad Y_1, Y_2 \text{ are independent} \quad \Leftrightarrow \quad f_{Y_1,Y_2}(y_1,y_2) \;=\; g_1(y_1) \cdot g_2(y_2) \tag{8.23}$$

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: Omitted. ■

8.4 The Expected Value of a Function of Several Random Variables

In this section we must work with vectors $(x_1, x_2, \dots, x_k)$ of fixed,but arbitrary dimension k, where each component x_j is a real number and thus, $(x_1, x_2, \dots, x_k) \in \mathbb{R}^k$. Since it is extremely space consuming to repeatedly write such lengthy objects, we remind you of the "arrow notation" that was introduced in Example 2.10 on p.27.

Notation 8.1 (Arrow notation for vectors)**.**

- We write $\vec{x}$ as an abbreviation for a vector $(x_1, x_2, \dots, x_n)$. 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

$$\int_A f(\vec{x})\, d\vec{x} \;=\; \int_{a_1}^{b_1} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n)\, dy_1 dy_2 \cdots dy_n$$

  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 $a_1 < b_1, a_2 < b_2, \dots, a_n < b_n$ for some $n \in \mathbb{N}$. Then $\vec{y} \in]a_1, b_1] \times \cdot \times]a_d, b_d]$ denotes the following: $\vec{y} = (y_1, y_2, \dots, y_d)$ and $a_i < y_i \le b_i$ for $i = 1, \dots, d$.

Here are some examples.

(a) $\vec{z} \in \mathbb{R}^m$ means: $\vec{z} = (z_1, z_2, \dots, z_m)$ and $z_j \in \mathbb{R}$ for all j.
(b) If $f : \mathbb{R}^k \to \mathbb{R}$, then $g(\vec{y})$ means: $f(y_1, \dots, y_k)$.
(c) If $g : \mathbb{R}^d \to \mathbb{R}$, then $g(\vec{Y})$ means: $g(Y_1, \dots, Y_d)$; $g(\vec{Y}(\omega))$ means: $g(Y_1(\omega), \dots, Y_d(\omega))$.
(d) If $\psi : \mathbb{R}^n \to \mathbb{R}$, then $E\left[\psi(\vec{Y})\right]$ means: $E\left[\psi(Y_1, \dots, Y_n)\right]$.

Definition 8.10 (Expected value of $g(\vec{Y})$)**.**

(a) Let $k \in \mathbb{N}$ and let $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ 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 $y_1, y_2, \dots, y_k$. Then

$$E\big[g(\vec{Y})\big] \;=\; E\big[g(Y_1, Y_2, \dots, Y_k)\big] \;:=\; \sum_{y_1,\, y_2,\, \dots,\, y_k} \!\!\cdots \sum g(\vec{y})\, p_{\vec{Y}}(\vec{y}) \tag{8.24}$$

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 $\vec{y} = (y_1, y_2, \dots, y_k)$ for which $p_{\vec{Y}}(\vec{y}) \neq 0$.

(b) Let $k \in \mathbb{N}$ and let $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ 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 $y_1, y_2, \dots, y_k$. Then

$$E\big[h(\vec{Y})\big] \;=\; E\big[h(Y_1, Y_2, \dots, Y_k)\big] \;:=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\vec{y})\, f_{\vec{Y}}(\vec{y}) d\vec{y} \tag{8.25}$$

is called the **expected value** or **mean** of the random variable $g(\vec{Y})$.

See Notations 8.1 (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 $\sum\cdots\sum |g(y_1,\dots,y_k)|\, p(y_1,\dots,y_k)$ is finite and that $E\big[h(\vec{Y})\big]$ exists if $\int\cdots\int |g(y_1,\dots,y_k)|\, f(y_1,\dots,y_k)\, dy_1\dots dy_k$ is finite. The functions g and h we deal with in this course will always satisfy that assumption. $\square$

Example 8.1. 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 8.10,

$$E[h(\vec{Y})] \;=\; \int_{\mathbb{R}^n}(y_1+\cdots+y_n)f_{\vec{Y}}(\vec{y})d\vec{y} \;=\; \sum_{j=1}^{n}\int_{\mathbb{R}^n} y_j f_{\vec{Y}}(\vec{y})d\vec{y}.$$

Let $\vec{\tilde{y}} := \big(y_1,\dots,y_{j-1},y_{j+1},\dots,y_n\big)$. 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 $E[Y_j] = \int\limits_{-\infty}^{\infty} y_j f_{Y_j} dy_j$ 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 $Y_1, Y_2,\dots$ are random variables on a common probability space (Ω, P)

Theorem 8.6 (WMS Ch.05.6, Theorem 5.6)**.**

$$c\in\mathbb{R} \;\;\Rightarrow\;\; E[c] \;=\; c\,. \tag{8.26}$$

PROOF: Trivial. ■

Theorem 8.7 (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

$$E[cg(Y_1,Y_2)] \;=\; cE[g(Y_1,Y_2)]\,. \tag{8.27}$$

PROOF: Trivial. ■

Theorem 8.8 (WMS Ch.05.6, Theorem 5.8)**.**

Let $g_1, g_2, \dots, g_k : \mathbb{R}^2 \to \mathbb{R}$ *Then the random variables* $g_j(Y_1, Y_2)$ $(j = 1, \dots, k)$ *satisfy*

$$\begin{aligned} E[g_1(Y_1,Y_2) + g_2(Y_1,Y_2) + \cdots + g_k(Y_1,Y_2)] \\ = E[g_1(Y_1,Y_2)] + E[g_2(Y_1,Y_2)] + \cdots + E[g_k(Y_1,Y_2)]\,. \end{aligned} \tag{8.28}$$

PROOF: Omitted. ■

The next theorem generalizes Theorem 8.4 on p.114. That one stated that, for independent random variables, the expectation of the product is the product of the expectations.

Theorem 8.9.

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*

$$E[g(Y_1)\,h(Y_2)] \;=\; E[g(Y_1)]\,E[h(Y_2)]\,. \tag{8.29}$$

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

$$f_{Y_1,Y_2}(y_1, y_2) \;=\; f_{Y_1}(y_1)\, f_{Y_2}(y_2)\,.$$

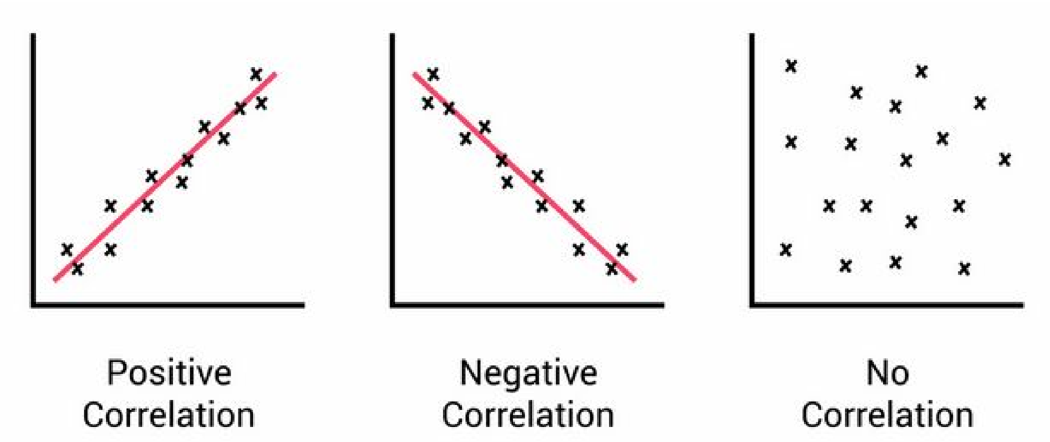
The product $g(Y_1)h(Y_2)$ is a function $\varphi(Y_1, Y_2)$ of Y_1 and Y_2. Hence, by Definition 8.10 (Expected value of $g(\vec{Y})$) on p.115,

$$\begin{aligned} E[g(Y_1)h(Y_2)] \;&=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f_{Y_1,Y_2}(y_1,y_2)\,dy_2\,dy_1 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2)f_{Y_1}(y_1)\,f_{Y_2}(y_2)\,dy_2\,dy_1 \\ &= \int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\,\left[\int_{-\infty}^{\infty} h(y_2)f_{Y_2}(y_2)\,dy_2\right]\,dy_1 \\ &= \int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\,E[h(Y_2)]\,dy_1 \\ &= E[h(Y_2)]\int_{-\infty}^{\infty} g(y_1)f_{Y_1}(y_1)\,dy_1 \;=\; E[g(y_1)]\,E[h(Y_2)]\,. \end{aligned}$$ ■

8.5 The Covariance of Two Random Variables

Introduction 8.2. 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. $\square$

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 8.11 (Covariance)**.**

The **covariance** of Y_1 and Y_2 is

$$Cov[Y_1, Y_2] = E\big[(Y_1 - E[Y_1])\,(Y_2 - E[Y_2])\big] = E\big[(Y_1 - \mu_1)\,(Y_2 - \mu_2)\big]. \quad \square \tag{8.30}$$

Remark 8.3. $Cov[Y_1, Y_2]$ 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. $\square$

Definition 8.12 (Correlation coefficient)**.**

The **correlation coefficient**, of Y_1 and Y_2 is

$$\rho \;=\; \frac{Cov(Y_1, Y_2)}{\sigma_1\sigma_2} \quad \square \tag{8.31}$$

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 8.2. *The correlation coefficient satisfies the inequality*

$$-1 \;\leq\; \rho \;\leq\; 1 \quad \square \tag{8.32}$$

PROOF: Omitted ■

The next formula often makes it easier to compute the covariance.

Theorem 8.10.

$$Cov[Y_1, Y_2] \;=\; E[(Y_1 - \mu_1)\,(Y_2 - \mu_2)] \;=\; E[Y_1Y_2] \;-\; E[Y_1]\,E[Y_2]\,. \tag{8.33}$$

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,

$$\begin{aligned}
Cov[Y_1, Y_2] &= E[(Y_1 - \mu_1)\,(Y_2 - \mu_2)] \\
&= E(Y_1Y_2 \;-\; \mu_1Y_2 \;-\; \mu_2Y_1 \;+\; \mu_1\mu_2) \\
&= E[Y_1Y_2] \;-\; \mu_1E[Y_2] \;-\; \mu_2E[Y_1] \;+\; \mu_1\mu_2 \\
&= E[Y_1Y_2] \;-\; \mu_1\mu_2 \;-\; \mu_2\mu_1 \;+\; \mu_1\mu_2 \;=\; E[Y_1Y_2] \;-\; \mu_1\mu_2\,. \quad \blacksquare
\end{aligned}$$

Theorem 8.11.

Independent random variables are uncorrelated.

PROOF: By Theorem 8.4 on p.114, independent random variables Y_1 and Y_2 satisfy $E[Y_1Y_2] = E[Y_1]E[Y_2]$. Together with (8.33), we obtain

$$Cov[Y_1, Y_2] = E[Y_1Y_2] - E[Y_1]\,E[Y_2] = 0\,. \blacksquare$$

Example 8.2 (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

	$\mathbf{Y_2}$	
$\mathbf{Y_1}$	-1	1
-1	0	$1/4$
0	$1/2$	0
1	0	$1/4$

$E[Y_1] = (-1)\frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 0$,
$E[Y_2] = (-1)\frac{1}{2} + 1 \cdot \frac{1}{2} = 0$,
$E[Y_1Y_2] = (-1)(-1)0 + 0(-1)\frac{1}{2} + (1)(-1)0$
$\quad + (-1)(1)\frac{1}{4} + 0 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot \frac{1}{4} = 0$.

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 8.13 (Linear function)**.** $\boxed{\star}$

Let $n \in \mathbb{N}$. We call a function $\varphi : \mathbb{R}^n \to \mathbb{R}$; $\vec{x} = (x_1, \dots, x_n) \mapsto \varphi(\vec{x})$, a **linear function**, if there are constants $a_1, \dots, a_n \in \mathbb{R}$ such that

$$\varphi(\vec{x}) = a_1x_1 + a_2x_2 + \dots + a_nx_n = \sum_{j=1}^{n} a_jx_j\,. \quad \square \tag{8.34}$$

Remark 8.4. Note that if $\vec{Y} = (Y_1, \dots, Y_n)$ is a vector of random variables, then the function φ of (8.34) defines a random variable $V = \varphi(\vec{Y}) = \sum_{j=1}^{n} a_jY_j$. $\square$

Theorem 8.12 (WMS Ch.05.8, Theorem 5.12)**.** *Let $\vec{X} = X_1, \dots, X_m$ and $\vec{Y} = Y_1, \dots, Y_n$ be random variables on a probability space (Ω, P). For $i = 1, \dots, m$ and $j = 1, \dots, n$, let $\xi_i := E(X_i)$ and $\mu_j := E(Y_j)$. Further, let*

$$U := \sum_{i=1}^{m} a_iX_i \qquad \text{and} \qquad V := \sum_{j=1}^{n} b_jY_j\,,$$

where $\vec{a} = (a_1, a_2, \dots, a_m)$ and $\vec{b} = (b_1, b_2, \dots, b_m)$ are two constant vectors. Then

$$E[U] = \sum_{i=1}^{m} a_i\xi_i\,, \tag{8.35}$$

$$Var[U] = \sum_{i=1}^{m} a_i^2 Var[X_i] + 2\sum\sum_{1\leq i<j\leq m} a_ia_jCov[X_i, X_j]\,. \tag{8.36}$$

$$Cov[U, V] = \sum_{i=1}^{m}\sum_{j=1}^{n} a_ib_jCov[X_i, Y_j]\,. \tag{8.37}$$

In (8.36), $\sum\sum_{1\le i<j\le m} \cdots$ *refers to summation over all pairs* (i,j) *satisfying* $i<j$.

PROOF: The theorem consists of three parts, of which (8.35) follows directly from Theorems 8.7 and 8.8.

Proof of (8.36): From the definition of variance we obtain

$$\begin{aligned} Var[U] &= E[U - E[U]]^2 \;=\; E\left[\sum_{i=1}^n a_i X_i \,-\, \sum_{i=1}^n a_i\xi_i\right]^2 \;=\; E\left[\sum_{i=1}^n a_i(X_i \,-\, \xi_i)\right]^2 \\ &= E\left[\sum_{i=1}^n a_i^2 (X_i - \xi_i)^2 \,+\, \sum_{i=1}^n \sum_{\substack{j=1 \\ i\ne j}}^n a_i a_j (X_i - \xi_i)(X_j - \xi_j)\right] \\ &= \sum_{i=1}^n a_i^2 E[X_i - \xi_i]^2 \,+\, \sum_{i=1}^n \sum_{\substack{j=1 \\ i\ne j}}^n a_i a_j E[(X_i - \xi_i)(X_j - \xi_j)]\,. \end{aligned}$$

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,

$$Var[U] \;=\; \sum_{i=1}^n a_i^2 Var[X_i] \,+\, \sum_{i=1}^n \sum_{\substack{j=1 \\ i\ne j}}^n a_i a_j Cov[X_i, X_j]\,.$$

We apply symmetry $Cov[X_i, X_j] = Cov[X_j, X_i]$ to the double summation and obtain

$$Var[U] \;=\; \sum_{i=1}^n a_i^2 Var[X_i] \,+\, 2\sum\sum_{1\le i<j\le n} a_i a_j Cov[X_i, X_j]\,.$$

We have shown (8.36). To prove (8.37), we proceed in a similar fashion: We have

$$\begin{aligned} Cov[U,V] &= E\big[\,(U - E[U])(V - E[V])\,\big] \\ &= E\left[\left(\sum_{i=1}^n a_i X_i \,-\, \sum_{i=1}^n a_i\xi_i\right)\left(\sum_{j=1}^m b_j X_j \,-\, \sum_{j=1}^m b_j\xi_j\right)\right] \\ &= E\left[\left(\sum_{i=1}^n a_i(X_i \,-\, \xi_i)\right)\left(\sum_{j=1}^m b_j(Y_j \,-\, \mu_j)\right)\right] \\ &= E\left[\sum_{i=1}^n \sum_{j=1}^m a_i b_j (X_i \,-\, \xi_i)(Y_j \,-\, \mu_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j E[(X_i \,-\, \xi_i)(Y_j \,-\, \mu_j)] \;\blacksquare \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j Cov[X_i, Y_j)\,. \end{aligned}$$

Remark 8.5. Note the following about Theorem 8.12:

(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 $Cov[Y_i, Y_i] = Var[Y_i]$, (8.36) is a particular version of (8.37). $\square$

We are now in a position to prove (7.19) of Theorem 7.6 on p.93 (and thus, also (6.13) of Theorem 6.5 on p.75) 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 8.11.)

Corollary 8.1 (Bienaymé formula for uncorrelated variables). $\boxed{\star}$

Let $Y_1, Y_2, \dots, Y_n : \Omega \to \mathbb{R}$ be uncorrelated random variables (which all are defined on the same probability space (Ω, P) ($n \in \mathbb{N}$. Then

$$Var\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var[Y_j]\,. \tag{8.38}$$

PROOF: Since $Y_1, \dots, Y_n$ are uncorrelated, $Cov[Y_i, Y_j] = 0$ for $1 \le i, j \le n$ and $i \ne j$. We employ (8.36) on p.120 with $a_1 = a_2 = \cdots = a_n = 1$ and obtain

$$Var\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} Var[Y_i] + 2\sum\sum_{1\le i<j\le n} Cov[Y_i, Y_j] = \sum_{i=1}^{n} Var[Y_i] + 0\,. \blacksquare$$

Example 8.3 (Variance of the sample mean [26]). This example belongs thematically to Section 5.2 (Random 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 $\vec{Y}(\omega) = \big(Y_1(\omega), \dots, Y_n(\omega)$ of n independent random variables $Y_1, \dots Y_n$ which have the same distribution as Y. In other words, the Y_j are a (finite) iid sequence in the sense of Definition 6.5 on p.76.
- 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 5.3 on p.70 and 5.4 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. [27]

In this example we specifically consider the mean of the population data.

[26] This is a modified version of WMS, Example 5.27.
[27] See parts **(c)** and **(d)** of Remark 5.2 on p.69.

- 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 $Y_1, \dots Y_n$ 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 $Y_1, Y_2, \dots, Y_n$ be independent random variables with common expectation $E[Y_j] = \mu$ and variance $Var[Y_j] = \sigma^2$ $(j = 1, \dots, n)$. Let

$$\bar{Y} \;:=\; \frac{1}{n}\sum_{j=1}^{n} Y_j\,. \tag{8.39}$$

It follows from (8.35) on p.120 and Corollary 8.1 on p.122 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 $\sqrt{K}$. $\square$

8.6 Conditional Expectations and Conditional Variance

8.6.1 The Conditional Expectation With Respect to an Event $\boxed{\star}$

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 3.2 on p.50 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. [28]

Assumption 8.3.

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

$$A \mapsto Q(A) := P(A \mid B), \quad \text{where } A \subseteq \Omega. \tag{8.40}$$

The symbols $X, X_1, X_2, \dots$ denote random items and $X, X_1, X_2, \dots$ denote random variables on Ω. We need not be specific about whether we mean (Ω, P) or (Ω, Q), because the definition of random item and random variable does not involve the probability measure, only the carrier space Ω. $\square$

Remark 8.6. 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

[28] 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.

BEWARE: The above does not apply to cases where a specific probability measure is considered. An example for this would be, e.g., Proposition 7.7 on p.101 (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 remar. □

Remark 8.7. 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 item [29] 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 $f_Y^Q(y)$. [30]

(d) We skip joint Q–CDFs and joint Q–PDFs and only elaborate on the joint Q–PMF. of two random items X_1 and X_2. It is, as one should expect, defined as
$p^Q_{X_1,X_2}(x_1, x_2) = Q\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1, X_2 = x_2 \mid B\}$.

(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 $E^Q[Y] = \int\limits_{-\infty}^{\infty} y \cdot f_Y^Q(y)dy$.

(g) The Q–variance of a random variable Y is $Var^Q[Y] = E^Q\big[(Y - E^Q[Y])^2\big]$.

(h) The Q–MGF of a random variable Y is $m_Y^Q(t) = E^Q\big[e^{tY}\big]$.

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 $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ of random variables. (See Definition 8.10 on p.115.)

(i) If the Y_j are discrete and $g : \mathbb{R}^k \to \mathbb{R}$, then $E^Q\big[g(\vec{Y})\big] = \sum\limits_{y_1,\, y_2,\, \dots,\, y_k}\!\!\cdots\sum g(\vec{y})\, p_{\vec{Y}}^Q(\vec{y})$.

(j) If the Y_j are continuous and $h : \mathbb{R}^k \to \mathbb{R}$, then $E^Q\big[h(\vec{Y})\big] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\vec{y})\, f_{\vec{Y}}^Q(\vec{y})d\vec{y}$. □

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 8.13.

If $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ is a vector of k discrete or Q–continuous random variables, then

$$E^Q\left[\sum_{j=1}^n Y_j\right] = \sum_{j=1}^n E^Q[Y_j]\,. \tag{8.41}$$

[29] Since $P\{X = x\} \cap B \le P\{X = x\}$, $P\{X = x\} = 0$ implies $Q\{X = x\} = 0$. Thus, any P–discrete random item also is Q–discrete.

[30] 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.

PROOF: This follows from Theorem 8.13 on p.125. ■

Theorem 8.14. *If Y is a discrete or Q–continuous random variable and $\vec{Y} = \big(Y_1, Y_2, \dots, Y_k\big)$ is a vector of k Q–independent discrete or Q–continuous random variables, then*

$$Var^Q[Y] = E^Q\left[Y^2\right] - \left(E^Q[Y]\right)^2, \tag{8.42}$$

$$Var^Q\left[aY+b\right] = a^2 Var^Q[Y], \tag{8.43}$$

$$Var^Q\left[\sum_{j=1}^n Y_j\right] = \sum_{j=1}^n Var^Q[Y_j]. \tag{8.44}$$

PROOF: This follows from Theorem 7.6 on p.93. ■

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 8.15.

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

$$\begin{aligned} & \textbf{(a)} \quad P(A_1 \cap A_2 \mid B) = P(A_1 \mid B) \cdot P(A_2 \mid B) \\ \Leftrightarrow \quad & \textbf{(b)} \quad P(A_1 \mid A_2 \cap B) = P(A_1 \mid B) \\ \Leftrightarrow \quad & \textbf{(c)} \quad P(A_2 \mid A_1 \cap B) = P(A_2 \mid B). \end{aligned} \tag{8.45}$$

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)**:

$$\begin{aligned} P(A_1 \cap A_2 \mid B) &= \frac{P(A_1 \cap A_2 \cap B)}{P(B)} = \frac{P(A_1 \cap A_2 \cap B)}{P(A_2 \cap B)} \cdot \frac{P(A_2 \cap B)}{P(B)} \\ &= P(A_1 \mid A_2 \cap B) \cdot P(A_2 \mid B) \overset{\textbf{(b)}}{=} P(A_1 \mid B) \cdot P(A_2 \mid B). \ ■ \end{aligned}$$

8.6.2 The Conditional Expectation w.r.t a Random Variable or Random Item

Remark 8.8. $\boxed{\star}$ We mentioned at the beginning of the previous subsection 8.6.1 (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 item X_1, that another random item 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 items. This will allow us to contrast discrete and continuous random variables.

The method of subsection 8.6.1 (using the probability measure $Q(A) = P\{A \mid \tilde{Y} = \tilde{y})$ 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 7.1 on p.88), 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

$$Q\{Y_1 \in B_1\} \;=\; \sum_{y_1 \in B_1} P\{Y_1 = y_1 \mid Y_2 = y_2\} \;=\; \sum_{y_1 \in B_1} p_{Y_1|Y_2}(y_1 \mid y_2)\,. \;\square$$

Definition 8.14 (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

$$E[g(Y_1) \mid Y_2 = y_2] := \sum_{y_1} g(y_1)\, p(y_1 \mid y_2) \quad \text{(discrete case)}, \tag{8.46}$$

$$E[g(Y_1) \mid Y_2 = y_2] := \int_{-\infty}^{\infty} g(y_1)\, f(y_1 \mid y_2)\, dy_1 \quad \text{(continuous case)}. \tag{8.47}$$

Then 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 8.9. 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

$$E[g(Y_1 \mid Y_2] \tag{8.48}$$

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

$$E[Z] = E\big[E[g(Y_1 \mid Y_2]\big]\,. \tag{8.49}$$

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 8.16 (WMS Ch.05.11, Theorem 5.14)**.**

Let Y_1 and Y_2 be either jointly continuous or jointly discrete random variables. Then

$$E[Y_1] = E\big[E[Y_1 \mid Y_2]\big]\,. \tag{8.50}$$

See Remark 8.9 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 8.14 to the functions $g(y_1) = y_1$ and $g(y_1) = y_1^2$.

Definition 8.15 (Conditional variance)**.**

Let Y_1 and Y_2 be two random variables which are either jointly discrete or jointly continuous. Let

$$Var[Y_1 \mid Y_2 = y_2] := E[Y_1^2 \mid Y_2 = y_2] - \big(E[Y_1 \mid Y_2 = y_2]\big)^2 . \tag{8.51}$$

(8.52)

Then we call $Var[Y_1 \mid Y_2 = y_2]$ the **conditional variance** of $g(Y_1)$, given that $Y_2 = y_2$. □

Theorem 8.17.

Let Y_1 and Y_2 be jointly discrete or jointly continuous random variables. Then

$$Var[Y_1 \mid Y_2] = E\big[(Y_1 - E[Y_1 \mid Y_2])^2 \mid Y_2\big] , \tag{8.53}$$

$$Var[Y_1] = E\big[Var[Y_1 \mid Y_2]\big] + Var\big[E[Y_1 \mid Y_2]\big] . \tag{8.54}$$

PROOF: We only give the proof of (8.54). 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,

$$\begin{aligned} Var[Y_1] &= E\big[Y_1^2\big] - \big(E[Y_1]\big)^2 \\ &= E\big[E[Y_1^2 \mid Y_2]\big] - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2 \\ &= E\big[E[Y_1^2 \mid Y_2]\big] - E\big[\big(E[Y_1 \mid Y_2]\big)^2\big] + E\big[\big(E[Y_1 \mid Y_2]\big)^2\big] - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2 \\ &= E\big[E[Y_1^2 \mid Y_2] - \big(E[Y_1 \mid Y_2]\big)^2\big] + \big\{E\big[\big(E[Y_1 \mid Y_2]\big)^2 - \big(E\big[E[Y_1 \mid Y_2]\big]\big)^2\big\} \\ &= E\big[V(Y_1 \mid Y_2)\big] + V\big[E(Y_1 \mid Y_2)\big] . \; \blacksquare \end{aligned}$$

8.7 The Multinomial Probability Distribution

Introduction 8.3. In Definition 4.3 (p.60) of Chapter 4 (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 6.2 (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 6.6 on p.76 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 $\vec{Y} = (Y_1, \dots, Y_k)$ 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 items, let us call them $X_1, \dots, X_n$, 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 8.16 (Multinomial Sequence)**.**

Let $X_1, X_2, \dots$ be a finite or infinite sequence of random items 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 8.17 (Multinomial distribution)**.**

Assume that $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ is a vector of random variables which possesses the joint probability mass function

$$p_{\vec{Y}}(y_1, y_2, \dots, y_k) \ = \ \binom{n}{y_1, \cdots, y_k} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}\,, \tag{8.55}$$

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 $p_{\vec{Y}}(y_1, y_2, \dots, y_k)$. □

Theorem 8.18.

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,

- $Y_i(\omega) = y_i \;\Leftrightarrow\; X_j(\omega) = \omega'_i$ *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) $$\{1, 2, \dots, n\} \;=\; J_1 \biguplus J_2 \biguplus \cdots \biguplus J_k$$

such that each J_i contains y_i of those n indices. It follows from Theorem 4.6 on p.62 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 (8.55). We conclude that $\vec{Y}$ has a multinomial distribution with parameters n and $p_{\vec{Y}}(y_1, y_2, \dots, y_k)$. ■

Theorem 8.19 (WMS Ch.05.9, Theorem 5.13).

Assume that $\vec{Y} = (Y_1, Y_2, \dots, Y_k)$ have a multinomial distribution with parameters n and $p_1, p_2, \dots, p_k$. Then, for $1 \le i, j \le k$ and $i \ne j$,

(a) $E[Y_i] = np_i$,

(b) $Var[Y_i] = np_iq_i$, *where* $q_i = 1 - p_i$,

(c) $Cov[Y_i, Y_j] = -np_ip_j$.

PROOF: See the proof of Theorem 5.13 in the WMS text. ■

8.8 The Bivariate Normal Distribution (Optional)

Definition 8.18 (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

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \qquad -\infty < y_1 < \infty,\ -\infty < y_2 < \infty, \tag{8.56}$$

where $Q = \dfrac{1}{1-\rho^2}\left[\dfrac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\,\dfrac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \dfrac{(y_2-\mu_2)^2}{\sigma_2^2}\right]$.

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 8.20.

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 (7.27) on p.96.
For the proof of **(b)**, see Casella, Berger [1]. ■

Theorem 8.21.

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 (8.56) is

$$f_{Y_1,Y_2}(y_1, y_2) \;=\; \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2}\,,$$

where $Q \;=\; \dfrac{(y_1-\mu_1)^2}{\sigma_1^2} \;-\; 0 \;+\; \dfrac{(y_2-\mu_2)^2}{\sigma_2^2}$. 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 8.20**(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 8.3 on p.113. ■

Remark 8.10. 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. □

9 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!

9.1 The Method of Distribution Functions

The Method of Distribution Functions is best explained by some examples.

Example 9.1. Find the CDF and PDF for $U := 2Y - 6$, where the density of the random variable Y is

$$f_Y(y) = \begin{cases} 4y\,, & \text{if } 0 \leq y \leq 1/2\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{9.1}$$

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

$$F_U(u) = P\{U \leq u\} = P\{2Y - 6 \leq u\} = P\left\{Y \leq \frac{u+6}{2}\right\} = F_Y\left(\frac{u+6}{2}\right).$$

Note that

$$0 \leq y \leq \frac{1}{2} \Leftrightarrow 0 \leq \frac{u+6}{2} \leq \frac{1}{2} \Leftrightarrow -6 \leq u \leq -5$$

Thus, $F_U(u) = 0$ for $u < -6$ and $F_U(u) = 1$ for $u > -5$.
For $-6 \leq u \leq -5$, i.e., $0 \leq y \leq \frac{1}{2}$, we must integrate:

$$P\left\{Y \leq \frac{u+6}{2}\right\} = \int_0^{(u+6)/2} f_Y(y)\,dy = \int_0^{(u+6)/2} 4y\,dy = \frac{4}{2}\left(\frac{u+6}{2}\right)^2.$$

We combine the cases $u < -6$; $-6 \leq u \leq -5$; $u > -5$ and obtain

$$F_U(u) = \begin{cases} 0\,, & \text{if } u < -6\,, \\ \dfrac{(u+6)^2}{2}\,, & \text{if } -6 \leq u \leq -5\,, \\ 1\,, & \text{if } u > -5\,. \end{cases}$$

We differentiate this CDF and otain the density function for U:

$$f_U(u) = \frac{dF_U(u)}{du} = \begin{cases} u+6\,, & \text{if } -6 \leq u \leq -5\,, \\ 0\,, & \text{else}\,. \end{cases} \quad \square$$

Example 9.2 (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, $F_U(u) \;=\; P\{Y_1 + Y_2 \le u\} \;=\; \displaystyle\iint_B f(y_1)f(y_2)\,dy_1\,dy_2$, 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 $\bullet\; u \le 0$ or $u \ge 2$ $\bullet\; 0 < u \le 1$ $\bullet\; 1 < u < 2$.

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, $\iint_B \cdots = \int\limits_0^1\int\limits_0^1 \cdots$ and thus, $F_U(u) = 1$.

Case 2: $\bullet\; 0 < u \le 1$.

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 $\iint_B \dots$ 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 $y_2 = u - y_1$

$$\begin{aligned} F_U(u) \;&=\; \iint_B 1\,dy_1\,dy_2 \;=\; \int_0^u \int_0^{u-y_1} 1\,dy_2\,dy_1 \\ &=\; \int_0^u (u - y_1)\,dy_1 \;=\; \left(uy_1 \;-\; \frac{u^2}{2}\right)\Big|_0^u \;=\; u^2 \;-\; \frac{u^2}{2} \;=\; \frac{u^2}{2}\,. \end{aligned}$$

Case 3: $\bullet\; 1 < u < 2$.

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

$$\textbf{(B)} \qquad \widetilde{B} \ = \big([0,1]\times[0,1]\big) \cap \{(y_1,y_2)\in\mathbb{R}^2 : y_1+y_2 \leq u\}\,,$$

$$\textbf{(C)} \qquad F_U(u) \ = 1 \ - \ P\{Y_1+Y_2 \geq u\} \ = \ 1 \ - \ \iint_{\widetilde{B}} 1\,dy_1\,dy_2$$

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,

$$\begin{aligned} 1 \ - \ F_U(u) \ &= \ \iint_{\widetilde{B}} 1\,dy_1\,dy_2 \ = \ \int_{u-1}^{1}\int_{u-y_1}^{1} 1\,dy_2\,dy_1 \\ &= \int_{u-y_1}^{1} (1-u+y_1)\,dy_1 \ = \ \left((1 \ - \ u) \ + \ \frac{y_1^2}{2}\right)\bigg|_{u-1}^{1} \\ &= (1-u)(2-u) \ + \ \frac{1}{2} \ - \ \frac{(u-1)^2}{2} \ = \ 2 \ - \ 2u \ + \ \frac{u^2}{2}\,. \end{aligned}$$

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 9.2: The Method of Transformations in One Dimension), because it is solved with the techniques of this section.

Example 9.3. Let Y_1 and Y_2 be jointly cntinuous 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 $u>0$,

$$\begin{aligned} P\{U\leq u\} \ &= \ \int_0^u \left[\int_0^{u-y_1} e^{-y_1-y_2}\,dy_2\right] dy_1 \ = \ \int_0^u e^{-y_1}\left[-e^{-y_2}\Big|_0^{u-y_1}\right] dy_1 \\ &= \ \int_0^u e^{-y_1}\left[1 \ - \ e^{-(u-y_1)}\right] dy_1 \ = \ \int_0^u e^{-y_1}\left[1 \ - \ e^{y_1}e^{-u}\right] dy_1 \\ &= \ \int_0^u e^{-y_1}\,dy_1 \ - \ \int_0^u e^{-u}\,dy_1 \ = \ -e^{-y_1}\Big|_0^u \ - \ u\,e^{-u} \\ &= \ -\left(e^{-u}-1\right) \ - \ u\,e^{-u} \ = \ 1 \ - \ (1+u)\,e^{-u}\,. \end{aligned}$$

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. $\square$

Remark 9.1. In the following we use the arrow notation $\vec{y} = \big(y_1, \dots, y_n\big)$, $\vec{Y} = \big(Y_1, \dots, Y_n\big)$, ...

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 = $\{g \leq u\} = g^{-1}\big(]-\infty, u]\big)$. (Thus, $R \subseteq \mathbb{R}^n$.)
- **(2)** Find the "boundary" $R^* = \{g = u\}$ of the region R.
- **(3)** Find the CDF $F_U(u) = P\{U \leq u\}$ 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

$$\begin{aligned} F_U(u) &= P\{U \leq u\} = P\{g\big(\vec{Y}\big) \leq u\} = P\big\{\omega : \vec{Y}(\omega) = \vec{y} \text{ such that } g(\vec{y}) \leq u\big\} \\ &= P\{Y \in R\} = \iint \cdots \int_R f_{\vec{Y}}(\vec{y})\,d\vec{y}. \ \square \end{aligned} \tag{9.2}$$

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 9.2. $\boxed{\star}$ 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 $y \geq 0$, i.e., $0 \leq y < y' \Rightarrow h(y) \neq h(y')$. (Thus, by symmetry, $h(y)$ also is injective for $y < 0$).

Then one can determine the CDF and PDF as follows:

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 $[-\infty, 0[$).
- **(2)** h is strictly decreasing on $[0, \infty[$ (and then, by symmetry, h is strictly increasing on $[-\infty, 0[$).

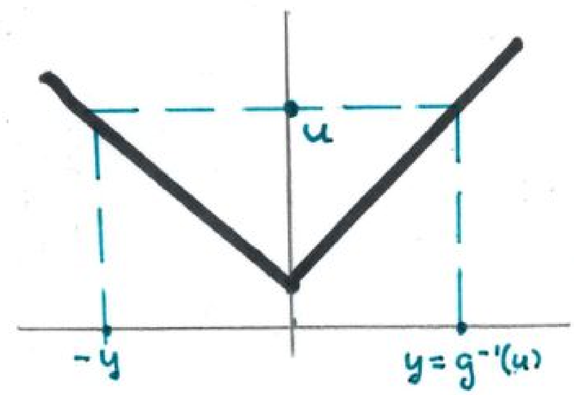
In either case, since there are no jumps for the continuous $h(\cdot)$, the following is true for any $u \in \mathbb{R}$:

(3) u is so small that $h(y) > u$ for all y. Note that then $P\{U \le u\} = 0$.

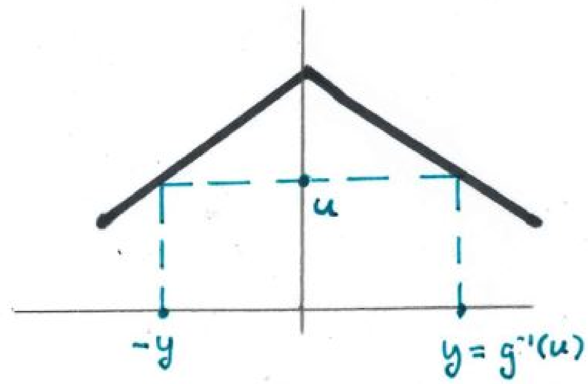
(4) u is so big that $h(y) < u$ for all y. Note that then $P\{U \le u\} = 1$.

(5) There is some y such that $h(y) = u$. Then, by symmetry, $h(-y) = u$. Moreover, the injectivity assumptions guarantee that $\left[\, \tilde{y} \neq y \text{ and } \tilde{y} \neq -y \,\right] \;\Rightarrow\; h(\tilde{y}) \neq u$.
Thus, for such u, there is a unique $y \ge 0$ such that $h(y) = u$. We write $y = h^{-1}(u)$. [31]

For such a u of **(5)** with corresponding $y = h^{-1}(u)$, we obtain (see Figures 9.1 and 9.2) [32]



9.1 (Figure). **symmetric $g(y)$, increasing for $y \ge 0$**



9.2 (Figure). **symmetric $g(y)$, decreasing for $y \ge 0$**

(6) If h is strictly increasing on $[0, \infty[$ (case **(1)**), then $U(\omega) \le u \;\Leftrightarrow\; |Y(\omega)| \le y = h^{-1}(u)$. Thus,

$$\begin{aligned} F_U(u) &= P\{U \le u\} \;=\; P\{|Y| \le h^{-1}(u)\} \;=\; P\{-h^{-1}(u) \le Y \le h^{-1}(u)\} \\ &= F_Y\big(h^{-1}(u)\big) \;-\; F_Y\big(-h^{-1}(u)\big)\,. \end{aligned}$$

Thus, if h is strictly increasing on $[0, \infty[$, we obtain for general $-\infty < u < \infty$ the following:

$$F_U(u) \;=\; \begin{cases} 1\,, & \text{if } h(y) < u \text{ for all } y\,, \\ F_Y\big(h^{-1}(u)\big) \;-\; F_Y\big(-h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{if } h(y) > u \text{ for all } y\,. \end{cases} \tag{9.3}$$

We differentiate $\dfrac{d}{du}$ to obtain the density. We write $h^{-1\prime}(u) = \dfrac{dh^{-1}(u)}{du}$:

- $f_U(u) \;=\; h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big) \;-\; (-1)h^{-1\prime}(u) f_Y\big(h^{-1}(u)\big) \;=\; 2\,h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big)\,.$
  Thus, $\quad f_U(u) \;=\; \begin{cases} 2\,h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{else}\,. \end{cases}$

(7) If h is strictly decreasing on $[0, \infty[$ (case **(2)**), then $U(\omega) \le u \;\Leftrightarrow\; |Y(\omega)| \ge y = h^{-1}(u)$. Thus,

$$\begin{aligned} F_U(u) &= P\{U \le u\} \;=\; P\{|Y| \ge h^{-1}(u)\} \;=\; P\{Y \le -h^{-1}(u)\} \;+\; P\{Y \ge h^{-1}(u)\} \\ &= F_Y\big(-h^{-1}(u)\big) \;+\; 1 \;-\; F_Y\big(h^{-1}(u)\big)\,. \end{aligned}$$

---

[31] $\boxed{\star}$ Matter of fact, $u \mapsto y = h^{-1}(u)$ is the inverse function of the function f, which becomes bijective, if we restrict its domain to $[0, \infty[$ and its codomain to $f([0, \infty[) = \{f(y) : 0 \le y < \infty\}$.

[32] In those figures, the name of the function is $g(y)$ rather than $h(y)$.

Thus, if h is strictly decreasing on $[0, \infty[$, we obtain for general $-\infty < u < \infty$ the following:

$$F_U(u) \;=\; \begin{cases} 1\,, & \text{if } h(y) < u \text{ for all } y\,, \\ F_Y\big(-h^{-1}(u)\big) \;+\; 1 \;-\; F_Y\big(h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{if } h(y) > u \text{ for all } y\,. \end{cases} \tag{9.4}$$

Again, we differentiate $\dfrac{d}{du}$ to obtain the density.

- $f_U(u) \;=\; (-1)h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big) \;-\; h^{-1\prime}(u) f_Y\big(h^{-1}(u)\big) \;=\; -2\, h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big)\,.$
  Thus, $f_U(u) \;=\; \begin{cases} (-2)\, h^{-1\prime}(u)\, f_Y\big(h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{else}\,. \end{cases}$

(8) Since h is strictly increasing on $[0, \infty[\;\Rightarrow\; h^{-1\prime}(u) > 0 \;\Rightarrow\; h^{-1\prime}(u) \;=\; |h^{-1\prime}(u)|$ and h is strictly decreasing on $[0, \infty[\;\Rightarrow\; h^{-1\prime}(u) < 0 \;\Rightarrow\; -h^{-1\prime}(u) \;=\; |h^{-1\prime}(u)|$, we can combine the results for the density $f_U(u)$ into a single formula:

$$f_U(u) \;=\; \begin{cases} 2\left|h^{-1\prime}(u)\right| f_Y\big(h^{-1}(u)\big)\,, & \text{if there is } y = h^{-1}(u)\,, \\ 0\,, & \text{else}\,. \quad \square \end{cases} \tag{9.5}$$

Remark 9.3. Examples of symmetric functions $u = h(y)$ are

$$u \;=\; y^2;\; u \;=\; cos(y);\; u \;=\; e^{cos(y)};\; u \;=\; \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2}\,.$$

The last two examples illustrate that ANY function $\psi\big(h(y)\big)$ of a symmetric function $h(\cdot)$ is symmetric. That is a triviality: Given any y,

$$g \text{ symmetric } \;\Rightarrow\; h(-y) \;=\; h(y) \;\Rightarrow\; \psi\big(h(-y)\big) \;=\; \psi\big(h(y)\big)\,.$$

Note that the last example (the standard normal PDF) is symmetric, because it is a function of y^2.

You are strongly encouraged to verify by direct computation that the results of Remark 9.2 are correct, if $h(y) = y^2$. You can do so without trying to understand the math that leads to the derivation of the formulas for $F_U(u)$ and $f_U(u)$. $\square$

Example 9.4. 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

$$\phi(y) \;:=\; f_Y(y) \;=\; \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,, \tag{9.6}$$

$$\Phi(y) \;:=\; \int_{-\infty}^{y} \phi(t)\, dt\,. \tag{9.7}$$

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 $u \geq 0$.

Then, $F_U(u) = P\{-\sqrt{u} \le Y \le \sqrt{u}\} = \Phi(\sqrt{u}) - \Phi(-\sqrt{u})$ and thus,

$$\begin{aligned} f_U(u) &= F_U'(u) = \frac{d}{du}\left[\Phi(\sqrt{u}) - \Phi(-\sqrt{u})\right] \\ &= \phi(\sqrt{u})\,\frac{1}{2\sqrt{u}} + \phi(-\sqrt{u})\,\frac{1}{2\sqrt{u}} = \phi(\sqrt{u})\,\frac{1}{\sqrt{u}} = \frac{1}{\sqrt{2\pi}}\,e^{-(\sqrt{u})^2/2}\,\frac{1}{\sqrt{u}} \end{aligned}$$

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 $\Gamma(1/2) = \sqrt{\pi}$. [33] We use that 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 9.5. 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, $F_U(u) = \begin{cases} 1 - e^{-\frac{u+4}{2}}\,, & \text{if } u \ge -2\,, \\ 0\,, & \text{else}\,. \end{cases}$

(6) We have obtained $F_U(u)$ without integrating a PDF.

(7) The density is $f_U(u) = F_U'(u) = \begin{cases} \frac{1}{2}\,e^{-\frac{u+4}{2}}\,, & \text{if } u \ge -2\,, \\ 0\,, & \text{else}\,. \end{cases}$ □

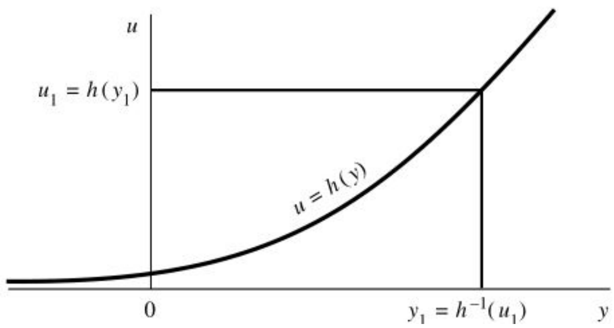
[33] See, e.g., https://en.wikipedia.org/wiki/Gamma_function or Shilov, G. [4].

9.2 The Method of Transformations in One Dimension

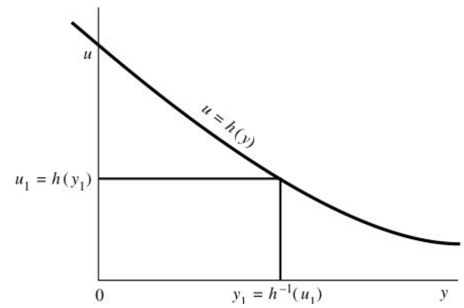
Introduction 9.1. We already encountered the method of transformations in Remark 9.2 on p.137. There we computed the CDF and PDF of the random variable $U = h(Y)$ for a continuous random variable Y and a symmetric and differentiable function $h(y)$ which was injective on the interval $B_1 = [0, \infty[$. (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 9.1 and 9.2 in that remark and also figures 9.3 and 9.4 below.



9.3 (Figure). **Strictly increasing function.**
Source: WMS Ch.6.4



9.4 (Figure). **Strictly decreasing function.**
Source: WMS Ch.6.4

Those figures illustrate the following.

(1) If h is strictly increasing, then $h(y) \leq u_1 \Leftrightarrow y \leq h^{-1}(u_1)$. Thus,

$$P\{U \leq u\} = P\{h(Y) \leq u\} = P\{h^{-1}[h(Y)] \leq h^{-1}(u)\} = = P\{Y \leq h^{-1}(u)\}\,, \quad \text{i.e.,} \quad F_U(u) = F_Y\big(h^{-1}(u)\big)\,. \tag{9.8}$$

We differentiate 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,

$$f_U(u) = f_Y\big(h^{-1}(u)\big)\cdot |h^{-1\prime}(u)|\,. \tag{9.9}$$

(2) If h is strictly decreasing, then $h(y) \leq u_1 \Leftrightarrow y \geq h^{-1}(u_1)$. Thus,

$$P\{U \leq u\} = P\{h(Y) \leq u\} = P\{Y \geq h^{-1}(u)\} = 1 - P\{Y \leq h^{-1}(u)\}\,, \quad \text{i.e.,} \quad F_U(u) = 1 - F_Y\big(h^{-1}(u)\big)\,. \tag{9.10}$$

We differentiate 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, ${h^{-1}}'(u) < 0$. Thus, $-{h^{-1}}'(u) = |{h^{-1}}'(u)|$. Thus,

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big) \cdot |{h^{-1}}'(u)|\,. \tag{9.11}$$

(3) We compare (**??**) and (9.11) 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,)

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big) \cdot |{h^{-1}}'(u)| \;=\; f_Y\big(h^{-1}(u)\big) \cdot \left| \frac{d[h^{-1}(u)]}{du} \right|\,. \tag{9.12}$$

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 the set $\mathrm{suppt}(f_Y) := \{\tilde y : f(\tilde y) \neq 0\}$. It is customary to call $\mathrm{suppt}(f_Y)$ the **support** of the density $f_Y(y)$. [34] □

The following theorem summarizes the observations of those introductory results:

Theorem 9.1.

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 $y \in$ suppt(f_Y), i.e., for all y that satisfyy $f_Y(y) > 0$. Then the PDF of $U := h(Y)$ is

$$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big) \cdot |{h^{-1}}'(u)| \;=\; f_Y\big(h^{-1}(u)\big) \cdot \left| \frac{d[h^{-1}(u)]}{du} \right|\,. \tag{9.13}$$

PROOF: See the introduction 9.1. ■

Example 9.6 (Increasing function)**.** Given is a random variable Y with the following PDF:

$$f_Y(y) \;=\; \begin{cases} 2y\,, & \text{if } 0 \le y \le 1\,, \\ 0\,, & \text{else}\,. \end{cases}$$

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, ${h^{-1}}'(u) = 1/4$. Since $0 \le (u+3)/4 \le 1 \;\Leftrightarrow\; -3 \le u \le 1$,

$$f_U(u) \;=\; \begin{cases} \dfrac{2(u+3)}{4} \cdot \dfrac{1}{4}\,, & \text{if } -3 \le u \le 1\,, \\ 0\,, & \text{else}\,. \end{cases} \;=\; \begin{cases} \dfrac{u+3}{8}\,, & \text{if } -3 \le u \le 1\,, \\ 0\,, & \text{else}\,. \;\square \end{cases}$$

[34] ⋆ In general, one defines the support $\mathrm{suppt}(g) := \{\tilde x : f(\tilde x) \neq 0\}$ for any real valued function $x \mapsto g(x)$.

Example 9.7 (Decreasing function)**.** Given is a random variable Y with the same PDF as in Example 9.6:

$$f_Y(y) \;=\; \begin{cases} 2y\,, & \text{if } 0 \le y \le 1\,,\\ 0\,, & \text{else}\,. \end{cases}$$

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 $0 \le (2-u)/3 \le 1 \;\Leftrightarrow\; 0 \ge (u-2) \ge -3 \;\Leftrightarrow\; -1 \le u \le 2$,

$$f_U(u) \;=\; \begin{cases} \dfrac{2(2-u)}{3}\cdot\left|\dfrac{-1}{3}\right|\,, & \text{if } -1 \le u \le 2\,,\\ 0\,, & \text{else}\,. \end{cases} \;=\; \begin{cases} \dfrac{4-2u}{9}\,, & \text{if } -3 \le u \le 1\,,\\ 0\,, & \text{else}\,. \;\square \end{cases}$$

Example 9.8 (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) $(y_1, y_2) \in B \;\Rightarrow\; 0 < 1 - y_1 < 1 \;\Rightarrow\; 0 > y_1 - 1 > -1 \;\Rightarrow\; 0 < y_2 < 1$.
$0 < y_2 < 1$ 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)**, $u \ge 1 \;\Rightarrow\; B \cap C = B \;=\; \text{suppt}(f_U) \;\Rightarrow\; F_U(u) = 1$.
(9) Thus, from **(3)** & **(7)** & **(8)**, $f_U(u) \;=\; \begin{cases} 0\,, & \text{if } u \le 0\,,\\ u^2\,, & \text{if } 0 < u < 1\,,\\ 1\,, & \text{if } u \ge 1\,. \;\square \end{cases}$

Remark 9.4. In the following we use the arrow notation $\vec{y} = (y_1, \dots, y_n)$, $\vec{Y} = (Y_1, \dots, Y_n)$, ...

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 $y \in D$ 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 $\quad \dfrac{dh^{-1}}{du} = \dfrac{dh^{-1}(u)}{du} = {h^{-1}}'(u)$.

(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 9.5. 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 $\quad u := h(y) := \begin{cases} y, & \text{if } y \leq 0\,, \\ e^{-y}, & \text{if } y > 0\,. \end{cases}$

- On $]-\infty, 0]$, h is strictly increasing with inverse $y = h^{-1}(u) = u$. This inverse has derivative ${h^{-1}}'(u) = 1 > 0$.
- On $]0, \infty[$, h is strictly decreasing with inverse $y = h^{-1}(u) = -\ln(u)$. This inverse has derivative ${h^{-1}}'(u) = -1/u < 0$.
- Obviously if $y \leq 0$, then $y \leq 0 \Leftrightarrow u \leq 0$. Moreover, $y > 0 \Leftrightarrow 0 < u = e^{-y} < 1$.
- Thus, $\quad f_U(u) = \begin{cases} f_Y(u) \cdot |1| = f_Y(u), & \text{if } u \leq 0\,, \\ f_Y\big(e^{-u}\big) \cdot |-1/u| = \dfrac{f_Y\big(-\ln(u)\big)}{u}, & \text{if } 0 < u < 1\,. \end{cases}$ ☐

9.3 The Method of Transformations in Multiple Dimension

Introduction 9.2. In Chapter 9.2 (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 9.1 on p.142 provided us with an explicit formula for the PDF $f_U(u)$ of the transformed random variable $U = h(Y)$:

(9.14) $$f_U(u) \;=\; f_Y\big(h^{-1}(u)\big)\cdot|{h^{-1}}'(u)| \;=\; f_Y\big(h^{-1}(u)\big)\cdot\left|\frac{d[h^{-1}(u)]}{du}\right|.$$

(1) Since $|{h^{-1}}'(u)|$ 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 ${h^{-1}}'(u)$ can be zero, since $h'(y)\cdot{h^{-1}}'(u) = 1$.

Existence of $h^{-1}(u)$ requires h to be injective on the support of the PDF f_Y:

(4) If u is the function value $u = h(y)$ of some argument y that satisfies $f_Y(y) > 0$,
• then there is no other argument $\tilde{y}$ that satisfies the conditions given in **(4)**.

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 lok 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 $\vec{Y} = \big(Y_1,\dots,Y_n\big)$, where each coordinate Y_j is a random variable.
- functions $\vec{u} = \vec{h}(\vec{y})$ 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 $\vec{u}$ is a function value $\vec{u} = \vec{h}(\vec{y})$ of some argument $\vec{y}$ that satisfies $f_{\vec{Y}}(\vec{y}) > 0$, (here, $f_{\vec{Y}}(\vec{y})$) is the PDF of the jointly continuous random variables $Y_1,\dots,Y_n$),
• then there is no other argument $\vec{\tilde{y}}$ that satisfies all those conditions.

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 $h_1^{-1}(\cdot),\ h_2^{-1}(\cdot),\dots,h_n^{-1}(\cdot)$. They are defined by the equations

(9.15) $$h_1^{-1}(\vec{u}) \;=\; y_1,\quad h_2^{-1}(\vec{u}) \;=\; y_2,\ \dots,\ h_n^{-1}(\vec{u}) \;=\; y_n.$$

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,

(9.16) $$\frac{dh^{-1}}{dy} \;=\; 1\Big/\frac{dh}{du}\,.$$

In the n–dimensional case, we must replace the condition $\left|\dfrac{dh}{du}\right| \neq 0$ with the condition

(5) $$J^{-1} := \det \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \dots & \frac{\partial h_1}{\partial y_n} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \dots & \frac{\partial h_2}{\partial y_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n}{\partial y_1} & \frac{\partial h_n}{\partial y_2} & \dots & \frac{\partial h_n}{\partial y_n} \end{bmatrix} \neq 0\,.$$

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,

$$J = \det \begin{bmatrix} \frac{\partial h_1^{-1}}{\partial u_1} & \frac{\partial h_1^{-1}}{\partial u_2} & \dots & \frac{\partial h_1^{-1}}{\partial u_n} \\ \frac{\partial h_2^{-1}}{\partial u_1} & \frac{\partial h_2^{-1}}{\partial u_2} & \dots & \frac{\partial h_2^{-1}}{\partial u_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n^{-1}}{\partial u_1} & \frac{\partial h_n^{-1}}{\partial u_2} & \dots & \frac{\partial h_n^{-1}}{\partial u_n} \end{bmatrix} \neq 0\,. \tag{9.17}$$

Moreover, the determinants J^{-1} and J satisfy the analogue of (9.16):

$$J^{-1} = \frac{1}{J}\,. \ \square \tag{9.18}$$

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 9.9 (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:

$$u_1 := h_1(y_1, y_2) := 2y_1 + y_2\,, \tag{9.19}$$
$$u_2 := h_2(y_1, y_2) := y_1 - 2y_2\,. \tag{9.20}$$

(1) We show that this function can be inverted by solving these equations for $\vec{y} = (y_1, y_2)$.

- $u_1 - 2u_2 - 2u_2 \overset{(9.19)}{=} y_2 + 4y_2 = 5y_2 \Rightarrow y_2 = u_1/5 - 2u_2/5\,.$
- Thus, $y_1 \overset{(9.20)}{=} u_2 + 2y_2 = u_2 + (1/5)[2u_1 - 4u_2] = (2u_1)/5 + u_2/5.$

We have found the inverse function $\vec{h}^{-1} = \big(h_1^{-1}, h_2^{-1}\big)$ to be

$$h_1^{-1}(u_1, u_2) = y_1 = \frac{1}{5}(2u_1 + u_2)\,, \tag{9.21}$$
$$h_1^{-1}(u_1, u_2) = y_2 = \frac{1}{5}(u_1 - 2u_2)\,. \tag{9.22}$$

We will continue in Example9.10 on p.148. $\square$

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 9.2.

- *Let $\vec{Y} = (Y_1, \dots, Y_n)$ 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 $1 \le i, j \le n$.*

⊡ *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 $\vec{y}$ that satisfies all those conditions.*

Then $\vec{h}$ has an inverse $\vec{h}^{-1} = h_1^{-1}, h_2^{-1}, \dots, h_n^{-1}$ 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:

$$u_1 \ = \ h_1(\vec{y}),\ \dots,\ u_n \ = \ h_n(\vec{y}) \quad \text{and} \quad y_1 \ = \ h_1^{-1}(\vec{u}),\ \dots,\ y_n \ = \ h_n^{-1}(\vec{u}). \tag{9.23}$$

Also, all partial derivatives $\dfrac{\partial h_i^{-1}}{u_j}$ exist and are continuous for $1 \le i, j \le n$.

$$\text{Let} \quad \frac{d\vec{h}}{d\vec{y}} := \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \cdots & \frac{\partial h_2}{\partial y_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n}{\partial y_1} & \frac{\partial h_n}{\partial y_2} & \cdots & \frac{\partial h_n}{\partial y_n} \end{bmatrix}, \qquad \frac{d\vec{h}^{-1}}{d\vec{u}} := \begin{bmatrix} \frac{\partial h_1^{-1}}{\partial u_1} & \frac{\partial h_1^{-1}}{\partial u_2} & \cdots & \frac{\partial h_1^{-1}}{\partial u_n} \\ \frac{\partial h_2^{-1}}{\partial u_1} & \frac{\partial h_2^{-1}}{\partial u_2} & \cdots & \frac{\partial h_2^{-1}}{\partial u_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h_n^{-1}}{\partial u_1} & \frac{\partial h_n^{-1}}{\partial u_2} & \cdots & \frac{\partial h_n^{-1}}{\partial u_n} \end{bmatrix}. \tag{9.24}$$

$$\text{Let} \quad J^{-1} \ := \ J^{-1}(\vec{y}) \ := \ det\left(\frac{d\vec{h}}{d\vec{y}}\right), \qquad J \ := \ J(\vec{u}) \ := \ det\left(\frac{d\vec{h}^{-1}}{d\vec{u}}\right). \tag{9.25}$$

- *We add another assumption: $J^{-1}(\vec{y}) \ \neq \ 0$ for all y that satisfy $f_{\vec{Y}}(\vec{y}) > 0$.*

$$\text{Then} \quad J\big(h(\vec{y})\big) \neq 0 \quad \text{and} \quad J\big(h(\vec{y})\big) \ = \ 1 \Big/ J^{-1}(\vec{y})\,. \tag{9.26}$$

Further, the density of the transform $\vec{U} = h(\vec{Y})$ is computed as

$$f_{\vec{U}}(\vec{u})) \ = \ f_{\vec{Y}}\big(h^{-1}(\vec{u})\big) \cdot J(\vec{u})\,. \tag{9.27}$$

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 $\dfrac{d\vec{h}}{d\vec{y}}$ of the partial derivatives of h.

Remark 9.6. In the onedimensional case ($n = 1$), the situation is as follows.

- $\mathbb{R}^n$ is the set $\mathbb{R}$ of real numbers, • $\vec{u} = \vec{h}(\vec{y})$ becomes $u = h(y)$ for real numbers y and u,
- the 1×1 "matrix" of "partial" derivatives is $h'(y) = \dfrac{dh}{dy}$.

Considering that last point, it seems natural to write $\dfrac{d\vec{h}}{d\vec{y}}$ for the $n \times n$ matrix of partial derivatives $\dfrac{\partial h_i}{\partial y_j}$ and this author chose to do so. However, you will find either different notation [35] or, like in the WMS text, no dedicated symbols at all. That works well enough with 2×2 matrices. □

Definition 9.1 (Jacobian and Jacobian matrix)**.**

The matrix $\dfrac{d\vec{h}}{d\vec{y}}$ 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, $J^{-1}(\vec{y}) = \det\left(\dfrac{d\vec{h}}{d\vec{y}}\right)$, as the **Jacobian**, sometimes also the **Jacobian determinant**, of $\vec{h}(\cdot)$. □

Notation 9.1 (Jacobian: WMS definition)**.**

- Stewart writes $\dfrac{\partial(u_1,\dots,u_n)}{\partial(y_1,\dots,y_n)} := \det\left(\dfrac{d\vec{h}^{-1}}{d\vec{u}}\right)$ and $\dfrac{\partial(y_1,\dots,y_n)}{\partial(u_1,\dots,u_n)} := \det\left(\dfrac{d\vec{h}^{-1}}{d\vec{u}}\right)$
- This author follows the great majority of books on multivariable calculus in defining the the Jacobian as the determinant of $\dfrac{d\vec{h}}{d\vec{y}}$.
- Be aware that WMS chooses instead to call $J = \det\dfrac{d\vec{h}^{-1}}{d\vec{u}}$ the Jacobian.
- The reason seems to be that most books on probability and statistics agree on using the letter J for $\det\dfrac{d\vec{h}^{-1}}{d\vec{u}}$ (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 9.10 (The joint PDF of two independent, exponential random variables – Part 2)**.** In Example 9.9 on p.146, we defined $\vec{u} = \vec{h}(\vec{y})$ as follows:

$$u_1 = h_1(y_1, y_2) = 2y_1 + y_2\,, \qquad u_2 = h_2(y_1, y_2) = y_1 - 2y_2\,.$$

[35] For example, Williamson, Richard E. and Trotter, Hale [6] uses the notation $\vec{h}'(\vec{y})$, the multidimensional analogue of $h'(y)$.

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 (9.26) on p.147.

We will continue in Example9.11. □

Example 9.11 (The joint PDF of two independent, exponential random variables – Part 3)**.** In Example 9.9 on p.146, we defined $\vec{u} = \vec{h}(\vec{y})$ as follows:

(9.28) $$u_1 = h_1(y_1, y_2) = 2y_1 + y_2, \qquad u_2 = h_2(y_1, y_2) = y_1 - 2y_2.$$

In its continuation, Example 9.10 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 9.2 on p.147 to compute the joint density $f_{\vec{U}}(u_1, u_2)$ of $\vec{U} = \vec{h}(\vec{Y})$.

Solution:

(a) $$f_{\vec{Y}}(\vec{y}) = f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{2}\, e^{-(y_1+y_2)/2}, & \text{if } y_1, y_2 > 0, \\ 0, & \text{else}. \end{cases}$$

(b) We recall that $y_1 = \frac{1}{5}(2u_1 + u_2)$ and $y_2 = \frac{1}{5}(u_1 - 2u_2)$. Thus,

$$f_{\vec{U}}(\vec{u}) = f_{U_1,U_2}(u_1, u_2) = \frac{1}{2} \exp\left\{ - \left(\frac{1}{5}(2u_1 + u_2) + \frac{1}{5}(u_1 - 2u_2) \right) \Big/ 2 \right\} \cdot \left| -\frac{1}{5} \right|$$

$$= \frac{1}{10} \exp\left\{ \frac{-1}{10}\,(2u_1 + u_2 + u_1 - 2u_2) \right\} = \frac{1}{10} \exp\left\{ \frac{3u_1 - u_2}{-10} \right\} = \frac{1}{10} \exp\left\{ \frac{u_2 - 3u_1}{10} \right\}.$$

- **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 $y_2 \geq 0 \overset{(9.28)}{\Rightarrow} u_1 = 2y_1 + y_2 \geq 0$.

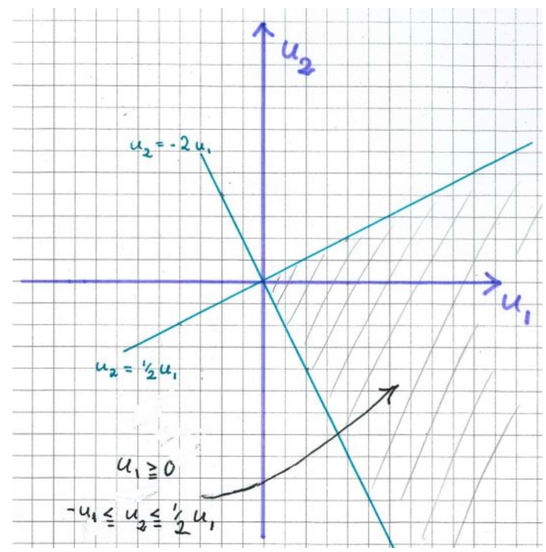
(e) From **(c)**: $2u_1 + u_2 \geq 0 \;\Rightarrow\; u_2 \geq -2u_1$

(f) From **(c)**: $u_1 - 2u_2 \geq 0 \;\Rightarrow\; u_1 \geq 2u_2$ $\Rightarrow\; u_2 \leq \dfrac{u_2}{2}$

(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 $-2u_1 \leq u_2 \leq \dfrac{u_2}{2}$.

- 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,
$$f_{\vec{U}}(\vec{u}) \;=\; \begin{cases} \dfrac{1}{10}\, e^{(u_2 - 3u_1)/10}, & \text{if } \vec{u} \in R\,, \\ 0, & \text{else}\,. \end{cases}$$



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,

$$\iint_{\mathbb{R}^2} \cdots f_{\vec{U}}(\vec{U})\, d\vec{u} \;=\; \iint_R \cdots \tfrac{-1}{10}\, e^{(u_2 - 3u_1)/10}\, d\vec{u} \;=\; \int_0^\infty \int_{-2u_1}^{u_2/2} \cdots \tfrac{-1}{10}\, e^{(u_2 - 3u_1)/10}\, du_2\, du_1$$

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

$$E[g(\vec{U})] \;=\; = \int_0^\infty \int_{-2u_1}^{u_2/2} g(\vec{u})\tfrac{-1}{10}\, e^{(u_2 - 3u_1)/10}\, du_2\, du_1 \quad \square$$

9.4 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 n random variables $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$, 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 9.12. Here are some examples.

(a) 70 students are randomly selected when exiting lecture hall and their age is measured in years. Those 70 ages, $A_1(\omega), \dots, A_{70}(\omega)$, 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, $A_{(j)} = A_{(j+1)}$, 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, $H_1(\omega), \dots, H_{70}(\omega)$, 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 9.2 (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

$$Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \cdots \leq Y_{(n)}. \tag{9.29}$$

We call $Y_{(j)}$ the **jth order statistic** of $\vec{Y}$.

See Example 9.12**(b)** why we may consider strictly increasing rather than nondecreasing. $\square$

Assumption 9.1.

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. $\square$

Remark 9.7. 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:

$$Y_{(1)}(\omega) > y \Leftrightarrow \min\big(Y_1(\omega), \dots, Y_n(\omega)\big) > y \Leftrightarrow Y_j(\omega) > y \text{ for all } j \in [1,n]_{\mathbb{Z}}\,, \tag{9.30}$$

$$Y_{(n)}(\omega) \le y \Leftrightarrow \max\big(Y_1(\omega), \dots, Y_n(\omega)\big) \le y \Leftrightarrow Y_j(\omega) \le y \text{ for all } j \in [1,n]_{\mathbb{Z}}\,. \quad \square \tag{9.31}$$

Theorem 9.3 (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:

$$F_{Y_{(1)}(y)} = 1 - [1 - F(y)]^n\,, \tag{9.32}$$

$$F_{Y_{(n)}(y)} = [F(y)]^n\,, \tag{9.33}$$

$$F_{Y_{(k)}(y)} = 1 - \sum_{j=0}^{k-1} \binom{n}{j} [F(y)]^j\, [1 - F(y)]^{n-j}\,. \tag{9.34}$$

For $y \in \mathbb{R}$, the PDF of the kth order statistic ($k = 1, \dots, n$) satisfies the following:

$$f_{Y_{(1)}(y)} = n\,[1 - F(y)]^{n-1}\, f(y)\,, \tag{9.35}$$

$$f_{Y_{(n)}(y)} = n\,[F(y)]^{n-1}\, f(y)\,, \tag{9.36}$$

$$f_{Y_{(k)}(y)} = \sum_{j=0}^{k-1} \binom{n}{j} f(y)\,\big(j\,[F(y)]^{j-1} - n[F(y)]^{n-1}\big)\,. \tag{9.37}$$

$$f_{Y_{(k)}(y)} = \binom{n-1}{k-1} f(y)[F(y)]^{k-1}\,[1 - F(y)]^{n-k}\,. \tag{9.38}$$

Note that the proofs are not given in the order of the seven formulas of the theorem.
PROOF of (9.33):

$$\begin{aligned} F_{Y_{(n)}(y)} &\overset{(9.31)}{=} P\big(\{Y_1 \le y\} \cap \{Y_2 \le y\} \cap \dots \cap \{Y_n \le y\}\big) \\ &\overset{\textbf{indep}}{=} P\{Y_1 \le y\} \cdot P\{Y_2 \le y\} \cdots P\{Y_n \le y\} = [F(y)]^n\,. \end{aligned}$$

PROOF of (9.32):

$$\begin{aligned} P\{Y_{(1)} > y\} &\overset{(9.30)}{=} P\big(\{Y_1 > y\} \cap \{Y_2 > y\} \cap \dots \cap \{Y_n > y\}\big) \\ &\overset{\textbf{indep}}{=} P\{Y_1 > y\} \cdot P\{Y_2 > y\} \cdots P\{Y_n >\} = [1 - F(y)]^n\,. \end{aligned}$$

Thus, $F_{Y_{(1)}(y)} = 1 - P\{Y_{(1)} > y\} = 1 - [1 - F(y)]^n$.
PROOF of (9.35) and (9.36):
This follows from $\dfrac{d}{dy}\big(1 - [1 - F(y)]^n\big) = -n[1 - F(y)]^{n-1}\big(-f(y)\big)$

and $\dfrac{d}{dy}\big([F(y)]^n\big) = n[F(y)]^{n-1}\, f(y)$.

PROOF of (9.34):

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 $j = 1, \dots, n$, let $X_j(\omega) := \begin{cases} 1 & \text{if } Y_j(\omega) \le y\,, \\ 0 & \text{else}\,. \end{cases}$ Let $U(\omega) := \sum_{j=1}^{n} X_j(\omega)$.
- Thus, if we interpret $Y_j(\omega) \le y$ as a success and $Y_j(\omega) > y$ as a failure, $X_1, \dots, X_n$ are a 0–1 encoded Bernoulli sequence [36] and $U \sim$ 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 $U \sim$ binom(n, p) and $p = F(y)$, $F_{Y_{(k)}}(y) = 1 - \sum_{j=0}^{k-1} \binom{n}{j} [F(y)]^j \, [1 - F(y)]^{n-j}$.

PROOF of (9.37):

This is done by differentiation. For each $j = 0, \dots, k-1$,

$$\textbf{(A)} \qquad \begin{aligned} \frac{d}{dy}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j} &= \binom{n}{j}\frac{d}{dy}\left([F(y)]^j \;-\; F(y)]^n\right) \\ &= \binom{n}{j}\;\left(j\,[F(y)]^{j-1}\,f(y) \;-\; n\,F(y)]^{n-1}\,f(y)\,\right) \end{aligned}$$

$$\text{Thus,} \qquad \begin{aligned} f_{Y_{(k)}} &= \frac{d}{dy}\left[1 - \sum_{j=0}^{k-1}\binom{n}{j}[F(y)]^j\,[1-F(y)]^{n-j}\right] \\ &= -\sum_{j=0}^{k-1}\frac{d}{dy}\binom{n}{j}\left(\,[F(y)]^j\,[1-F(y)]^{n-j}\,\right) \\ &\overset{\text{(A)}}{=} \sum_{j=0}^{k-1}\binom{n}{j}\,f(y)\,\left(\,j\,[F(y)]^{j-1} \;-\; n\,F(y)]^{n-1}\,\right)\,. \end{aligned}$$

This finishes the proof of (9.37).

The proof of (9.38) is based on an entirely different approach. Before we do that proof, we first illustrate that approach by redoing those of (9.35) and (9.36). Those proofs are much simpler and are a good preparation for that of (9.38).

ALTERNATE PROOF of (9.36):

First we note the following for a continuous random variable U with density $f_U(u)$ Assume that $\Delta u > 0$ 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 $]u, u + \Delta u$.

(a) Thus, $P\{u < U \le u + \Delta u\} = \int_u^{u+\Delta u} f_U(t)\,dt \approx f_U(u) \cdot \Delta$.

[36]See Definition 6.6 (Bernoulli items and variables) on p.76.

(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 items $X_1, \dots, X_n$ as follows.
$\boxdot$ $X_j(\omega) = L \Leftrightarrow Y_j(\omega) \le y$. Then $P\{X_j = L\} = P\{Y_j \le y\} = F(y)$.
$\boxdot$ $X_j(\omega) = I \Leftrightarrow y < Y_j(\omega) \le y + \Delta y$, Then $P\{X_j = I\} = P\{y < Y_j \le y + \Delta y\} \overset{\textbf{(a)}}{\approx} f_U(u) \cdot \Delta$.
$\boxdot$ $X_j(\omega) = R \Leftrightarrow Y_j(\omega) > y + \Delta y$. Then $P\{X_j = R\} = P\{Y_j > y\} = 1 - F(y)$.

(c) By construction, the X_j form a multinomial sequence. Let $\vec{U} := (U_1, U_2, U_3)$, 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, $p_1 = F(y)$, $p_2 = f(y)\Delta y$, $p_3 = 1 - F(y)$.

(e) Since we assume that $Y_{(j)}(\omega)$ is strictly increasing with j for all ω, it seems reasonable that, for "really small" Δy, the following is true:

- If $Y_{(1)}(\omega) > y$, then $Y_{(j)}(\omega) > y + \Delta y$ for all $j > 1$.

(f) Thus, $f_{Y_{(1)}}(y) \cdot \Delta y \overset{\textbf{(a)}}{\approx} P\{\, y < Y_{(1)} \le y + \Delta y \,\}$
$= P\{$ exactly one of $Y_1, \dots, Y_n \in]y, y + \Delta y]$ and $Y_j > y + \Delta y$ for all other $j\,\}$.
$= P\{$ none of the X_j are L and exactly one is I and $n-1$ are $R\,\}$.
$= P\{U_1 = 0,\, U_2 = 1,\, U_3 = n-1,\,\} \overset{\textbf{(d)}}{=} \binom{n}{0, 1, n-1} [F(y)]^0 \left[f(y)\Delta y\right]^1 [1 - F(y)]^{n-1}$.

(g) Since $\binom{n}{0, 1, n-1} = \dfrac{n!}{0! \cdot 1! \cdot (n-1)!} = n$,
we obtain $f_{Y_{(1)}}(y) \cdot \Delta y \approx n\, [1 - F(y)]^{n-1} f(y)\Delta y$.

We divide both expressions by Δy and obtain the density of $Y_{(1)}$ as

$$f_{Y_{(1)}}(y) \approx n\,[1 - F(y)]^{n-1}\, f(y)\,.$$

ALTERNATE PROOF of (9.36):

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') $f_{Y_{(n)}}(y) \cdot \Delta y \overset{\textbf{(a)}}{\approx} P\{\, y < Y_{(n)} \le y + \Delta y \,\}$
$= P\{$ exactly one of $Y_1, \dots, Y_n \in]y, y + \Delta y]$ 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 $L\,\}$.
$= P\{U_1 = n-1,\, U_2 = 1,\, U_3 = 0,\,\} \overset{\textbf{(d)}}{=} \binom{n}{n-1, 1, 0} [F(y)]^{n-1} \left[f(y)\Delta y\right]^1 [1 - F(y)]^0$.

(g') Since $\binom{n}{n-1, 1, 0} = \dfrac{n!}{(n-1)! \cdot 1! \cdot 0!} = n$,
we obtain $f_{Y_{(n)}}(y) \cdot \Delta y \approx n\, [F(y)]^{n-1} f(y)\Delta y$.

We divide both expressions by Δy and obtain the density of $Y_{(n)}$ as

$$f_{Y_{(n)}}(y) \approx n\,[F(y)]^{n-1}\, f(y)\,.$$

PROOF of (9.37):

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′) $f_{Y_{(n)}}(y) \cdot \Delta y \overset{\textbf{(a)}}{\approx} P\{\, y < Y_{(n)} \leq y + \Delta y \,\}$
$= P\{$ exactly one of $Y_1, \dots, Y_n \in]y, y+\Delta y]$ and $Y_j \leq y$ for all other $j\,\}$.
$= P\{$ none of the X_j are R and exactly one is I and $n-1$ are $L\,\}$.
$= P\{U_1 = n-1,\, U_2 = 1,\, U_3 = 0,\,\} \overset{\textbf{(d)}}{=} \binom{n}{n-1, 1, 0} [F(y)]^{n-1} \left[f(y)\Delta y\right]^1 [1 - F(y)]^0.$

(g′) Since $\binom{n}{n-1,1,0} = \dfrac{n!}{(n-1)! \cdot 1! \cdot 0!} = n,$
we obtain $f_{Y_{(n)}}(y) \cdot \Delta y \approx n\,[F(y)]^{n-1} f(y)\Delta y.$

We divide both expressions by Δy and obtain the density of $Y_{(n)}$ as

$$f_{Y_{(n)}}(y) \approx n\,[F(y)]^{n-1} f(y)\,.$$

PROOF of (9.38):

This time 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″) $f_{Y_{(k)}}(y) \cdot \Delta y \overset{\textbf{(a)}}{\approx} P\{\, y < Y_{(k)} \leq y + \Delta y \,\}$
$= P\{$ exactly one of $Y_1, \dots, Y_n \in]y, y+\Delta y]$ and $Y_j \leq y$ 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
$= P\{U_1 = k-1,\, U_2 = 1,\, U_3 = n-k,\,\} \overset{\textbf{(d)}}{=} \binom{n}{k-1,1,n-k} [F(y)]^{k-1} \left[f(y)\Delta y\right]^1 [1 - F(y)]^{n-k}.$

(g″) Since $\binom{n}{k-1,1,n-k} = \dfrac{n \cdot (n-1)!}{(k-1)! \cdot 1! \cdot (n-k)!} = n \cdot \binom{n-1}{k-1},$
we obtain $f_{Y_{(n)}}(y) \cdot \Delta y \approx n \cdot \binom{n-1}{k-1} [F(y)]^{k-1} f(y)\Delta y\, [1 - F(y)]^{n-k}.$

We divide both expressions by Δy and obtain the density of $Y_{(k)}$ as

$$f_{Y_{(n)}}(y) \approx n\,[F(y)]^{n-1} f(y)\,.$$

■

Remark 9.8. (9.34) yields (9.32) for $k = 1$ and (9.33) 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 (9.34) for $k = 1$ and $k = n$, we obtain

$$\begin{aligned} F_{Y_{(1)}(y)} &= 1 - \binom{n}{0} [F(y)]^0\, [1 - F(y)]^n = 1 - 1 \cdot 1 \cdot [1 - F(y)]^n = [1 - F(y)]^n\,, \\ F_{Y_{(n)}(y)} &= 1 - \sum_{j=0}^{n-1} \binom{n}{j} [F(y)]^j\, [1 - F(y)]^{n-j} \overset{\textbf{(A)}}{=} \binom{n}{0} [F(y)]^0\, [1 - F(y)]^n = [1 - F(y)]^n\,. \quad \square \end{aligned}$$

Remark 9.9. You may have noticed that there are two formulas for $f_{Y_{(k)}(y)}$.
(9.38) was shown by means of the "density approach" that utilized a limiting process $\triangle y \to 0$ in conjunction with the multinomial distribution. The proof was harder than that of (9.37). In return, (9.38) has computational advantages, since no more summation $\sum_{j=0}^{k-1}$ is required. $\square$

@@Author

The next remark belongs into Section 4.2 (Permutations) of Chapter 4. It has been placed here, since every order statistic

$$\vec{Y}_{(\bullet)} \;=\; \big(Y_{(1)}, \dots, Y_{(n)}\big).$$

is a (specific) permutation of $\vec{Y} = \big(Y_1, \dots, Y_n\big)$, and every other permutation

$$\big(Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}\big)$$

of $\vec{Y} = \big(Y_1, \dots, Y_n\big)$, possesses the same order statistic.

Remark 9.10. If we deal with a list $\vec{a} = (a_1, a_2, \dots, a_n)$ of distinct numbers, e.g.,

$$\textbf{(A)} \qquad \vec{a} \;=\; (13.2,\ -3,\ 6.6,\ 2,\ -1.5)\,,$$

then there is a uniquely determined permutation, $\vec{a}_{(\bullet)} = (a_{(1)}, a_{(2)}, \dots, a_{(n)})$ of $\vec{a}$, which has those a_j in increasing order. In other words,

$$a_{(1)} \;<\; a_{(2)} \;< \dots <\; a_{(n)}\,.$$

In the specific example **(A)**, we obtain

$$\vec{a}_{(\bullet)} \;=\; (-3,\ -1.5,\ 2,\ 6.6,\ 13.2)\,.$$

If $\vec{b} = (b_1, b_2, \dots, b_n)$ 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. $\square$

Theorem 9.4 (Joint PDF of the order statistic)**.**

Let $\vec{y} \in \mathbb{R}^n$ satisfy

$$y_1 < y_2 < \cdots < y_n . \tag{9.39}$$

For the vector $\vec{Y} = \big(Y_1, \dots, Y_n\big)$, let $\vec{Y}_{(\bullet)}$ be the vector of its associated order statistics, i.e.,

$$\vec{Y}_{(\bullet)} = \big(Y_{(1)}, \dots, Y_{(n)}\big). \tag{9.40}$$

Then its density function at $\vec{y}$ is given by

$$f_{\vec{Y}_{(\bullet)}}(\vec{y}) = n! \cdot \prod_{j=1}^{n} f(y_j) = n!\, f(y_1) \cdots f(y_n) . \tag{9.41}$$

If $\vec{y}$ does not satisfy (9.39), then $f_{\vec{Y}_{(\bullet)}}(\vec{y}) = 0$.

FIRST PROOF:
Let Δ be a "small" cube that is centered at $\vec{y}$. Study the proof of (9.34) of Theorem 9.3 on p.152. It explains (in the onedimensional case), why one can approximate

$$P\{\vec{Y} \in \Delta\} \approx f_{\vec{Y}}(\vec{y}) \cdot \Delta ,$$
$$P\{\vec{Y}_{(\bullet)} \in \Delta\} \approx f_{\vec{Y}_{(\bullet)}}(\vec{y}) \cdot \Delta .$$

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 all intervals $[\,y_j - \varepsilon, y_j + \varepsilon\,]$ have empty intersection. For the following, see Remark 9.10 on p.156.
Note that

$$\textbf{(A)} \qquad \begin{aligned} \vec{Y}_{(\bullet)}(\omega) \in \Delta &\Leftrightarrow y_k - \varepsilon \le Y_{(k)}(\omega) \le y_k + \varepsilon \text{ for all } k, \\ &\Leftrightarrow y_k - \varepsilon \le Y_j(\omega) \le y_k + \varepsilon \text{ for all } k, \\ &\qquad \text{where } j \text{ can be chosen depending on } k. \end{aligned}$$

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) $2.35 \le Y_3(\omega) \le 2.85,\ 3.95 \le Y_2(\omega) \le 4.45,\ 7.55 \le Y_1(\omega) \le 8.05,$

We use this illustrates the phrase "for all k, where j can be chosen depending on k".

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 9.10 on p.156 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.
- Since the Y_j are iid, each one has the same approximate probability $\prod_{j=1}^n f(y_j)$.
- Thus, $f_{\vec{Y}_{(\bullet)}}(\vec{y}) \cdot \Delta \ \approx \ n! \cdot \prod_{j=1}^n f(y_j) \cdot \Delta$
- 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, $\big[y_j \leq Y_{(j)}(\omega) \leq y_j + d\,t_j$ for all $j \,\big] \ \Leftrightarrow \ \big[$ there is a permutation $i_1, i_2, \dots, i_n$ of the indices $1, 2, \dots, n$ such that $y_j \leq Y_{i_j}(\omega) \leq y_j + d\,t_j$ for all $j \,\big]$

(c) Thus, $\big[y_j \leq Y_{(j)}(\omega) \leq y_j + d\,t_j$ for all $j \,\big] \ \Leftrightarrow \ \big[$ among the $X_i(\omega)$, 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 $X_1, \dots, X_n$, where $X_k(\omega)$ results in outcome $\#j$, if $y_j \leq Y_k \leq y_j + d\,t_j$.

- The probabilities $p_j = P\{X_k$ results in $\#j\}$ are, for small enough dt_j, equal to

$$p_j \, = \, P\{Y_i \in [y_j, y + dt_j]\,\} \ = \ \int\limits_{y_j}^{y+dt_j} f(t)\,dt \ \approx \ f(t_j)\,dt_j.$$

(e) From **(b)**, **(c)**, **(d)**:

$$\begin{aligned} f_{\vec{Y}_{(\bullet)}}(\vec{y})\,dt_1 \cdots\, dt_n \ &= P\{y_j \leq Y_{(j)}(\omega) \leq y_j + d\,t_j \text{ for all } j\}\\ &= P\{ \text{ there is a permutation } i_1, i_2, \dots, i_n \text{ of the indices } 1, 2, \dots, n\\ &\qquad\quad \text{such that } y_j \leq Y_{i_j} \leq y_j + d\,t_j \text{ for all } j\}\\ &= P\{ \text{ each } X_k \text{ has exactly one outcome } \#j \text{ for each } j = 1, \dots, n\}\\ &= \binom{n}{1, 1, \dots, 1} p_1^1 p_2^1 \cdots p_n^1 \ = \ \frac{n!}{1! \cdots 1!} \prod_j \big(f(t_j)\,dt_j\big)\,. \end{aligned}$$

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 9.13. Find the formula for the joint density of $Y_{(1)}$ and $Y_{(n)}$.

Solution:

(a) Note that, since the Y_j are continuous, "$<$" and "$\leq$" can be interchanged and the same is true for "$>$" and "$\geq$" when computing probabilities.

(b) Also, applying $A = (A \cap B) \uplus A \cap B^{\complement}$ with $A = \{Y_{(n)} \leq y_n\}$ and $B = \{Y_{(1)} \leq y_1\}$ yields $P\{Y_{(n)} \leq y_n\} \ = \ P\{Y_{(n)} \leq y_n,\, Y_{(1)} \leq y_1\} \ + \ P\{Y_{(n)} \leq y_n,\, Y_{(1)} > y_1\}$.

We find the CDF as follows:

$$\begin{aligned} F_{Y_{(1)},Y_{(n)}}(y_1,y_n) &\overset{\textbf{(b)}}{=} P\{Y_{(n)} \le y_n\} - P\{Y_{(1)} > y_1, Y_{(n)} \le y_n\} \\ &= P\{Y_j \le y_n \text{ for all } j\} - P\{y_1 < Y_j \le y_n \text{ for all } j\} \\ &= \prod_{j=1}^n P\{Y_j \le y_n\} - \prod_{j=1}^n P\{y_1 < Y_j \le y_n\} = \big[F(y_n)\big]^n - \big[F(y_n) - F(y_1)\big]^n . \end{aligned}$$

We used first independence, then identical distribution in the last line.
Differentiation of the above then gives us $f_{Y_{(1)},Y_{(n)}}(y_1,y_n)$

Alternatively, the PDF can be found by interpreting certain events related to finding the density as the outcomes of the following multinomial sequence, $\vec{X} = (X_1,\dots,X_n)$,

(c) For a given j, the outcomes ω_i' and associated probabilities p_i for X_j are
⊡ ω_1': Y_j is close to y_1 $\Rightarrow$ $p_1 = f(y_1)\,dy_1$ ⊡ ω_2': Y_j is close to y_n; $\Rightarrow$ $p_2 = f(y_n)\,dy_n$
⊡ ω_3': Neither ω_1' nor ω_2' happens and $y_1 < Y_j > y_n$.; $\Rightarrow$ $p_3 = F(y_n) - F(y_1)$.

(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 [37] 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 9.3 (CDF and PDF of the jth order statistic) and 9.4 (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

$$= P\{W_1 = 1, W_2 = 1, W_3 = n-2\} = p_{\vec{W}}(1,1,n-2) = \binom{n}{1,1,n-2} p_1^1\, p_2^1\, p_k^{n-2} .$$

$$= n(n-2)\cdot f(y_1)\,dy_1 \cdot f(y_n)\,dy_n \cdot F(y_n) - F(y_1).$$

(f) Thus, $f_{Y_{(1)},Y_{(n)}}(y_1,y_n)\,dy_1\,dy_n \overset{\textbf{(e)}}{=} n(n-2)\cdot f(y_1)\cdot f(y_n)\cdot \big[F(y_n) - F(y_1)\big]^{n-2} dy_1\,dy_n$.

- We cancel $dy_1\,dy_n$ in that last equation and obtain

(g) $$f_{Y_{(1)},Y_{(n)}}(y_1,y_n) = n(n-2)\cdot f(y_1)\cdot f(y_n)\cdot \big[F(y_n) - F(y_1)\big]^{n-2}. \quad \square$$

Remark 9.11 (Sample median)**.** Recall from Definition 7.4 (pth quantile) on p.90 that the median of a random variable U with CDF $F_U(\cdot)$ was its 0.5th quantile

$$\phi_{0.5} = \min\{u \in \mathbb{R} : F_U(u) \ge 0.5\} .$$

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 $\phi_{0.5}$.

Assume that $\vec{Y} = (Y_1,\dots,Y_n)$ 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 $\vec{y} = (y_1,\dots,y_n)$, where

$$y_1 = Y_1(\omega),\; y_2 = Y_2(\omega),\; \dots\; y_n = Y_n(\omega) .$$

[37] See Definition 8.17 (Multinomial distribution) on p.130.

Further assume that the Y_j are continuous. Then we can assume that all sample picks $Y_1, \dots, Y_n$ 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 is odd, then the sample median of $\vec{Y}$ is is the $(n+1)$th order statistic $Y_{(n+1)}$.
- **(b)** If n is even, then the sample median of $\vec{Y}$ is is the (random) average $\dfrac{Y_{n/2} + Y_{n/2+1}}{2}$.

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 sampling action is an order statistic, whereas that of an even sized one is not.

Let us assume that the the sample picks of an odd sized sample $\vec{Y} = \big(Y_1, \dots, Y_{2n+1}\big)$ are not only continuous random variables, but also iid. 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 9.12. Here are two observations about n iid random variables $Y_1, \dots, Y_n$.

- **(a)** Assume that $Y_{k_1}, \dots, Y_{k_n}$ is a permutation (ANY permutation!!) of $Y_1, \dots, Y_n$. 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
$$P\{Y = Y_{(1)}\} \;=\; P\{Y = Y_{(2)}\} \;=\; \dots\; P\{Y = Y_{(n)}\}\,.$$
Since there are n such arrangements, each one of those probabilities equals $\dfrac{1}{n}$. $\square$

9.5 The Method of moment–generating Functions

Assumption 9.2. Unless stated otherwise, we will assume in this entire section that

(a) $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ 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) $h : D \to \mathbb{R}; \quad \vec{y} \mapsto u = h(\vec{y}) = h(y_1, \dots, y_n)$
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, $m_{Y_j}(t) = E\left[e^{tY_j}\right]$ and $m_U(t) = E\left[e^{tU}\right]$ 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. $V = g(\vec{\widetilde{Y}}) = g(\widetilde{Y}_1, \dots, \widetilde{Y}_k)$. □

Introduction 9.3. The moment–generating function method for finding the probability distribution of a function of random variables $Y_1, Y_2, \dots, Y_n$ is based on Proposition 6.4 on p.85 (Section 6.5: 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 9.2:

Assume that two random variables Y and $\widetilde{Y}$ possess identical kth moments about the origin for all $k = 1, 2, \dots$. In other words, assume that

$$E[Y^1] = E[\widetilde{Y}^1],\ E[Y^2] = E[\widetilde{Y}^2],\ E[Y^3] = E[\widetilde{Y}^3], \dots$$

Then $P_Y = P\widetilde{Y}$, i.e., Y and $\widetilde{Y}$ have the same distribution. □

We have the following uniqueness theorem.

Theorem 9.5 (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 6.18 on p.85 and 7.7 on p.94 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 6.4 on p.85 that $P_Y = P_{\widetilde{Y}}$ ■

Remark 9.13.

To find the distribution of $U = h(\vec{Y}) = h\big(Y_1, Y_2, \dots, Y_n\big)$ by means of the MGF method, proceed as follows:
- Compute the MGF $m_U(t) = E\left[e^{tU}\right]$ 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 [5] Wackerly, Mendenhall, Scheaffer, R.L.
- Then you are done, since Theorem 9.5 (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 $U = a_1Y_1 + \cdots + a_nY_n$, with constant $a_j \in \mathbb{R}$.
- The random variables $Y_1, \dots, Y_n$ 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. $\square$

Example 9.14 (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 7.4 on p.96, $m_Y(t) = e^{\mu t + (\sigma^2 t^2)/2}$.

(b) Any random variable W is independent from any constant (real number) a.

(c) Thus, according to Theorem 8.9 on p.117, 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, $m_U(t) = m_Y(t)\, e^{-t\mu} \overset{\textbf{(a)}}{=} e^{\mu t + (\sigma^2 t^2)/2} = e^{(\sigma^2 t^2)/2}$.
- 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 7.4 once more and see that $t \mapsto e^{t^2/2}$ is the MGF of a standard normal random variable. Thus, $Z \sim \mathcal{N}(0,1)$. $\square$

Example 9.15 (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

$$\textbf{(A)} \quad \begin{aligned} m_{Z^2}(t) &= E(e^{tZ^2}) = \int_{-\infty}^{\infty} e^{tz^2} f(z)\,dz = \int_{-\infty}^{\infty} e^{tz^2} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(z^2/2)(1-2t)}\,dz\,. = \int_{-\infty}^{\infty} \psi(z)\,dz\,, \end{aligned}$$

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 7.6 on p.99, $t \mapsto \dfrac{1}{(1-2t)^{1/2}}$ is the MGF of a random variable which follows a gamma$(1/2, 2)$ distribution which is, by definition 7.11 on p.100, also known as a χ^2 distribution with one degree of freedom. We obtained this result previously in Example 9.4 on p.139 by the method of distribution functions. □

Theorem 9.6 (MGF of a sum of functions of independent variables)**.**

Given are n independent random variables $Y_1, Y_2, \dots, Y_n$ with MGFs $m_{Y_1}(t), m_{Y_2}(t), \dots, m_{Y_n}(t)$. and n real–valued functions $h_1(y_1), \dots, h_n(y_n)$ of real numbers $y_1, \dots, y_n$.
Let $U := h_1(Y_1) + h_2(Y_2) + \cdots + h_n(Y_n)$. Then (under the conditions of Assumption 9.2 on 160)

(9.42) $$m_U(t) = m_{h_1(Y_1)+\cdots+h_n(Y_n)} = \prod_{j=1}^{n} m_{h_j(Y_j)}(t)\,.$$

PROOF:
For each $j = 1, \dots, n$, 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 8.9 on p.117 and obtain

$$\begin{aligned} m_U(t) &= E\big[e^{t(V_1+V_2+\cdots+V_n)}\big] \\ &= E\big[e^{tV_1}\big]\cdots E\big[e^{tV_n}\big]\,. = E\big[e^{th_1(Y_1)}\big]\cdots E\big[e^{th_n(Y_n)}\big] \\ &= m_{h_1(Y_1)}(t)\cdot m_{h_1(Y_1)}(t)\cdots m_{h_1(Y_n)}(t)\,. \end{aligned}$$ ■

Corollary 9.1 (WMS Ch.06.5, Theorem 6.2)**.**

Let $Y_1, Y_2, \dots, Y_n$ be independent random variables with moment–generating functions $m_{Y_1}(t), m_{Y_2}(t), \dots, m_{Y_n}(t)$, respectively. Then

(9.43) $$m_{Y_1+\cdots+Y_n}(t) = \prod_{j=1}^{n} m_{Y_j}(t) = m_{Y_1}(t)\cdot m_{Y_2}(t)\cdots m_{Y_n}(t)\,.$$

PROOF:
This follows from applying Theorem 9.6 to the functions $h_j(y_j) = y_j$. ■

Next, we generalize But its great importance gives it the status of a theorem.

Theorem 9.7 (Linear combinations of uncorrelated normal variables are normal)**.**

Given are n uncorrelated, $\mathcal{N}(\mu_j, \sigma_j^2)$ random variables Y_j, $(j = 1, \dots, n$. In other words, each Y_j is normal with expectation μ_j and standard deviation σ_j. Let $a_1, \dots, a_n \in \mathbb{R}$. Then

$$\sum_{j=1}^{n} a_j Y_j \sim \mathcal{N}\left(\sum_{j=1}^{n} a_j \mu_j \,, \sum_{j=1}^{n} a_j^2 \sigma_j^2\right) . \tag{9.44}$$

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 9.6 (MGF of a sum of functions of independent variables) on p.163 with the functions $h_j(y_j) = ajy_j$ 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 $\mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ variable has MGF $e^{\widetilde{\sigma}^2 t^t + \widetilde{\mu} t}$. See Proposition 7.4 on p.96. 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 7.4, the last expression is the MGF of a $\mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ 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,

$$U \sim \mathcal{N}\left(\sum_{j=1}^{n} a_j \mu_j \,, \sum_{j=1}^{n} a_j^2 \sigma_j^2\right) . \ \blacksquare$$

Theorem 9.8.

Given are n independent, gamma(α_j, β) random variables Y_j, $(j = 1, \dots, n$. In other words, each Y_j is gamma with the same scale parameter β. Then

$$\sum_{j=1}^{n} Y_j \sim gamma\left(\sum_{j=1}^{n} \alpha_j\,, \beta\right). \tag{9.45}$$

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 9.6 (MGF of a sum of functions of independent variables) on p.163 and recall that the MGF of a gamma$(\widetilde{\alpha}, \widetilde{\beta})$ variable $\widetilde{Y}$ is, according to Proposition 7.6 on p.99, $m_{\widetilde{Y}} = (1 - \widetilde{\beta}t)^{\widetilde{\alpha}}$. We obtain

$$\begin{aligned} m_U(t) &= \prod_{j=1}^{n} m_{a_j Y_j}(t) = \prod_{j=1}^{n} m_{Y_j}(a_j t) \\ &= \prod_{j=1}^{n} \frac{1}{(1-\beta t)^{\alpha_j}} = \frac{1}{(1-\beta t)^{\sum_{j=1}^{n} \alpha_j}}\,. \end{aligned}$$

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 9.2.

Let $Y_1, Y_2, \dots, Y_n$ be independent χ^2 variables such that each Y_j has ν_j degrees of freedom. Then

$$m_{Y_1 + \dots + Y_n}(t) \sim \chi^2\left(\sum_{j=1}^{n} \nu_j \; df\right). \tag{9.46}$$

PROOF:
This follows immediately from Theorem 9.8, Since χ^2 variables with ν_j df are gamma$(\nu_j/2, 2)$. $\blacksquare$

10 Limit Theorems

Introduction 10.1. 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 $k = 1, 2, \dots$. 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. [38]

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" [39] of the functions f_n to the function f simply means that

(10.1) $$\lim_{n\to\infty} f_n(a) = f(a) \qquad \text{for all } a \in A\,.$$

- Let $f_n, f, g, h : [0,1] \to \mathbb{R}$ be the functions

(10.2) $$\boxdot\; f_n(x) := x^n \quad \boxdot\; f(x) := \begin{cases} 0\,, & \text{if } 0 \le x < 1, \\ 1\,, & \text{if } x = 1\,, \end{cases} \quad \boxdot\; g(x) := 0, \quad \boxdot h(x) := x.$$

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 $\{0\}$.

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: (10.1) becomes

  (10.3) $$\lim_{n\to\infty} f_n(\omega) = f(\omega) \qquad \text{for all } \omega \in \Omega$$

  and one certainly can examine whether or not the above is true for any kind of Ω.

We will not discuss vector–valued sequences. However, for completeness sake, we give the following example.

- If $\vec{y}_n = \big((-1)^n, \cos(2/n)\big)$, then $\lim\limits_{n\to\infty} \vec{y}_n$ 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.

[38] There is no such thing as divergence to $\pm\infty$. Thus, you must say that (y_n) diverges, **not** that (y_n) diverges to ∞.

[39] The formal definition of pointwise limits will be given in Section 10.1 (Four Kinds of Limits for Sequences of Random Variables).

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 $\lim\limits_{n\to\infty} Y_n(\omega)$ exist and will it be a real number (rather than $\pm\infty$) for all $\omega \in \Omega$?
  ⊡ If so, then the assignment $\omega \mapsto Y(\omega) := \lim\limits_{n\to\infty} Y_n(\omega)$ 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, [40] we will see in Section 10.1 (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. □

10.1 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 10.1 (Convergence of Random Variables)**.**

Let Y_n $(n \in \mathbb{N})$ and Y be random variables on a probability space (Ω, P). We define

(10.4) $Y_n \overset{\textbf{pw}}{\to} Y$ or $\text{pw} - \lim\limits_{n\to\infty} Y_n = Y\,,\quad$ if $\lim\limits_{n\to\infty} Y_n(\omega) = Y(\omega),\;$ for all $\omega \in \Omega\,,$

(10.5) $Y_n \overset{\textbf{a.s.}}{\to} Y$ or $\text{a.s.} - \lim\limits_{n\to\infty} Y_n = Y\,,\quad$ if $P\{\omega \in \Omega : \lim\limits_{n\to\infty} Y_n(\omega) \neq Y(\omega)\} = 0\,,$

(10.6) $Y_n \overset{\textbf{P}}{\to} Y$ or $\text{P} - \lim\limits_{n\to\infty} Y_n = Y\,,\quad$ if $\forall\, \varepsilon > 0\; \lim\limits_{n\to\infty} P\{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \varepsilon\} = 0\,,$

(10.7) $Y_n \overset{\textbf{D}}{\to} Y,\;$ if $\lim\limits_{n\to\infty} F_{Y_n}(y) = F_Y(y),\; \forall\, y \in \mathbb{R}$ where the CDF F_Y of Y is continuous.

[40] 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.

We also say:
If $Y_n \overset{\textbf{pw}}{\to} Y$, Y is the **pointwise limit** of the Y_n, or: Y_n **converges pointwise** to Y.
If $Y_n \overset{\textbf{a.s.}}{\to} Y$, Y is the **almost sure limit** of the Y_n, or: Y_n **converges almost surely** to Y.
If $Y_n \overset{\textbf{P}}{\to} Y$, Y is the **limit in probability**; of the Y_n, or: Y_n **converges in probability** to Y.
If $Y_n \overset{\textbf{D}}{\to} Y$, Y is the **limit in distribution** of the Y_n, or: Y_n **converges in distribution** to Y.

Example 10.1. 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 (10.2) 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

$$Y_n(\omega) \;:=\; \omega^n, \quad U(\omega) \;=\; 0, \quad V(\omega) := \omega, \quad (\text{for } 0 \le \omega \le 1) \quad Y(\omega) \;:=\; \begin{cases} 0\,, & \text{if } 0 \le \omega < 1, \\ 1\,, & \text{if } \omega = 1\,. \end{cases}$$

Part I: Pointwise and a.s convergence

Pointwise convergence behavior of the Y_n corrresponds to that of (10.2):
- 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 $[\Omega$,
- V is not the pointwise limit of the Y_n (except for $\omega = 0$).

With respect to almost sure convergence, we see that
- $Y_n \overset{\textbf{a.s.}}{\to} Y$, since $\{\lim\limits_{n\to\infty} Y_n = Y\} \;=\; [0,1] \;=\; \Omega$, and $P(\Omega) = 1$.
- $Y_n \overset{\textbf{a.s.}}{\to} U$, since $\{\lim\limits_{n\to\infty} Y_n \ne U\} \;=\; \{1\}$, and $P(\{1\}) = 0$.
- $(Y_n)_n$ does not converge to V a.s., since $P\{\lim\limits_{n\to\infty} Y_n = V\} \;=\; P\{0\} \ne 1$.

Part II: Convergence in probability

By definition of $P\text{–}\lim\limits_{n\to\infty} Y_n = \widetilde{Y}$, we must prove that, for any fixed, but arbitrary $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{|Y_n - \widetilde{Y}| > \varepsilon\} \;=\; 0. \qquad \text{See (10.6).}$$

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, since $Y_n(\omega) = \omega^n$ and $0 < \varepsilon < 1$,

(A)
$$\begin{gathered} \big[\, |Y_n(\omega) \ge \varepsilon| \;\Leftrightarrow\; \omega^n \ge \varepsilon \;\Leftrightarrow\; \omega \ge \varepsilon^{1/n} \,\big] \\ \Rightarrow\; \big[\, P\{|Y_n \ge \varepsilon|\} \;=\; P\{\varepsilon^{1/n}, 1]\} \;=\; 1 \;-\; \varepsilon^{1/n} \,\big]\,. \end{gathered}$$

(B)
$$\varepsilon^{1/n} \to 0, \text{ as } n \to \infty. \text{ Thus, } \lim_{n\to\infty} \big(1 - \varepsilon^{1/n}\big) \;=\; 1.$$

(C) $\omega \in \Omega \Rightarrow P\{\omega)\} = 0$. In particular, $P\{1\} = 0$.

Part II (1): We now prove that $P\text{–}\lim\limits_{n\to\infty} Y_n = Y$:

(a)
$$\big[\, |Y_n(\omega) - Y(\omega)| \geq \varepsilon \Leftrightarrow |Y_n(\omega)| \geq \varepsilon \text{ and } \omega \neq 1 \big]$$
$$\Rightarrow \big[\, P\{|Y_n - Y| \geq \varepsilon\} \leq P\{|Y_n| \geq \varepsilon\} \overset{\textbf{(A)}}{=} 1 - \varepsilon^{1/n} \overset{\textbf{(B)}}{\to} 0, \text{ as } n \to \infty. \big].$$

Thus, $\lim\limits_{n\to\infty} P\{|Y_n - Y| \geq \varepsilon\} = 0$.

Part II (2): We now prove that $P\text{–}\lim\limits_{n\to\infty} Y_n = U$:

- 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 $P\text{–}\lim\limits_{n\to\infty} Y_n = U$
- 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 $\lim\limits_{n\to\infty} P\{|Y_n - Y| \geq \varepsilon\} = 0$,

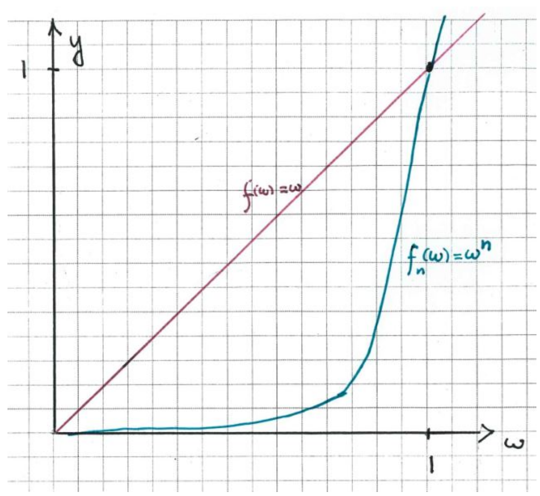
$$\lim_{n\to\infty} P\{|Y_n - U| \geq \varepsilon\} = \lim_{n\to\infty} P\{|Y_n - Y| \geq \varepsilon\} = 0.$$

Thus, $P\text{–}\lim\limits_{n\to\infty} Y_n = U$.

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, $\lim\limits_{n\to\infty} P\{|Y_n - V| \geq \varepsilon\} = 0$ is not true.
- Since $\lim\limits_{n\to\infty} P\{|Y_n - V| \geq \varepsilon\} = 0$ 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.

- Recall that $P]a,b] = b - a$ for all $0 \leq a < b \leq 1$. Let $y \in \mathbb{R}$.
- Since $V(\omega) = \omega$, $F_V(y) = P\{V \leq y\} = P\{\omega \in \Omega : V(\omega) \leq y\} = P]0,y] = y$.
- Since $Y_n(\omega) = \omega^n$, $F_{Y_n}(y) = P\{Y_n \leq y\} = P\{\omega \in \Omega : \omega^n \leq y\} = P]0, y^{1/n}] = y^{1/n}$.
- Thus, for $0 < y < 1$, $F_V(y) = y$, whereas, $\lim\limits_{n\to\infty} F_{Y_n}(y) = 0 \neq F_V(y)$.
- 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 converges in probability to those random variables. $\square$

Theorem 10.1 (Relationship between the modes of convergence).

Let Y and $Y_1, Y_2, \dots$ be random variables on a probability space (Ω, P). Then,

$$Y_n \overset{pw}{\to} Y \;\Rightarrow\; Y_n \overset{a.s.}{\to} Y \;\Rightarrow\; Y_n \overset{P}{\to} Y \;\Rightarrow\; Y_n \overset{D}{\to} Y\,. \tag{10.8}$$

PROOF:

I: It is obvious that $Y_n \overset{\textbf{pw}}{\to} Y \;\Rightarrow\; Y_n \overset{\textbf{a.s.}}{\to} Y$ for the following reason:

- For each $n \in \mathbb{N}$, let $A_n := \{\omega \in \Omega : \lim\limits_{n\to\infty} Y_n(\omega) \neq Y(\omega)\}$.
- Then, for $k \in \mathbb{N}$, $Y_n \overset{\textbf{pw}}{\to} Y \;\Rightarrow\; A_k = \emptyset \;\Rightarrow\; P(A_k) = 0 \;\Rightarrow\; \lim\limits_{n\to\infty} P(A_n) = 0 \;\Rightarrow\; Y_n \overset{\textbf{a.s.}}{\to} Y$.

II: Proof that $Y_n \overset{\textbf{a.s.}}{\to} Y \;\Rightarrow\; Y_n \overset{\textbf{P}}{\to} Y$:

- For $n \in \mathbb{N}$ and $\varepsilon > 0$, let

$$A_n := \{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) \neq Y(\omega)\}\,, \qquad A_n(\varepsilon) := \{\omega \in \Omega : \lim_{n\to\infty} |Y_n(\omega) - Y(\omega)| \geq \varepsilon\}\,.$$

- Then, $A_n = \{\omega \in \Omega : \lim\limits_{n\to\infty} |Y_n(\omega) - Y(\omega)| \neq 0\} \supseteq A_n(\varepsilon)$. $\Rightarrow P\big(A_n(\varepsilon)\big) \leq P(A_n)$.
  $\Rightarrow \lim\limits_{n\to\infty} P\big(A_n(\varepsilon)\big) \leq \lim\limits_{n\to\infty} P(A_n) = 0$ (since $Y_n \overset{\textbf{a.s.}}{\to} Y$).
- We are done, since, $\lim\limits_{n\to\infty} P\big(A_n(\varepsilon)\big) = 0 \;\Rightarrow\; Y_n \overset{\textbf{P}}{\to} Y$.

III: Proof that $Y_n \overset{\textbf{P}}{\to} Y \;\Rightarrow\; Y_n \overset{\textbf{D}}{\to} Y$:
Will not be given here. A very accessible proof can be found at this Wikipedia link. $\blacksquare$

Example 10.2 (Convergence in probability but not a.s.). $\boxed{\star}$

Consider the "sliding hump" example. [41] As our probability space we choose $\Omega := [0,1]$, the unit interval in $\mathbb{R}$, with the probability measure defined by $P\big(]a,b]\big) := b - a$.

[41] 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 $s_1 = 1,\ s_2 = s_1 + 2,\ s_3 = s_2 + 3, \dots, s_k = s_{k-1} + k = \sum\limits_{j=1}^{k} j = \dfrac{k \cdot (k+2)}{2}, \dots,$

(i) then $I_n = 1/k \Leftrightarrow n = s_{k-1} + 1,\ s_{k-1} + 2, \dots, s_{k-1} + k \Leftrightarrow s_{k-1} < n \le s_k$.

(j) It should be clear that $\left[\, n \to \infty \,\right] \left[\, k \to \infty \,\right]$ For a proof: $\boxdot$ "$\Leftarrow$" follows from $n \ge k$.
$\boxdot$For the other direction, we observe that $n \overset{\textbf{(i)}}{\le} 2s_k = 2k(k+1) < 2(k+1)^2$,
i.e., $\sqrt{n/2} - 1 < k$. Thus, $\left[\, n \to \infty \,\right] \Rightarrow \left[\, k \to \infty \,\right]$ 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 $\left[\, n \to \infty \,\right] \overset{\textbf{(j)}}{\Rightarrow} \left[\, k \to \infty \,\right]$, it follows that $\lim\limits_{n\to\infty} P\{|Y_n| > \varepsilon\} = 0$ and this shows that $Y_n \xrightarrow{P} 0$. $\blacksquare$

$\square$

10.2 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 10.2 (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,

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \text{ converges in probability to } \mu, \quad \text{i.e.,}$$

(10.9)
$$\left[\, \varepsilon > 0 \,\right] \quad \Rightarrow \quad \left[\lim_{n\to\infty} P\left\{ \left| \frac{1}{n} \sum_{j=1}^{n} Y_j - \mu \right| > \varepsilon \right\} = 0. \right]$$

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 8.3 (Variance of the sample mean) on p.122, that

(A)
$$\mu_{\bar{Y}_n} = E\left[\bar{Y}_n\right] = \mu, \qquad \text{and} \qquad \sigma^2_{\bar{Y}_n} = Var\left[\bar{Y}_n\right] = \frac{\sigma^2}{n}.$$

We apply Tchebysheff's inequality 7.43 on p.103 with $k = \varepsilon\sqrt{n}/\sigma$ and obtain from **(A)**, that

$$P\left\{ |\bar{Y}_n - \mu| > \varepsilon \right\} \leq \frac{1}{(n\,\varepsilon^2/\sigma)^2} = \frac{\sigma^2}{n\,\varepsilon^2} \to 0, \text{ as } n \to \infty$$

This proves that $P\text{–}\lim\limits_{n\to\infty} \bar{Y}_n = \mu$. ■

Remark 10.1. 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 8.3 (Variance of the sample mean) on p.122.

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 10.2. We have learned in Theorem 10.1 (Relationship between the modes of convergence) on p.170, 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 10.2 on p.172. It is called the **weak** law of large numbers because there also is a **strong** law of large numbers which replaces the conclusion $P\!-\!\lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{j=1}^{n} Y_j = \mu$ with a.s.$-\lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{j=1}^{n} Y_j = \mu$. We will study that next.

□

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 10.3 (Strong Law of Large Numbers)**.**

Let $Y_1, Y_2, \dots$ be an iid sequence of random variables on a probability space (Ω, P).
Let $\mu := E[Y_n]$. Then,

(10.10)
$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \text{ converges almost surely to } \mu, \quad \text{i.e.,}$$
$$P\left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} Y_j \neq \mu \right\} = 0\,.$$

PROOF:

Outside the scope of these lecture notes. ■

Example 10.3 (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 $X_1, X_2, \dots$ of key strokes is an iid sequence (a Bernoulli sequence) with success probability $p = 1/K$.
- We consider the indices $1, 2, 3, \dots$ 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 consists of, but this certainly is a finite number. Let us denote it by N.

Let $Y_1 := 1$, if $X_1, X_2, \dots, X_N$ form S-C-W. Let $Y_1 := 0$, else.
Let $Y_2 := 1$, if $X_{N+1}, X_{N+2}, \dots, X_{2N}$ form S-C-W. Let $Y_2 := 0$, else.
- -
Let $Y_j := 1$, if $X_{(j-1)N+1}, X_{(j-1)N+2}, \dots, X_{jN}$ form S-C-W. Let $Y_j := 0$, else.
- -

- If $i \neq j$, then Y_i and Y_j depend on "disjoint" chunks $\big(X_{(i-1)N+1}, X_{(i-1)N+2}, \dots, X_{iN}\big)$ and $\big(X_{(j-1)N+1}, X_{(j-1)N+2}, \dots, X_{jN}\big)$ 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 1. Thus,

$$\omega \in A \;\Rightarrow\; Y_j(\omega) \;=\; 1\,, \text{ infinitely often!}$$

- Since $P(A) = 1$ and Y_i denotes the completion of the nth collection of Shakespeare's works:
- With probability 1, the monkey will produce an infinite number of Shakespeare's entire collection! $\square$

10.3 Sampling Distributions

Introduction 10.2. Back in Chapter 5.2 (Random Sampling and Urn Models With and Without Replacement), we gave Definition 5.2 (Sampling as a Random item) on p.68 of a sampling action.

- A sampling action of size n was nothing but a vector $\vec{X} = \big(X_1, X_2, \dots, X_n\big)$ of random items. What makes it a sampling action 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 we may consider the sample picks as having identical distribution. Thus,

$$F_{X_1}(x) \;=\; F_{X_2}(x) \;=\cdots=\; F_{X_n}(x) \quad (x \in \mathbb{R})\,;$$

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 and variance and .....

Moreover, nothing is assumed about the independence of the sample picks. To have it would be extremely desirable from a mathematical perspective. For example, if the X_j are jointly continuous random variables, knowledge of the marginal densities yields the joint density, because,

$$f_{\vec{X}}(\vec{x}) \;=\; f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n) \quad (\vec{x} \in \mathbb{R}^n)\,.$$

Unfortunately, dentical distribution and independence are simplifications of the real world. This is even true when one considers n rolls of a die. [42] 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 action according to Definition 5.3 (Simple Random Sample) on p.70:

- A sampling action $\vec{X} = \big(X_1, X_2, \dots, X_n\big)$ of size n from a population of size $N \geq n$ is called a simple random sampling action (SRS action), 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 action is large, but small when compared to the size of the population, then treating it as iid will result in insignificant domputational differences. [43] This observation is one of the reasons that even the more restrictive definition of an SRS action is of a generality we are not looking for in this chapter. We follow [2] 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 do not even know the type of distribution: Does Y follow a Poisson distribution or is it normal or exponential or .....?

(II) 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.

We deal in this section with problem **(I)**. $\square$

Example 10.4. 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 to make explicit the role of the unknown parameter, p.

(b) $Y \sim \mathscr{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We write $f_Y(y; \mu, \sigma)$ for the PDF to make explicit the role of the unknown parameters, μ and σ.

(c) $Y \sim$ expon(β), with unknown β. We write $f_Y(y; \mu, \beta)$ for the PDF.

(d) $Y \sim$ gamma$(\alpha, 3)$, with unknown α. We write $f_Y(y; \alpha)$ for the PDF. $\square$

[42] Interpret X_j as the jth pick from the population of all rolls of that die.

[43] We mentioned this in Remark 5.2 on p.69.

Remark 10.3. 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 $\vec{theta}$ 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.

Thus, in Example 10.3**(a)**, $\Theta = [0,1]$. In Example 10.3**(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.
- How can we find a good, possibly optimal, procedure to estimate 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 function

$$\theta \;=\; T(\vec{y}) \;=\; T(y_1,\dots,y_n) \;=\; T\big(\vec{Y}(\omega)\big) \;=\; T\big(Y_1(\omega),\dots,Y_n(\omega)\big)\,.$$

In the context of a sampling action, we refer to the specific list of numbers, $\vec{y} = (y_1,\dots,y_n)$, as the values or **realizations**, of the sampling action. $\square$

We had stated in the introduction that we will restrict the scope of the sampling actions in this section to the iid case.

Definition 10.2 (Random sampling action from a distribution)**.**

Let Y be a random variable on a probability space (Ω, P). We call a vector $\vec{Y} = \big(Y_1,\dots,Y_n\big)$ a **random sampling action from the distribution of** Y, or also, a **random sampling action on** Y, if

- each Y_j has the same distribution as Y
- the random variables $Y_1,\dots,Y_n$ are iid. $\square$

That definition allows us to restate the essence of Remark 10.4 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 sampling action on Y.

Definition 10.3 (Statistic)**.**

Let Y be a random variable on a probability space (Ω, P) and $\vec{Y} = \big(Y_1, \dots, Y_n\big)$ 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,

$$B \mapsto P_{T\circ\vec{Y}}(B) = P\{T(\vec{Y}) \in (B)\} = P\{\omega \in \Omega : T\big(\vec{Y}(\omega)\big) \in B\} \tag{10.11}$$

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 10.4.

Let Y be a random variable on a probability space (Ω, P) and $\vec{Y} = \big(Y_1, \dots, Y_n\big)$ a random sampling action on Y. Let $T_1, T_2, \dots, T_k : \mathbb{R}^n \mapsto \mathbb{R}$ be statistics for that sample action. Let

$$\widetilde{T} : \mathbb{R}^k \mapsto \mathbb{R}\,; \qquad (t_1, \dots, t_k) \mapsto \widetilde{T}\big((t_1, \dots, t_k)\big).$$

Then, setting $\vec{t} = (t_1, \dots, t_k)$ and $\vec{T} = (T_1, \dots, T_k)$, the composition

$$T^* \circ \vec{T}(\vec{Y}) : \omega \mapsto T^*\big(T_1(\vec{Y}), \dots, T_k(\vec{Y})\big)$$

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. ■

Here is an example of a statistic which is so important that it deserves its own definition. It also is used to illustrate Theorem 10.4.

Definition 10.4 (Sample variance)**.**

Let $\vec{Y} = \big(Y_1, \dots, Y_n\big)$ be a random sample action on a random variable Y.
The **sample variance** is defined as the random variable

$$\omega \mapsto S^2(\omega) := \frac{1}{n-1} \sum_{j=1}^{n} \big(Y_j(\omega) - \bar{Y}(\omega)\big)^2 . \tag{10.12}$$

We further call $\omega \mapsto S(\omega) := \sqrt{S^2(\omega)}$ the The **sample standard deviation**.
We write s^2 and s for the realizations $S^2(\omega)$ and $S(\omega)$ that result from creating the sample.
□

Example 10.5. For the following examples assume that $\vec{Y} = (Y_1, \dots, Y_n)$ is a random sample action on a random variable Y.

(a) In Example 8.3 (Variance of the sample mean) on p.122, we considered the sample mean $\omega \mapsto \bar{Y}(\omega) = \frac{1}{n}\sum\limits_{j=1}^{n} Y_j(\omega)$. $\bar{Y}$ is a statistic: The transform is $T(\vec{Y}) = \frac{1}{n}\sum\limits_{j=1}^{n} Y_j$.
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 10.4 on p.177 as follows. Let

$$t_1 = T_1(\vec{y}) = y_1,\ t_2 = T_2(\vec{y}) = y_2, \dots, t_n = T_n(\vec{y}) = y_n,\ t_{n+1} = T_{n+1}(\vec{y}) = \bar{y}.$$

$$T^*(t_1, \dots, t_n, t_{n+1}) = \frac{1}{n-1}\sum_{j=1}^{n}(t_j - t_{n+1})^2$$

Then $S^2 = T^*\big(T_1(\vec{Y}), \dots, T_n(\vec{Y})), T_{n+1}(\vec{Y})\big)$. By Theorem 10.4, S^2 is a statistic for the randpom sampling action $\vec{Y}$. 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 $y_1, \dots, y_n$ 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 10.6 (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 10.5 (WMS Ch.07.2, Theorem 7.1)**.** ()

Let $Y_1, Y_2, \dots, Y_n$ be a random sampling action 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 9.7 (Linear combinations of uncorrelated normal variables are normal) on p.163. ■

Theorem 10.6 (WMS Ch.07.2, Theorem 7.2)**.**

Let $\vec{Y} = (Y_1, \dots, Y_n)$ be a random sampling action on $Y \sim \mathcal{N}(\mu, \sigma^2)$. Let $Z_j = (Y_j - \mu)/\sigma$ for $j = 1, 2, \dots, n$. Then $\vec{Z} = (Z_1, \dots, Z_n)$ is a random sampling action on a standard normal variable. (In particular, the Z_+j are iid.) Further,

$$\sum_{j=1}^{n} Z_i^2 = \sum_{j=1}^{n}\left(\frac{Y_j - \mu}{\sigma}\right)^2 \tag{10.13}$$

follows a χ^2 distribution with n degrees of freedom.

PROOF: It follows from Theorem 9.7 (Linear combinations of uncorrelated normal variables are normal) on p.163 that the linear combination $Z_j = (Y_j - \mu/\sigma)$ is standard normal. It follows from Theorem 9.6 (MGF of a sum of functions of independent variables) on p.163 that the Z_j are iid. It follows from Theorem 9.8 on p.165 that $\sum\limits_{j=1}^{n} Z_i^2 \sim \chi^2(\mathrm{df} = n)$. ■

The following is Example Example 6.13 of the WMS text.

Proposition 10.1. $\boxed{\star}$

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 10.7 (WMS Ch.07.2, Theorem 7.3). $\boxed{\star}$

Let $\vec{Y} = (Y_1, \dots, Y_n)$ be a random sample action on a $\mathcal{N}(\mu, \sigma^2$ random variable Y and let $Z_j = (Y_j - \mu)/\sigma$. Then $\vec{Z} = (Z_1, \dots, Z_n)$ is a random sample action on a standard normal variable Z. Further,

(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 10.5 (Student's t–distribution [44]). $\boxed{\star}$

Let Z and W be independent random variables such that Z is standard normal and$W$ is χ^2 with ν df. Let

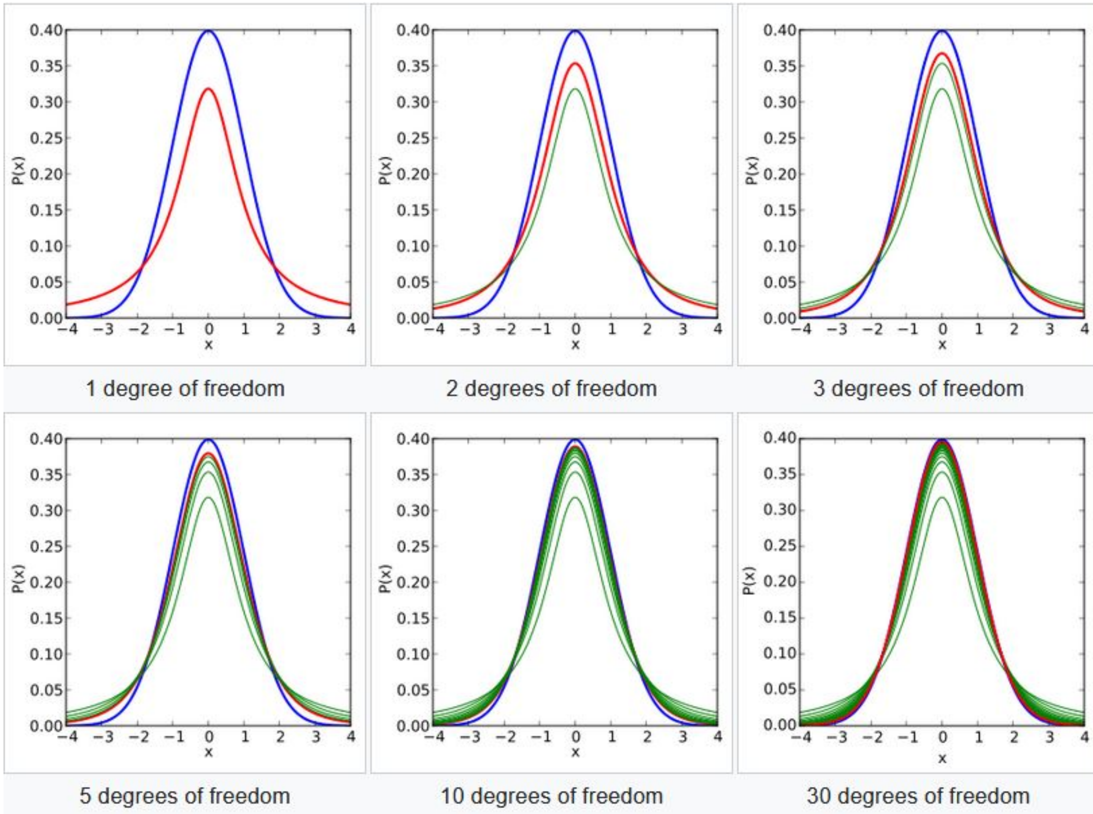
$$T = \frac{Z}{\sqrt{W/\nu}} \tag{10.14}$$

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 $T \sim t(\mathrm{df} = \nu)$. □

[44] 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".

Remark 10.4. 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 differences:

- The mean $E[T]$ is not a parameter: it is 0.
- The tails are fatter than those of a $\mathcal{N}(0,1)$ variable. See Figure 10.1. □



10.1 (Figure). **densities of the standard normal and t distribution.** Source: Wikipedia.

Theorem 10.8. $\boxed{\star}$

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $\vec{Y} = \big(Y_1, \dots, Y_n\big)$ be a random sample action on Y. Let

$$T := \frac{\bar{Y} - \mu}{S/\sqrt{n}}. \tag{10.15}$$

Then T follows a t–distribution with $(n-1)$ df.

PROOF: Let

$$Z := \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \qquad \text{and} \qquad W := \frac{(n-1)S^2}{\sigma^2}.$$

We have seen that $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2(\mathrm{df} = \nu)$. Since $\bar{Y}$ and S^2 are independent by Theorem

10.7 on p.179, Z as a function of $\bar{Y}$ only and W as a function of S^2 only also are independent. Thus,

$$T \;=\; \frac{Z}{\sqrt{W/\nu}} \;=\; \frac{\sqrt{n}(\bar{Y}-\mu)/\sigma}{[\sqrt{(n-1)S^2/\sigma^2}]/(n-1)} \;=\; \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right)$$

has at distribution with $(n-1)$ df. ■

Example 10.7 (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 sampling actions 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 10.6 (F–distribution)**.** $\boxed{\star}$

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 10.5. $\boxed{\star}$

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 10.9. $\boxed{\star}$

Consider two random sampling actions of sizes n_1 and n_2 on random variables $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ from two independent populations, with sample variances S_1^2 and S_2^2. Let

$$F \;:=\; \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\,. \tag{10.16}$$

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 sampling actions 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 Definition 10.6 of an F distribution,

$$\frac{W_1/\nu_1}{W_2/\nu_2} = \frac{[(n_1 - 1)S_1^2/\sigma_1^2]/[(n_1 - 1)}{[(n_2 - 1)S_2^2/\sigma_2^2/(n_2 - 1)]} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows an F distribution with $(n_1 - 1)$ numerator df and $(n_2 - 1)$ denominator df. ∎

Example 10.8 (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. □

10.4 The Central Limit Theorem

Introduction 10.3. In section10.3 (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 action $\vec{Y}$ 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. The limit we can show to exist is the least desirable of the four, the limit in distribution. But that is not as bad as it sounds

- 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 $P\{U \in B\} \approx P\{U_n \in B\}$, □

The limit theorem alluded to in the introduction stated and proven 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 10.10 (Lévy's continuity theorem)**.** $\boxed{\star}$

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

(10.17)
$$\begin{aligned} &\left[m_{Y_n}(t) \to m_Y(t) \text{ as } n \to \infty, \ \text{for all } t \right] \\ \Rightarrow\ &\left[F_{Y_n}(y) \to F_Y(y) \text{ as } n \to \infty, \ \text{for all } y \text{ where } F_Y(\cdot) \text{ is continuous.} \right] \end{aligned}$$

PROOF: Outside the scope of this course. ■

Theorem 10.11 (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

$$U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \qquad \text{where } n \in \mathbb{N}\,,\ \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i\,.$$

Then U_n converges to Z in distribution as $n \to \infty$. In other words,

$$\lim_{n\to\infty} P\{U_n \leq u\} = P\{Z \leq u\} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad \text{for all } u\,.$$

PROOF:

Let $U_n := \dfrac{\bar{Y}_n}{\sigma/\sqrt{n}}$.

(1) Since the $\widetilde{Y}_n := Y_n - \mu$ are iid, they have a common MGF, $m(t) = m_{\widetilde{Y}_n}(t)$.
By Corollary 9.1 on p.163, $m_{\widetilde{Y}_1 + \cdots \widetilde{Y}_n}(t) = \big[m(t)\big]^n$. Thus.

(2)
$$m_{U_n}(t) = E\left[\sum_{j=1}^{n} \widetilde{Y}_j \cdot \frac{t}{\sigma/\sqrt{n}}\right] = m_{\widetilde{Y}_1+\cdots\widetilde{Y}_n}\left(\frac{t}{\sigma/\sqrt{n}}\right) = \left[m\left(\frac{t}{\sigma/\sqrt{n}}\right)\right]^n.$$

(3) According to Theorem 10.10 (Lévy's 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 $x \mapsto \ln(x)$ is continuous and injective and it's inverse, $x \mapsto e^x$ also 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,

$$\textbf{(6)} \qquad \lim_{n\to\infty} \ln\, m_{U_n}(t) \;=\; \frac{t^2}{\sigma^2} \lim_{h\to 0} \frac{\ln m(h)}{h^2}\,.$$

Since $m(0) = e^0 = 1$, the right–hand limit is of the form $0/0$. We use L'Hôpital's rule [45] twice in a row and obtain, since $m(t) = m_{\widetilde{Y}_n}(t)$ and hence, $m''(0) = E[\widetilde{Y}_n^2]$,

$$\begin{aligned} \lim_{h\to 0} \frac{\ln m(h)}{h^2} \;&=\; \lim_{h\to 0} \frac{m(h)m'(h)}{2h} \;=\; \lim_{h\to 0} \frac{m'(h)}{2hm(h)} \\ \textbf{(7)} \qquad &=\; \lim_{h\to 0} \frac{m''(h)}{2m(h) + 2hm'(h)} \;=\; \frac{m''(0)}{2m(0)+0} \;=\; m''_{\widetilde{Y}_n}(0) \;=\; E[\widetilde{Y}_n^2] \end{aligned}$$

(8) Since $E[\widetilde{Y}_n] = \mu - \mu = 0$ and $Var[\widetilde{Y}_n] = Var[Y_n] = \sigma^2$, we obtain from **(7)** that $\lim\limits_{h\to 0} \dfrac{\ln m(h)}{h^2} \;=\; \dfrac{\sigma^2}{2}$.
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. ■

Example 10.9 (WMS Ch.07.3, Example 7.8)**.** Example 7.8 of the WMS text discusses a practical example of the use of the CLT (SAT scores). You are strongly encouraged to study it. □

Example 10.10 (WMS Ch.07.3, Example 7.9)**.** Example 7.9 of the WMS text discusses another practical example of the use of the CLT (checkout counter service times). You are strongly encouraged to study it. □

Example 10.11 (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 10.12 (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. □

[45] in the form $\lim\limits_{x\to 0} \dfrac{f(x)}{g(x)} \;=\; \lim\limits_{x\to 0} \dfrac{f'(x)}{g'(x)}$

11 Sample Problems for Exams

11.1 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 11.1. $\boxed{\text{Practice Midterm 1 (C. Haines) – \# 01}}$
SKIPPED □

Answer: N/A ■

Exercise 11.2. $\boxed{\text{Practice Midterm 1 (C. Haines) – \# 02}}$
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 $\vec{x} = (x_1, x_2, \dots, x_n)$, where $n \le 7$ 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 $x_2, \dots, x_4$ is H and the other three and x_5 are $L \Rightarrow \binom{4}{1} = 4$ choices
- or two of $x_2, \dots, x_5$ 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 $x_1, \dots, x_5$ 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 $x_1, \dots, x_6$ 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 $x_1, \dots, x_4$ is H and $x_5 = L \quad\Rightarrow 4\;$ 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,

$$P(A \mid B) \;=\; \frac{P(A \cap B)}{1 - P(B^{\complement})} \;=\; \frac{10 \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^4 \;+\; 20 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^4}{1 \;-\; 2\left(\frac{3}{4}\right)^4} \quad \blacksquare$$

12 Other Appendices

12.1 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 $\quad B =$ 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		

12.2 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 12.1. **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 $A + B$. Both $A^{\complement}$ and, occasionally, $\complement A$ denote the complement $\Omega \setminus A$ of A.

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] George Casella and Roger L. Berger. Statistical Inference. Cengage, 2nd edition, 2001.

[2] Robert Hogg, Joseph McKean, and Allen Craig. Introduction to Mathematical Statistics. Pearson, 8th edition.

[3] Sheldon M. Ross. A First Course in Probability. Macmillan, New York, 3rd edition, 198.

[4] Georgi Evgen'evitch Shilov. Elementary Real and Complex Analysis. Dover, Mineola, 1st edition, 1996.

[5] D. Wackerly, W. Mendenhall, and R.L. Scheaffer. Mathematical Statistics with Applications. Thomson Brooks/Cole, 7th edition, 2008.

[6] Richard E. Williamson and Hale F. Trotter. Multivariable Mathematics. Prentice Hall, 3rd edition, 1995.

List of Symbols

Index