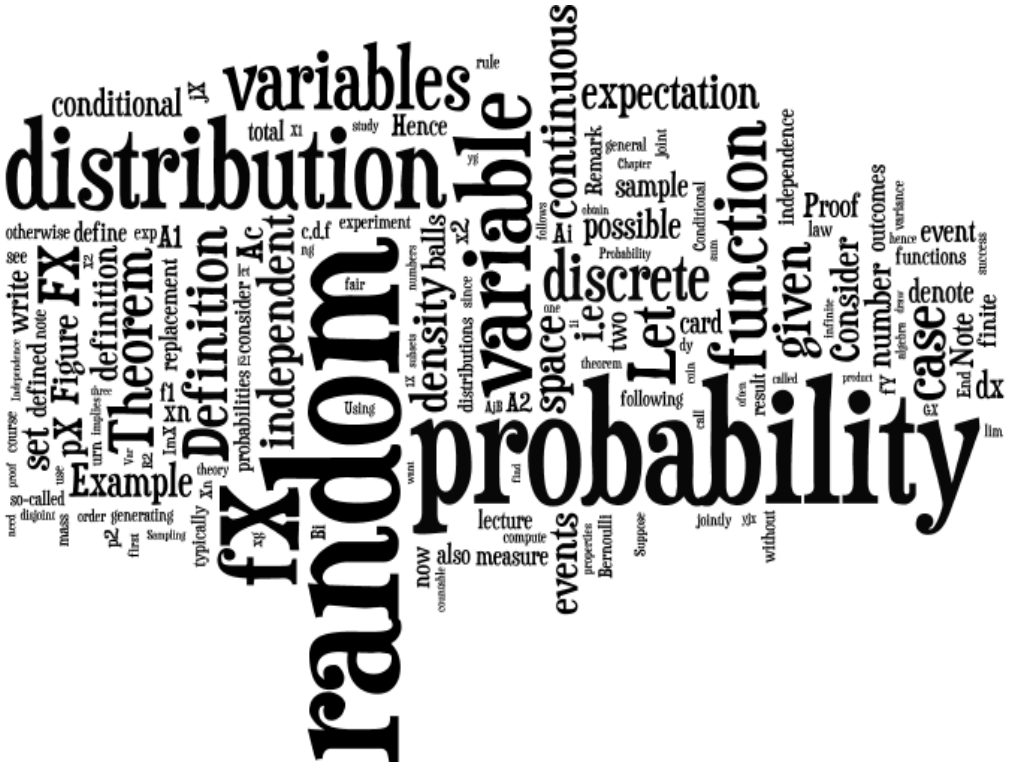
# MATH40005: Probability and Statistics

Almut E. D. Veraart
Room 551, Huxley Building
Department of Mathematics
Imperial College London
180 Queen's Gate, London, SW7 2AZ
E-Mail: a.veraart@imperial.ac.uk

Autumn 2022




(Last Update: Monday 3rd October, 2022)

# Contents

# Chapter 1

# Introduction

## 1.1 Why do we study probability?

Probability

- is a beautiful branch of mathematics with a long history going back to the early works by Cardano (16th century), Fermat and Pascal (17th century), Laplace (19th century). Modern (axiomatic) probability theory, however, is a much younger discipline which goes back to the influential work by Kolmogorov published in 1933,
- is a very dynamic discipline with a strong interplay between theory and applications,
- is ubiquitous in every day life and in most sciences,
- is the foundation for statistics,
- enables us to interpret and quantify uncertainty.

## 1.2 Complementary reading

- These lecture notes are self contained. They are mainly based on the textbooks Grimmett & Welsh (1986), Blitzstein & Hwang (2019) and Anderson et al. (2018).

  You can get the first and third book from our library and the second book is available on-line at https://projects.iq.harvard.edu/stat110/about, where you can also find additional exercises and solutions.
- Complementary reading material can be found in the following textbooks: Ross (2014).

## 1.3 Course overview (Autumn term)

1. Interpretations of probability; limiting frequency; classical (symmetry between equally likely outcomes) ; subjective (degree of personal belief)
2. Counting: multiplication principle; binomial coefficients; the inclusion-exclusion principle; stars and bars arguments
3. Formal probability: probability axioms; conditional probability; Bayes' theorem; independence
4. Random variables: mass and density functions; common discrete and continuous distributions, transformations of random variables, expectation and variance; probability and moment generating functions

5. Multivariate random variables: Joint mass and density functions; independence; covariance
6. Conditional distribution: Conditional probability mass function, conditional density, conditional expectation, law of total expectation

# Chapter 2

# Sample spaces and interpretations of probability

*The material of this chapter is based on Blitzstein & Hwang (2019), p.1-8, Anderson et al. (2018), p.1-5, Proschan & Shaw (2016), p.9-10.*

## 2.1 Notation

Throughout the lecture notes we denote the natural numbers by $\mathbb{N} = \{1, 2, \dots\}$ and we define $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Moreover, we denote the integers by $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, the real numbers by $\mathbb{R}$. For real numbers $a < b$ we write $[a, b]$ for closed intervals and $(a, b)$ for open intervals.

## 2.2 The sample space $\Omega$

Probability theory is based on *set theory* which was introduced in the course *Introduction to University Mathematics*. We will now explain how set theory enters in probability theory and review some of the key concepts briefly.

**Definition 2.2.1** (Sample space)**.** *The* sample space $\Omega$ *is defined as the set of all possible outcomes of an experiment. The elements of* $\Omega$ *are typically denoted by* $\omega$ *and called* sample points.

**Example 2.2.2.** *We start with the classical example of flipping a (fair) coin. We write* $H$ *for heads and* $T$ *for tails. The sample space is given by*

$$\Omega = \{H, T\}.$$

Figure 2.1: Flipping a fair coin.

**Example 2.2.3.** *Consider the experiment where we roll a standard six-sided (fair) die. The sample space associated with this experiment is given by*

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

### 2.2.1 Notation from set theory and Venn diagrams

Let us recap some concepts from set theory which you studied in MATH40001.

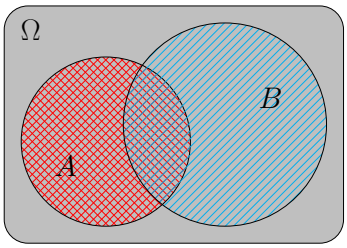
- Subsets of $\Omega$ are collections of elements of $\Omega$ and called *events*. Notation: $A$ is a subset of $\Omega$ can be written as $A \subseteq \Omega$ meaning that every element of $A$ is also an element of $\Omega$.
- We write $\omega \in A$ if the element $\omega$ is a member of $A$ and $\omega \notin A$ if the element $\omega$ is not a member of $A$.
- We denote the empty set by $\emptyset$. Note that the empty set contains no points, i.e. $\omega \notin \emptyset$ for all $\omega \in \Omega$.
- Every subset $A$ of the sample space $\Omega$ satisfies $\emptyset \subseteq A \subseteq \Omega$.

**Example 2.2.4.** *Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. For instance, we can say that $1 \in \Omega$ and $\{1\} \subseteq \Omega$.*

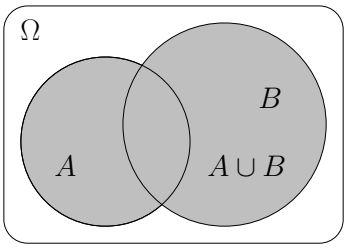
Suppose that $A, B \subseteq \Omega$ are events, then

- the union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that at least one of $A$ and $B$ occurs (this is the *inclusive* "or"),
- the intersection $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the event that both $A$ and $B$ occur,
- the complement $A^c = \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$ is the event that occurs if and only if $A$ does not occur.

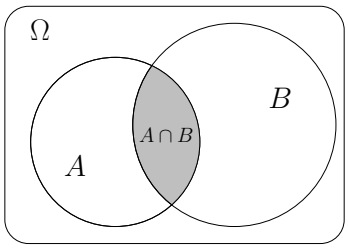
We typically use so-called *Venn diagrams* to illustrate concepts from set theory such as the union, intersection and complement of sets introduced above. Consider a sample space $\Omega$ with subsets $A, B \subseteq \Omega$.



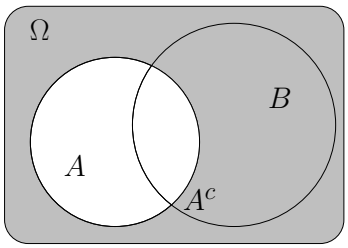

(a) $A, B \subseteq \Omega$




(b) $A \cup B$




(c) $A \cap B$




(d) $A^c$

Figure 2.2: We consider a sample space $\Omega$ with subsets $A, B \subseteq \Omega$ which are depicted in Figure 2.2a. The grey area in Figure 2.2b depicts the union $A \cup B$. The grey area in Figure 2.2c depicts the intersection $A \cup B$, and the grey area in Figure 2.2d depicts the complement $A^c$.

In MATH40001 you studied De Morgan's laws and distributivity for propositions. These results imply the following very useful identities for sets.

Let $\mathcal{I}$ denote a general index set, e.g. $\mathcal{I} = \{1, 2\}$ or $\mathcal{I} = \mathbb{N}$ or $\mathcal{I} = [0, \infty)$. Suppose that $A_i \subseteq \Omega$ for all $i \in \mathcal{I}$ and $B \subseteq \Omega$. Then the following identities hold.

- De Morgan's laws:

$$\left(\bigcap_{i\in\mathcal{I}} A_i\right)^c = \bigcup_{i\in\mathcal{I}} A_i^c, \qquad \text{and} \qquad \left(\bigcup_{i\in\mathcal{I}} A_i\right)^c = \bigcap_{i\in\mathcal{I}} A_i^c,$$

- Distributivity:

$$B \cap \left(\bigcup_{i\in\mathcal{I}} A_i\right) = \bigcup_{i\in\mathcal{I}} (B \cap A_i), \qquad \text{and} \qquad B \cup \left(\bigcap_{i\in\mathcal{I}} A_i\right) = \bigcap_{i\in\mathcal{I}} (B \cup A_i).$$

Let us prove the first De Morgan's law as an exercise. First we consider the easier case when $\mathcal{I} = \{1, 2\}$.

**Exercise 2.2.5.** *Using the notation above, show that* $(A_1 \cap A_2)^c = A_1^c \cup A_2^c$.

*Proof.* Here we want to prove an identity between two sets. As a general strategy, recall that we can prove the equality by first showing that the set on the left hand side is a subset of the set on the right hand side and then that the set on the right hand side is a subset of the set on the left hand side.

**Proof of $(A_1 \cap A_2)^c \subseteq A_1^c \cup A_2^c$:** Let $a \in (A_1 \cap A_2)^c$. Then $a$ is not an element of $(A_1 \cap A_2)$. That implies that there is at least one $i^* \in \mathcal{I} = \{1, 2\}$ such that $a \notin A_{i^*}$. This implies that there is at least one $i^* \in \mathcal{I} = \{1, 2\}$ such that $a \in A_{i^*}^c$. Since $A_{i^*}^c \subseteq A_1^c \cup A_2^c$, we deduce that $a \in A_1^c \cup A_2^c$.

**Proof of $(A_1 \cap A_2)^c \supseteq A_1^c \cup A_2^c$:** Let $a \in A_1^c \cup A_2^c$. Then there is at least one $i^* \in \mathcal{I}$ such that $a \in A_{i^*}^c$. I.e. there is at least one $i^* \in \mathcal{I}$ such that $a \notin A_{i^*}$. Since $A_1 \cap A_2 \subseteq A_{i^*}$, we deduce that $a \notin A_1 \cap A_2$. (Suppose that $a \in A_1 \cap A_2$. Then $a \in A_1$ and $a \in A_2$, which is a contradiction to the statement that there is at least one $i^* \in \mathcal{I}$ such that $a \notin A_{i^*}$.) Since the proposition $a \notin A_1 \cap A_2$ is equivalent to $a \in (A_1 \cap A_2)^c$, we are done with the proof.

□

We can use the same arguments to prove the general case:

**Exercise 2.2.6.** *Using the notation above, show that* $\left(\bigcap_{i\in\mathcal{I}} A_i\right)^c = \bigcup_{i\in\mathcal{I}} A_i^c$.

*Proof.* **Proof of $\left(\bigcap_{i\in\mathcal{I}} A_i\right)^c \subseteq \bigcup_{i\in\mathcal{I}} A_i^c$:** Let $a \in \left(\bigcap_{i\in\mathcal{I}} A_i\right)^c$. Then $a$ is not an element of $\left(\bigcap_{i\in\mathcal{I}} A_i\right)$. That implies that there is at least one $i^* \in \mathcal{I}$ such that $a \notin A_{i^*}$. This implies that there is at least one $i^* \in \mathcal{I}$ such that $a \in A_{i^*}^c$. Since $A_{i^*}^c \subseteq \bigcup_{i\in\mathcal{I}} A_i^c$, we deduce that $a \in \bigcup_{i\in\mathcal{I}} A_i^c$.

**Proof of $\left(\bigcap_{i\in\mathcal{I}} A_i\right)^c \supseteq \bigcup_{i\in\mathcal{I}} A_i^c$:** Let $a \in \bigcup_{i\in\mathcal{I}} A_i^c$. Then there is at least one $i^* \in \mathcal{I}$ such that $a \in A_{i^*}^c$. I.e. there is at least one $i^* \in \mathcal{I}$ such that $a \notin A_{i^*}$. Since $\bigcap_{i\in\mathcal{I}} A_i \subseteq A_{i^*}$, we deduce that $a \notin \bigcap_{i\in\mathcal{I}} A_i$ which is equivalent to $a \in \left(\bigcap_{i\in\mathcal{I}} A_i\right)^c$.

□

**End of lecture 1.** .....

### 2.2.2 Cardinality

**Definition 2.2.7** (Cardinality)**.** *For any set $A$, we define the* cardinality *of $A$ as the number of elements in $A$. We typically write $\text{card}(A)$ or simply $|A|$ (the latter should not be confused with the absolute value!).*

**Example 2.2.8.** *Let us compute the cardinality of the sample spaces $\Omega$ considered in the two examples above.*

- *Flipping of a (fair) coin. Here we have $\Omega = \{H, T\}$, then*

$$\text{card}(A) = 2.$$

- *Rolling a (fair) die. Here we have $\Omega = \{1, 2, 3, 4, 5, 6\}$, then*

$$\text{card}(A) = 6.$$

**Definition 2.2.9.** *Two sets have the same cardinality if there is a bijection between the two sets.*

Bijective functions were introduced in MATH40001. Let us recall their definition again.

**Definition 2.2.10** (Injective, surjective and bijective functions)**.**

- *A function $f : A \mapsto B$ is called* injective *if $\forall a_1, a_2 \in A, f(a_1) = f(a_2) \Rightarrow a_1 = a_2$.*
- *A function $f : A \mapsto B$ is called* surjective *if $\forall b \in B, \exists a \in A$ such that $f(a) = b$.*
- *A function $f : A \mapsto B$ is called* bijective *if it is both injective and surjective.*

**Definition 2.2.11** (Finite, countably infinite and uncountably infinite sets)**.**

- *A set $A$ is said to be* finite*, if it has a finite number of elements.*
- *A set $A$ is called* countably infinite *if there is a bijection between the elements of $A$ and the natural numbers $\mathbb{N} = \{1, 2, \dots\}$.*
- *A set $A$ is called* countable *if it is either finite or countably infinite.*
- *If the set $A$ is neither finite nor countably infinite, we call it* uncountable *or* uncountably infinite.

**Example 2.2.12.** *The sample space of an experiment can be*

- *finite (e.g.* $\Omega = \{1, \dots, 6\}$*),*
- *countably infinite (e.g.* $\Omega = \mathbb{N} = \{1, 2, \dots\}$*), or*
- *uncountably infinite (e.g.* $\Omega = [0, 1]$*).*

**Exercise 2.2.13.** *Show that* $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ *and* $\mathbb{N} \cup \{0\} = \{0, 1, \dots\}$ *have the same cardinality.*

*Proof.* We use the convention that 0 is even. We define a function $f : \mathbb{N} \cup \{0\} \to \mathbb{Z}$ such that

$$f(x) = \begin{cases} \frac{x}{2}, & \text{if } x \text{ is even,} \\ -\frac{(x+1)}{2}, & \text{if } x \text{ is odd.} \end{cases}$$

We need to show that this function is bijective.

First, we show that it is injective: For all $x, y \in \mathbb{N} \cup \{0\}$ with $f(x) = f(y)$ we have that $f(x)$ and $f(y)$ have the same sign. So either

$$f(x) = \frac{x}{2} = \frac{y}{2} = f(y) \Rightarrow x = y,$$

or

$$f(x) = -\frac{x+1}{2} = -\frac{y+1}{2} = f(y) \Rightarrow x = y.$$

Next, we show that $f$ is surjective. For all $y \in \mathbb{Z}, y < 0$, choose $x = -2y - 1$ (which is odd), then $f(x) = -\frac{-2y-1+1}{2} = y$. For all $y \in \mathbb{Z}, y \geq 0$, choose $x = 2y$ (which is even), then $f(x) = \frac{2y}{2} = y$. □

## 2.3 Interpretations of probability

Let us briefly discuss the three main interpretations of probability, for an extended survey please see Hájek (2012).

### 2.3.1 Naive definition of probability - classical interpretation

Consider the case when the sample space $\Omega$ is finite, i.e. $\text{card}(\Omega)$ and suppose you want to assign a probability to the event $A \subseteq \Omega$. A naive definition of a probability is obtained when we count the elements in $A$ and divide by the total number of elements in $\Omega$:

**Definition 2.3.1** (Naive definition of probability)**.** *Suppose that the sample space* $\Omega$ *is finite, i.e.* $\text{card}(\Omega) < \infty$ *and consider an event* $A \subseteq \Omega$*. Then the* naive probability *of* $A$ *is defined as*

$$\text{P}_{\text{Naive}}(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}.$$

In addition to the assumption that the sample space is finite, the naive definition of a probability also assumes that each possible outcome has the same weight. When is such a definition applicable? In *symmetric settings* when all outcomes are equally likely (for instance when we toss a fair coin or roll a fair die), or in settings where the outcomes are equally likely due to the *design* of a study (for instance when I randomly select 10 students out of the entire year group assuming that the selection mechanism is such that all subsets of 10 students are equally likely).

**Example 2.3.2.** *Consider the example of rolling a six-sided fair die. What is the (naive) probability that I roll either a 1 or a 2? We have* $\Omega = \{1, \ldots, 6\}$ *and* $A = \{1, 2\}$. *Hence*

$$\mathrm{P}_{\mathrm{Naive}}(A) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)} = \frac{2}{6} = \frac{1}{3}.$$

*Let us consider the complement:* $A^c = \{3, 4, 5, 6\}$. *The probability of the complement can be computed as*

$$\mathrm{P}_{\mathrm{Naive}}(A^c) = \frac{\mathrm{card}(A^c)}{\mathrm{card}(\Omega)} = \frac{\mathrm{card}(\Omega) - \mathrm{card}(A)}{\mathrm{card}(\Omega)} = 1 - \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)} = 1 - \mathrm{P}_{\mathrm{Naive}}(A) = \frac{2}{3}.$$

Note that for $A \subseteq \Omega$ we always have that $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$, not just in the case of the naive probability.

The classical interpretation applies when we have outcomes that are equally likely. In our naive definition above, we have only covered the case when $\mathrm{card}(\Omega) < \infty$. If $\Omega$ is uncountably infinite, but of finite area, e.g. choose a disk of radius 1: $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ and the event $A$ is some subset of $\Omega$, then we could assume that the probability of the event $A$ should be uniform on $\Omega$, i.e.

$$\mathrm{P}(A) = \frac{\text{area of } A}{\text{area of } \Omega}.$$

For instance, for $A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 0.5^2\}$, we have

$$\mathrm{P}(A) = \frac{\text{area of } A}{\text{area of } \Omega} = \frac{0.5^2\pi}{\pi} = 0.25.$$

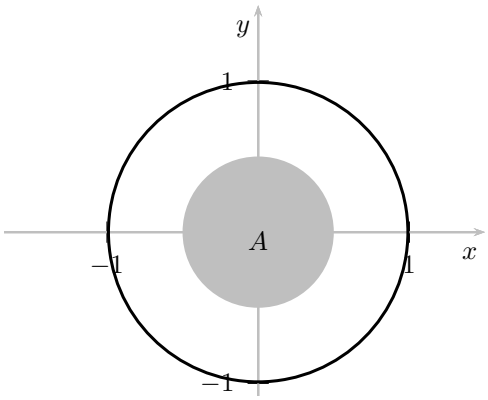



Figure 2.3: Illustration of the classical interpretation of probability in the case when $\Omega$ is uncountably infinite.

**Remark 2.3.3.** *In order for the classical/naive definition to work, we need that the number of elements in* $\Omega$ *is either finite, or, if* $\Omega$ *is uncountably infinite, then we require that the area of* $\Omega$ *is finite. In either scenario we can then define a uniform distribution, which we will study in more detail later in the course. Note that there is no uniform distribution on* $\mathbb{N}$ *or on* $\mathbb{R}$.

### 2.3.2 Limiting frequency

Consider $n_{\mathrm{total}}$ replications of an experiment and let $n_A$ denote the number of times event $A$ occurs (out of $n_{\mathrm{total}}$). Then we could interpret the probability of event $A$ occuring as

$$\mathrm{P}(A) = \lim_{n_{\mathrm{total}} \to \infty} \frac{n_A}{n_{\mathrm{total}}}.$$

The problem with this interpretation is that $n_{\mathrm{total}} \to \infty$ may be difficult to conceive, and any finite version may not be representative.

**Example 2.3.4.** *Consider a fair coin toss and let $A = \{H\}$ denote the event that Heads appears. Figure[1] 2.4 illustrates a possible outcome of the experiment.*

Figure 2.4: One possible outcome when tossing a fair coin repeatedly.

*Let us compute the relative frequency of $A$ and report and plot them in Table 2.1 and Figure 2.5, respectively.*

| $n_{\text{total}}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | $\cdots$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\frac{n_A}{n_{\text{total}}}$ | 0/1 | 0/2 | 1/3 | 2/4 | 2/5 | 3/6 | 4/7 | 5/8 | 6/9 | 6/10 | $\cdots$ |

Table 2.1: Relative frequencies of heads when repeatedly tossing a fair coin.

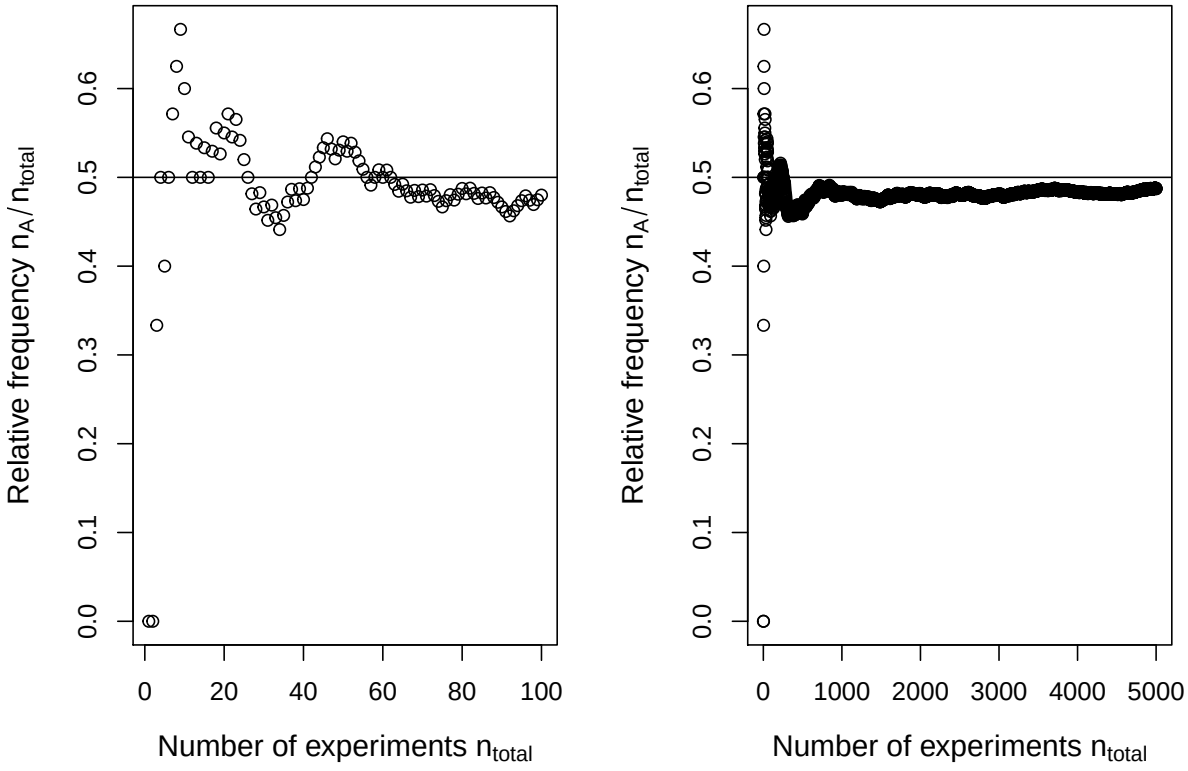



Figure 2.5: Relative frequencies of heads when repeatedly tossing a fair coin.

### 2.3.3 Subjective

For an event $A$, we can assign the probability $\mathrm{P}(A)$ according to our personal "degree of belief". This could be done according to historical information or local knowledge. This probability will not need to be

[1] The pictures of the one pound coin are attributed to Sir Magnus Fluffbrains [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)] https://commons.wikimedia.org/wiki/File:Pound_coin_front.png and https://commons.wikimedia.org/wiki/File:Pound_coin_back.png.

the same for each individual. The subjective approach may be difficult to implement in practice, but is a valid and universal interpretation of probability.

**Remark 2.3.5.** *It is important to remember that all three interpretations of probability depend on assumptions about experimental conditions.*

# Chapter 3

# Counting

*The material of this chapter is based on Blitzstein & Hwang (2019), p.8-28, Anderson et al. (2018), p.4-11.*

In order to compute (naive) probabilities we need to be able to count events in possibly large (but finite) sample spaces. The area of mathematics which deals with counting is called *Combinatorics*. We will study some key ideas from combinatorics and show their interplay with probability theory.

## 3.1 The multiplication principle

**Theorem 3.1.1** (The multiplication principle)**.** *Consider two experiments: Experiment A has $a$ possible outcomes, and Experiment B has $b$ possible outcomes. Then the compound experiment of performing Experiment A and B (in any order) has $ab$ possible outcomes.*

*Proof.* Without loss of generality we assume that we conduct Experiment A first. We draw a tree diagram consisting of $a$ branches with one branch for each possible outcome of Experiment A. For each of theses branches we then generate $b$ branches for each possible outcome of Experiment B. We can then directly read off that there are $\underbrace{b + \cdots + b}_{a} = ab$ possibilities. □

We illustrate the proof of the multiplication principle in Figure 3.1 in the case when the outcomes of Experiment A are labelled as $A_1, A_2$ (for $a = 2$) and the outcomes of Experiment B are labelled as $B_1, B_2, B_3$ (for $b = 3$).

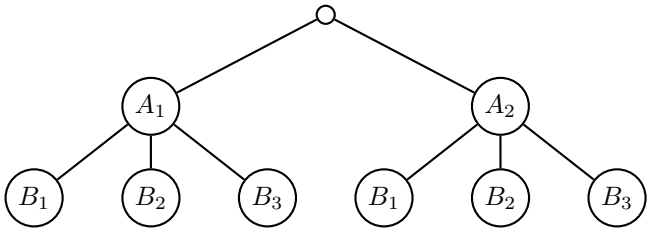



Figure 3.1: Illustration of the proof of the multiplication principle in the case when the outcomes of Experiment A are labelled as $A_1, A_2$ (for $a = 2$) and the outcomes of Experiment B are labelled as $B_1, B_2, B_3$ (for $b = 3$).

**Exercise 3.1.2.** *We flip a fair coin three times. We write "0" for heads and "1" for tails. Find the sample space $\Omega$ and $\mathrm{P}_{\mathrm{Naive}}(\{\omega\})$ for all $\omega \in \Omega$. Also, compute the probability that the first and the third flip are tails.*

*Proof.* Here the sample space is given by

$$\Omega = \{(0,0,0),(0,0,1),\ldots,(1,1,0),(1,1,1)\},$$

which are all the ordered triplets of zeros and ones. By the multiplication principle we have that $\mathrm{card}(\Omega) = 2^3 = 8$, hence $\mathrm{P}_{\mathrm{Naive}}(\{\omega\}) = 1/8$ for all $\omega \in \Omega$.

Let us now consider the event $B :=$ the first and the third flip are tails $= \{(1,0,1),(1,1,1)\}$. Then $\mathrm{P}_{\mathrm{Naive}}(B) = 2/8 = 1/4$. □

**Remark 3.1.3.** *If we deal with repetitions of experiments (coin toss, rolling a die), the corresponding sample spaces are given by* Cartesian product spaces.

*For sets $A_1,\ldots,A_n$, we define the Cartesian product as*

$$A_1 \times \cdots \times A_n = \{(x_1,\ldots,x_n) : x_i \in A_i \text{ for } i = 1,\ldots,n\}.$$

So in our example above, we can write $\Omega = \{0,1\} \times \{0,1\} \times \{0,1\} = \{0,1\}^3$.

**End of lecture 2.** .....

## 3.2 Power sets

**Exercise 3.2.1.** *Consider a set $\Omega$ with $\mathrm{card}(\Omega) = n \in \mathbb{N}$ elements. Use the multiplication principle to show that there are $2^n$ possible subsets of $\Omega$ if you include the empty set $\emptyset$ and $\Omega$.*

*Proof.* For each element in $\Omega$, you can choose whether or not to include it in the subset, so you have two options each. Hence you have $\underbrace{2\cdots 2}_{n} = 2^n$ possible outcomes. □

**Definition 3.2.2** (Power set)**.** *A* power set *of a set $A$, denoted as $\mathcal{P}(A)$ is defined as the set of all possible subsets of $A$ including $\emptyset$ and $A$.*

We have already proven the following result:

**Theorem 3.2.3** (Cardinality of the power set)**.** *Consider a sample space $\Omega$ with $\mathrm{card}(\Omega) < \infty$. Then $\mathrm{card}(\mathcal{P}(\Omega)) = 2^{\mathrm{card}(\Omega)}$.*

**Exercise 3.2.4.** *Let $\Omega = \{A,B,C\}$. Find $\mathrm{card}(\mathcal{P}(\Omega))$ and $\mathcal{P}(\Omega)$.*

*Proof.* From the theorem above, we know that the corresponding power set $\mathcal{P}(\Omega)$ consists of $\mathrm{card}(\mathcal{P}(\Omega)) = 2^3 = 8$ elements. Also, we have

$$\mathcal{P}(\Omega) = \{\emptyset,\{A\},\{B\},\{C\},\{A,B\},\{A,C\},\{B,C\},\Omega\}.$$

□

## 3.3 Sampling with and without replacement

We can use the multiplication principle to derive the number of outcomes when sampling with or without replacement. In the following, we will state the results in the context when we have an urn with $n \in \mathbb{N}$ balls which are labelled $\{1,2,\ldots,n\}$ and we draw $k \in \mathbb{N}$ balls from the urn.

Figure 3.2: Consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \ldots, n\}$. How many possible ways are there to draw $k \in \mathbb{N}$ balls with or without replacement?

### 3.3.1 Sampling with replacement – ordered

Consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \ldots, n\}$. Let $k \in \mathbb{N}$. Suppose that you take a ball out of the urn and write down its number. Then you put it back into the urn (i.e. you replace it). You do this $k$ times in total and write down the labels in the order in which they appear. The sample space $\Omega$ of this experiment can be expressed in the following way: We denote by $S = \{1, \ldots, n\}$ the labels of the balls in the urn. A possible outcome of the experiment can be written as $\omega = (s_1, \ldots, s_k)$, where $s_i$ denotes the number of the $i$th ball for $i \in \{1, \ldots, n\}$. Hence

$$\Omega = \underbrace{S \times \cdots \times S}_{k \text{ times}} = S^k = \{(s_1, \ldots, s_k) : s_i \in S \text{ for } i = 1, \ldots, k\}.$$

**Theorem 3.3.1** (Sampling with replacement)**.** *In the case of sampling $k$ balls with replacement from an urn containing $n$ balls as described above, there are $\mathrm{card}(\Omega) = n^k$ possible outcomes when the order of the objects matters.*

*Proof.* The result is a direct consequence of the multiplication principle: Each time we draw a ball, there are $n$ possible outcomes. We carry out this experiment $k$ times, so there are $n^k$ ways of obtaining a sample consisting of $k$ balls. □

### 3.3.2 Sampling without replacement – ordered

Again we consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \ldots, n\}$. Let $k \in \mathbb{N}$. Suppose that you take a ball out of the urn and write down its number. Then you remove the ball and do not put it back into the urn (i.e. you do not replace it). You do this $k$ times in total. Since you are removing balls from the urn permanently, $k$ cannot be larger than $n$.

The sample space $\Omega$ of this experiment can be expressed in the following way: We denote by $S = \{1, \ldots, n\}$ the labels of the balls in the urn. Then

$$\Omega = \{(s_1, \ldots, s_k) : s_i \in S \text{ for } i = 1, \ldots, k, \text{ and } s_i \neq s_j \text{ if } i \neq j\}.$$

**Theorem 3.3.2** (Sampling without replacement)**.** *In the case of sampling $k$ balls without replacement from an urn containing $n$ balls as described above, there are $\mathrm{card}(\Omega) = n(n-1)\cdots(n-(k-1)) = (n)_k$ possible outcomes when the order of the objects matters.*

Note that we use the convention that $(n)_1 = n$.

*Proof.* Also this result is a direct consequence of the multiplication principle: The first time, we draw a ball, there are $n$ possible outcomes, the second time, there are $n-1$ possible outcomes and so on, and when we draw the $k$th ball, there are $n-(k-1)$ possible labels left for the $k$th ball. Multiplying the number of possible outcomes for each sub-experiment together leads to the stated result. □

**Definition 3.3.3** (Factorial)**.** *Let $n \in \mathbb{N}$. The* factorial *of $n$, denoted by $n!$, is defined as the product of all natural numbers less than or equal to $n$, i.e. $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 = \prod_{i=1}^{n} i$. We define $0! = 1$.*

**Definition 3.3.4** (Descending factorial)**.** *For $k, n \in \mathbb{N}$ with $k \leq n$, we define the* descending factorial*, denoted by $(n)_k$ as $(n)_k = n(n-1)\cdots(n-k+1) = \prod_{i=0}^{k-1}(n-i) = \prod_{j=n-k+1}^{n} j$ with the convention that $(n)_1 = n$. We note that the descending factorial can be expressed as $(n)_k = \frac{n!}{(n-k)!}$.*

The factorial arises naturally in the context of so-called *permutations*. Consider the set of numbers $\{1, 2, \ldots, n\}$. A permutation brings these numbers into a certain order. As a consequence of Theorem 3.3.2 with $k = n$, we deduce that the numbers in the set $\{1, 2, \ldots, n\}$ can be arranged in exactly $n!$ possible ways.

**Example 3.3.5.** *Consider the set $\{1, 2, 3\}$. How many permutations (i.e. possible orderings) are there? We can write $(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$, so we have $3! = 6$ possible permutations.*

### 3.3.3 The birthday problem

Let us now study the famous birthday problem:

**Example 3.3.6.** *Assume there are $k \in \mathbb{N}$ people in a room and assume that each person's birthday is equally likely to be any of the 365 days of the year (with the 29th February excluded). What is the probability that at least two people in the room have the same birthday?*

- *Due to our assumption, the naive probability definition is applicable here. First we count how many possible ways there are to assign birthdays to the $k$ people in the room.*
- *This problem can be viewed as sampling with replacement, so we have $365^k$ possible birthday combinations.*
- *Next, we need to count how many scenarios there are such that at least two people have the same birthday. It appears that this is rather challenging...*
- *What is easier to compute is the complement, i.e. the number of scenarios such that no two people share the same birthday. This number can be computed using sampling without replacement, which leads to $(365)_k$ possible outcomes.*

*Combining the results, we get*

$$\begin{aligned}
&\mathrm{P}_{\mathrm{Naive}}(\text{At least two people in the room have the same birthday}) \\
&= 1 - \mathrm{P}_{\mathrm{Naive}}(\text{All people in the room have distinct birthdays}) \\
&= 1 - \frac{(365)_k}{365^k} = 1 - \frac{365}{365}\frac{364}{365}\cdots\frac{365-(k-1)}{365} =: f(k).
\end{aligned}$$

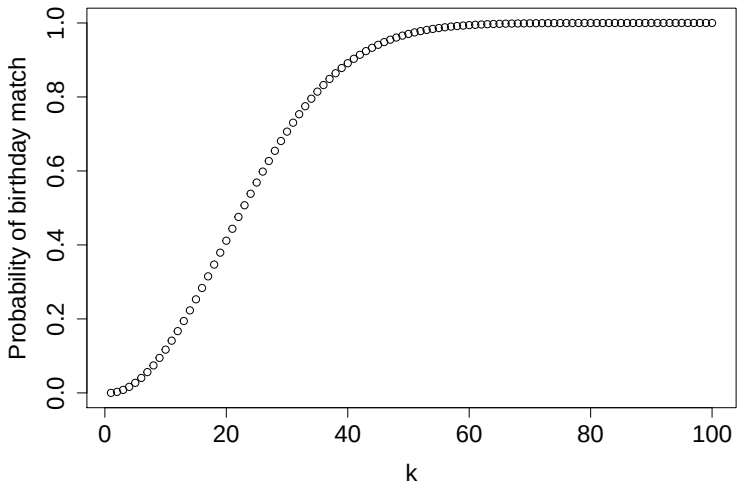



Figure 3.3: We compute the probability $f(k)$ that, out of $k$ people in a room, at least two share the same birthday.

*We plot the probabilities $f(k)$ for $k = 1, \dots, 100$ in Figure 3.3. Note that $f(22) \approx 0.476$ and $f(23) \approx 0.507$, so you need to have at least 23 people in the room to have a probability of at least $50\%$ such that at least two people share the same birthday.*

**Exercise 3.3.7.** *A college has 10 non-overlapping time slots for its courses and assigns courses to time slots randomly and independently. A student randomly chooses three of the courses to enroll in. What is the probability that there is a conflict in the student's schedule?*

*Proof.* Using the multiplication principle and the naive probability, we can compute the probability of no schedule conflict as $\frac{10 \cdot 9 \cdot 8}{10^3} (= 0.72)$. So the probability that there is at least one schedule conflict is given by $1 - \frac{10 \cdot 9 \cdot 8}{10^3} (= 0.28)$. □

**Exercise 3.3.8.** *A fair die is rolled 6 times. What is the probability that some value is repeated?*

*Proof.* There are $6^6$ possible outcomes when rolling a die 6 times. There are $6!$ configurations where each number appears exactly once. Hence $\mathrm{P}(\text{no value repeated}) = \frac{6!}{6^6}$ and $\mathrm{P}(\text{at least one value is repeated}) = 1 - \frac{6!}{6^6} (\approx 0.9845)$. □

**End of lecture 3.** .....

### 3.3.4 Sampling without replacement – unordered

Consider the case of an urn with $n \in \mathbb{N}$ balls, where we take out $k \in \mathbb{N}$ ($k \leq n$) balls and write down their labels, which are distinct numbers in $\{1, \dots, n\}$. We do not care about the order in which the balls are collected. (For instance, think of drawing the winning numbers in the lottery!) Note that you could view this experiment as drawing $k$ balls at once rather than one at a time. Hence the outcome of the experiment is a subset of size $k$ from $S = \{1, \dots, n\}$. Hence we can write $\Omega = \{\omega \subseteq S : \mathrm{card}(\omega) = k\}$.

**Definition 3.3.9** (Binomial coefficient)**.** *For any $k, n \in \mathbb{N} \cup \{0\}$, the* binomial coefficient *is defined as the number of subsets of size $k$ for a set of size $n$. It is denoted by $\binom{n}{k}$ and we say "$n$ choose $k$".*

**Theorem 3.3.10** (Binomial coefficient)**.** *For any $k, n \in \mathbb{N} \cup \{0\}$, we have*

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-(k-1))}{k!} = \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}.$$

*for $k \leq n$, and $\binom{n}{k} = 0$ for $k > n$.*

*Proof.* Consider the setting of the urn with $n$ balls, where we draw $k$ balls at once as described above. Clearly, if $k > n$, then $\binom{n}{k} = 0$. Suppose now that $k \leq n$. By Theorem 3.3.2, we know that there are $(n)_k$ possible choices if we draw $k$ balls without replacement and care about the ordering. Now we need to make an adjustment for the overcounting since we do not care about the order any more. For each subset of size $k$, we have $k!$ permutations. So we conclude that when we divide $(n)_k$ by $k!$ we have adjusted for the overcounting and obtain the result. □

**Example 3.3.11.** *Recall the binomial theorem, which states that for any $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

*The theorem can be proven as follows: We expand $(x+y)^n = (x+y)\cdots(x+y)$ in $n$ factors and then pick either the $x$ or the $y$ from the first factor, multiply this to either the $x$ or the $y$ of the second factor and so on. There are $\binom{n}{k}$ ways of picking exactly $k$ $x$'s and each of these choices leads to a term of the form $x^k y^{n-k}$. Summing over all possible $k$ leads the result.*

**Exercise 3.3.12.** *A family has 6 children consisting of 3 boys and 3 girls. Assuming that all birth orders are equally likely, what is the probability that the 3 eldest children are the 3 girls?*

*Proof.* Label the girls as 1, 2, 3, and the boys as 4, 5, 6. Then the birth order is a permutation of 1, 2, 3, 4, 5, 6. So, 236514 means that child 2 was born first, then child 3 etc. The number of possible permutations is $6!$. For the three girls to be the eldest children, we need a permutation of 1,2,3, followed by a permutation of 4,5,6. Hence

$$\mathrm{P}(\text{the 3 girls are the 3 eldest children}) = \frac{3!3!}{6!} = \frac{1}{20} = 0.05.$$

**Alternative proof:** There are $\binom{6}{3}$ ways to chose where the three girls appear in the birth order (without taking ordering of the girls into account). Of these cases, there is only one where the three girls are the three eldest children. Hence

$$\mathrm{P}(\text{the 3 girls are the 3 eldest children}) = \frac{1}{\binom{6}{3}} = \frac{3!3!}{6!} = \frac{1}{20} = 0.05.$$

□

**Exercise 3.3.13.** *Consider a group of four people.*

1. *How many ways are there to choose a two-person committee?*
2. *How many ways are there to break the people into two teams of two?*

*Proof.* **Part 1:** We use two approaches to show that there are 6 possibilities:

- We could list all possibilities: Label the people as 1, 2, 3, 4. Then the possibilities are

  $$\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\},$$

  i.e. we have 6 possibilities.

- Alternatively, we can use the multiplication rule and account for overcounting:

  There are $4$ possibilities for choosing the first person on the committee and $3$ to choose the second person. Note, however, that this counts every possibility twice since picking 1 and 2 is the same as picking 2 and 1. We have overcounted by a factor of 2. So the number of possibilities is given by

$$\frac{4 \cdot 3}{2} = 6 \left( = \binom{4}{2} \right).$$

**Part 2:** We use three approaches to show that there are $3$ possibilities:

- Again, we could list all possibilities and count them:

$$\{1,2\},\{3,4\};\{1,3\},\{2,4\};\{1,4\},\{2,3\},$$

  which is rather tedious...

- Alternatively, we could just specify the first person's teammate, then the other team is determined. There are $3$ ways of doing this.

- Or, we use Part 1) to deduce that there are $6$ possibilities of choosing one team. Here we overcount by a factor of $2$ since picking $\{1,2\}$ as a team is equivalent to picking $\{3,4\}$ as a team. Hence we have $6/2 = 3$ possibilities.

□

**Exercise 3.3.14.** *How many ways are there to permute the letters in the word STATISTICS? Note that there are 10 letter in total, "S" and "T" appear three times, "I" twice and "A" and "C" once.*

*Proof.*

- Approach 1: There are $10$ positions in total, first we choose the $3$ positions for the "S"s out of $10$, then the $3$ positions for the "T"s out of the remaining $7$ positions etc. Hence we get

$$\underbrace{\binom{10}{3}}_{\text{"}S\text{"}} \underbrace{\binom{7}{3}}_{\text{"}T\text{"}} \underbrace{\binom{4}{1}}_{\text{"}A\text{"}} \underbrace{\binom{3}{2}}_{\text{"}I\text{"}} \underbrace{\binom{1}{1}}_{\text{"}C\text{"}} = 50400$$

- Approach 2: Start with $10!$ permutations of the $10$ letters and adjust for overcounting:

$$\frac{10!}{3!3!2!} = 50400.$$

□

### 3.3.5 Sampling with replacement – unordered

As before, let $k, n \in \mathbb{N}$. Let us consider the case that we have an urn with $n$ balls with labels in $\{1,\dots,n\}$, and we want to choose $k$ balls one after the other with replacement. Assuming the order of the balls does not matter, how many possible outcomes are there? The sample space for our experiment is given by $\Omega = \{\omega : \omega \text{ is a } k - \text{element multiset with elements from } \{1,\dots,n\}\}$.

**Theorem 3.3.15** (Sampling with replacement when the order does not matter)**.** *In the sampling with replacement problem described above and assuming that the order of the balls does not matter, we have* $\text{card}(\Omega) = \binom{n+k-1}{k}$ *possibilities.*

The proof of the theorem relies on the so-called *stars and bars* argument presented in Feller (1957), which is a graphical tool for counting. Note that since we sample with replacement, we can have that $k > n$.

*Proof.* Consider $n$ distinguishable boxes representing the $n$ distinct labels of the balls in the urn. We can draw these $n$ boxes using $n+1$ bars which represent the walls of the boxes, see Figure 3.4 for an illustration. Next, we have $k$ indistinguishable balls, which we now draw as stars and place them into the boxes, i.e. between the bars.

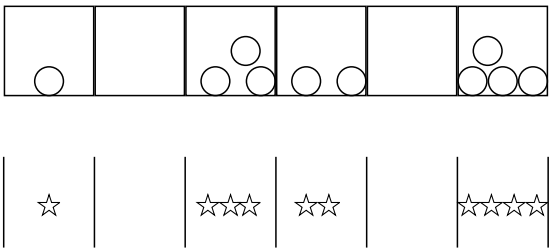


Figure 3.4: Illustration of the stars and bars method in the case when $n = 6$ and $k = 10$. Here we have $\binom{n+k-1}{k} = \binom{15}{10} = 3003$ possible outcomes.

I.e. we can view the stars as "check marks" which count how often a particular label gets selected. Between the two outer walls (i.e. the first and the last bar), which are fixed, there are $n-1$ bars and $k$ stars, i.e. $n+k-1$ symbols which can be arranged in any possible order. Out of the $n+k-1$ possible positions, we choose $k$ positions for the stars and fill the remaining positions with bars. Hence $\text{card}(\Omega) = \binom{n+k-1}{k}$. Note that we could have picked the $n-1$ bars instead and then filled in the remaining stars, which leads to the identity

$$\text{card}(\Omega) = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}.$$

□

It is important to remember that you should not use the above result in connection with the naive probability since the unordered samples are typically not equally likely.

**Exercise 3.3.16.** *[Exam question 2020] How many possibilities are there to write the number 7 as an ordered sum of 3 positive integers? [E.g. 7=1+3+3 would be one possible case and 7=3+1+3 would be another case.]*

*Proof.* We present an elementary solution here: We can write down all possible 3-tuples of numbers which sum up to 7:

- $(1,3,3)$ with $3!/2! = 3$ possible arrangements (where we adjusted for over-counting since the number 3 appears twice),
- $(1,1,5)$ with $3!/2! = 3$ possible arrangements,
- $(2,2,3)$ with $3!/2! = 3$ possible arrangements,
- $(1,2,4)$ with $3! = 6$ possible arrangements.

Hence there are $3+3+3+6 = 15$ possibilities of writing the number $7$ as a sum of $3$ positive integers. □

**Exercise 3.3.17.** *[Exam question 2020] Let $k, n \in \mathbb{N} = \{1,2,\ldots\}$. How many possibilities are there to write the number $k$ as an ordered sum of $n$ positive integers?*

*Proof.* We can use a stars and bars argument: There are $0$ possibilities if $k < n$.

Now suppose that $k \geq n$. Then we represent the number $k$ as $k$ stars which we would like to place in $n$ bins such that each bin contains at least one object (since we have the restriction that all addends are *positive* integers). We can first write the $k$ stars in one line. Then there are $k-1$ possible gaps between the stars, where a bar could be inserted to separate the bins. We need to select $n-1$ gaps out of the $k-1$ gaps, to create the $n$ bins, so in total we have $\binom{k-1}{n-1}$ possibilities. □

The above proof can be further illustrated by associating the stars with "1"s and then writing

$$k = 1 + 1 + 1 + \cdots + 1,$$

i.e. we express $k$ as a sum of $k$ "1"s. Each "+" sign is a possible location for a bar. There are $k-1$ "+"signs and we need to choose $n-1$ of them which leads to the result.

For instance, in the case of Exercise 3.3.16, one possible configuration is $*|***|**$, where we have 1 star in the first bin, followed by 3 stars in the next bin, followed by 3 stars in the last bin. This can be viewed is writing

$$7 = 1 + (1 + 1 + 1) + (1 + 1 + 1).$$

### 3.3.6 Summary table

We can summarise the preceding discussion as follows: Consider an urn with $n \in \mathbb{N}$ balls, where you draw $k \in \mathbb{N}$ balls. The number of possible outcomes is then given as follows:

| | Ordered | Unordered |
|---|---|---|
| With replacement | $n^k$ | $\binom{n+k-1}{k}$ |
| Without replacement (for $k \leq n$ only) | $(n)_k$ | $\binom{n}{k}$ |

**End of lecture 4.** .....

# Chapter 4

# Axiomatic definition of probability

*The material of this chapter is based on Blitzstein & Hwang (2019), p.21-26, Anderson et al. (2018), p.1-21, Grimmett & Welsh (1986), p.3-9.*

In this chapter we will focus on an axiomatic definition of probability and derive some of the key properties of the probability measure.

## 4.1 The event space $\mathcal{F}$

Recall Definition 2.2.1, which states that the *sample space* $\Omega$ is defined as the set of all possible outcomes of an experiment. In our previous discussion we assigned (naive) probabilities to *events* which were subsets of $\Omega$. We typically denote by $\mathcal{F}$ the *event space*, which contains the events we are allowed to consider. This is a rather vague statement, which we will need to make precise! In fact, in probability theory, we always require that the event space $\mathcal{F}$ is a so-called $\sigma$-algebra (which is the same as a $\sigma$-field).

**Definition 4.1.1** (Algebra and $\sigma$-algebra)**.** *A collection of subsets of $\Omega$ denoted by $\mathcal{F}$ is called*

1. *an* algebra *(or a* field*) on $\Omega$ if*
   (a) $\emptyset \in \mathcal{F}$,
   (b) *$\mathcal{F}$ is closed under complements, i.e. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, and*
   (c) *$\mathcal{F}$ is closed under unions of pairs of members, i.e. $A_1, A_2 \in \mathcal{F} \Rightarrow A_1 \cup A_2 \in \mathcal{F}$.*
2. *an $\sigma$-*algebra *(or a $\sigma$-*field*) on $\Omega$ if*
   (a) $\emptyset \in \mathcal{F}$,
   (b) *$\mathcal{F}$ is closed under complements, i.e. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, and*
   (c) *$\mathcal{F}$ is closed under countable union, i.e. for any countable [1] index set $\mathcal{I}$, we have $A_i \in \mathcal{F}$, for all $i \in \mathcal{I} \Rightarrow \cup_{i \in \mathcal{I}} A_i \in \mathcal{F}$.*

Let us consider two examples of algebras, which are not a $\sigma$-algebras.

**Example 4.1.2.** *Let $\Omega = \mathbb{R}$ and define $\mathcal{F}$ as the collection of finite disjoint unions of sets of the form $(a, b], (-\infty, a], (b, \infty)$ for all $a, b \in \mathbb{R}$.*
*Then $\mathcal{F}$ is an algebra:*

1. *$\emptyset = (a, b] \in \mathcal{F}$ for $b < a$,*

[1] Recall Definition 2.2.11: A countable index set could have countably infinitely many values, e.g. when $\mathcal{I} = \mathbb{N}$ or finitely many values, e.g. when $\mathcal{I} = \{1, 2\}$.

2. $\emptyset^c = \mathbb{R} = (-\infty, a] \cup (a, \infty) \in \mathcal{F}$. *Also, for all* $a < b$ *we have*

$$(a, b]^c = (-\infty, a] \cup (b, \infty) \in \mathcal{F},$$
$$(-\infty, a]^c = (a, \infty) \in \mathcal{F},$$
$$(b, \infty)^c = (-\infty, b] \in \mathcal{F}.$$

*These results can be further extended to the corresponding finite disjoint unions of sets specified above and imply that for any* $A \in \mathcal{F}$*, we also have that* $A^c \in \mathcal{F}$.

3. *The closedness under unions of pairs follows directly from the definition of* $\mathcal{F}$.

*Hence* $\mathcal{F}$ *is an algebra.*

*However, it is NOT a* $\sigma$*-algebra. To see this, note that*

$$A_i = \left(0, 1 - \frac{1}{i}\right] \in \mathcal{F},$$

*but*

$$\bigcup_{i=1}^{\infty} A_i = \left(0, 1 - \frac{1}{1}\right] \cup \left(0, 1 - \frac{1}{2}\right] \cup \left(0, 1 - \frac{1}{3}\right] \cup \cdots = (0, 1) \notin \mathcal{F}.$$

**Example 4.1.3.** *Let* $\Omega = \mathbb{R}$ *and* $\mathcal{F} = \{A \subset \Omega : A \text{ is finite or } A^c \text{ is finite}\}$. *Then* $\mathcal{F}$ *is an algebra since:*

1. $\emptyset \in \mathcal{F}$ *since it is finite,*

2. *For any* $A \in \mathcal{F}$*, we have either*

$$A \text{ is finite} \Rightarrow (A^c)^c = A \text{ is finite, hence } A^c \in \mathcal{F},$$

*or*

$$A^c \text{ is finite} \Rightarrow A^c \in \mathcal{F}.$$

3. *For any* $A_1, A_2 \in \mathcal{F}$*, we have either*

$$A_1, A_2 \text{ are both finite} \Rightarrow A_1 \cup A_2 \text{ finite } \Rightarrow A_1 \cup A_2 \in \mathcal{F},$$

*or at least one* $A_i^c$ *is finite for* $i = 1, 2$*. Without loss of generality assume that* $A_2^c$ *is finite. Then, by De Morgan's law,*

$$(A_1 \cup A_2)^c = (A_1^c \cap A_2^c) \subseteq A_2^c \text{ is finite} \Rightarrow A_1 \cup A_2 \in \mathcal{F}.$$

*Hence* $\mathcal{F}$ *is an algebra.*

*However,* $\mathcal{F}$ *is NOT a* $\sigma$*-algebra. To see this, note that*

*$A_i = \{i\} \in \mathcal{F}$ since it is finite, but $\cup_{i=1}^{\infty} A_i = \mathbb{N} \notin \mathcal{F}$ since it is not finite and $\mathbb{N}^c = \mathbb{R} \setminus \mathbb{N}$ is not finite either!*

*Note that if we work with $\Omega = \mathbb{N}$ instead in the above example, we still get that $\mathcal{F}$ is an algebra, but not a $\sigma$-algebra. Here we could take $A_i = \{2i\}$, the $\cup_{i=1}^{\infty} A_i$ are the even natural numbers, which are infinite, and $(\cup_{i=1}^{\infty} A_i)^c$ are the odd natural numbers, which are infinite, too. So $\cup_{i=1}^{\infty} A_i \notin \mathcal{F}$.*

**Remark 4.1.4.** 1. *Any algebra is closed under finite unions and finite intersections.*

2. *Any* $\sigma$*-algebra is closed under countable intersections since* $\cap_{i=1}^{\infty} A_i = (\cup_{i=1}^{\infty} A_i^c)^c$ *and each* $A_i^c \in \mathcal{F}$.

3. *Any (*$\sigma$*-)algebra contains* $\Omega = \emptyset^c$.

**Example 4.1.5.** *Consider any sample space* $\Omega$*. Then the so-called* trivial $\sigma$*-algebra is defined as* $\mathcal{F}_{trivial} = \{\emptyset, \Omega\}$ *and the* total *or* power $\sigma$*-field is defined as* $\mathcal{F} = \mathcal{P}(\Omega) = \{\text{all subsets of } \Omega\}$.

**Example 4.1.6.** *Consider a sample space $\Omega$ with $A \subseteq \Omega$. Then $\{\emptyset, \Omega, A, A^c\}$ is a $\sigma$-algebra (in fact the smallest $\sigma$-algebra including $A$).*

Throughout the course, we shall assume that $\mathcal{F}$ is a $\sigma$-algebra. This will allow us to consider countable infinite rather than finite unions. Note that this is clearly more restrictive than assuming that $\mathcal{F}$ is an algebra.

So why do we care about algebras at all? In probability we typically define a probability measure first on an algebra and extend it to a $\sigma$-algebra. The details behind this construction are beyond the scope of this introductory course, but you can learn more about this in our measure theory course.

## 4.2 Definition of probabilities and basic properties

### 4.2.1 Probability measure and probability space

**Definition 4.2.1** (Probability measure)**.** *A mapping $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ is called a* probability measure *on $(\Omega, \mathcal{F})$ if it satisfies three conditions:*

*(i)* $\mathrm{P}(A) \geq 0$ *for all events $A \in \mathcal{F}$,*

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

*(iii) For any countable*[2] *sequence of disjoint events $(A_i)_{i\in\mathcal{I}}$ with $A_i \in \mathcal{F}$, for all $i \in \mathcal{I}$, we have*

$$\mathrm{P}\left(\bigcup_{i\in\mathcal{I}} A_i\right) = \sum_{i\in\mathcal{I}} \mathrm{P}(A_i).$$

*[Note that by "disjoint events" we mean that $A_i \cap A_j = \emptyset$ for all $i \neq j$.]*

**Definition 4.2.2** (Probability space)**.** *We define a probability space as the triplet $(\Omega, \mathcal{F}, \mathrm{P})$, where $\Omega$ is a set (the sample space), $\mathcal{F}$ is a $\sigma$-algebra (the event space) consisting of subsets of $\Omega$ and $\mathrm{P}$ is a probability measure on $(\Omega, \mathcal{F})$.*

### 4.2.2 Basic properties of the probability measure

**Theorem 4.2.3.** *Consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then, for any events $A, B \in \mathcal{F}$, we have*

1. $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
2. *If $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$.*
3. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

*Proof.* 1. Show that $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$: Since $A$ and $A^c$ are disjoint and $\Omega = A \cup A^c$, the second axiom (ii) leads to $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c)$ and the third axiom (iii) leads to $\mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$. Altogether, we have $\mathrm{P}(A) + \mathrm{P}(A^c) = 1$.

2. Show that, if $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$: We can express $B$ as a union of two disjoint sets: $B = (B \cap A) \cup (B \cap A^c)$. Since $A \subseteq B$, we have that $B \cap A = A$. So, using the axiom (iii), we have that

$$\mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c) = \mathrm{P}(A) + \mathrm{P}(B \cap A^c).$$

Using the fact that the probability measure is nonnegative (axiom (i)), we conclude that $\mathrm{P}(B \cap A^c) \geq 0$ and hence

$$\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \cap A^c) \geq \mathrm{P}(A).$$

[2] Recall Definition 2.2.11: A countable index set could have countably infinitely many values, e.g. when $\mathcal{I} = \mathbb{N}$ or finitely many values, e.g. when $\mathcal{I} = \{1, 2\}$.

3. Show $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$: We express $A$ and $B$ in terms of disjoint unions:

$$A = (A \cap B) \cup (A \cap B^c), \qquad B = (B \cap A) \cup (B \cap A^c).$$

By axiom (iii), we have that

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c), \qquad \mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c).$$

Hence

$$\mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) + \mathrm{P}(B \cap A^c).$$

Also the union $A \cup B$ can be expressed as a union of disjoint sets

$$A \cup B = (A \cap B) \cup (A \cap B^c) \cup (B \cap A^c).$$

By axiom (iii), we have that

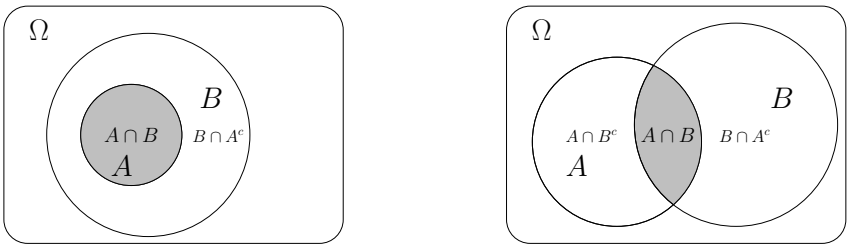
$$\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) + \mathrm{P}(B \cap A^c).$$

So, indeed,

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

□

Some graphical illustrations for the arguments presented in the above proof are given in Figure 4.1.




(a) (b)

Figure 4.1: We consider a sample space $\Omega$ with subsets $A, B \subseteq \Omega$. Figures 4.1a and 4.1b depict the settings we are considering in the proofs of Theorem 4.2.3 part 2 and 3, respectively.

**End of lecture 5.** .....

**Remark 4.2.4.** *The above theorem implies that* $\mathrm{P}(\emptyset) = 0$. *To see this, note that* $\Omega$ *and* $\emptyset$ *are disjoint since* $\Omega \cap \emptyset = \emptyset$. *Also,* $\Omega = \Omega \cup \emptyset$. *So, altogether we have* $1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega \cup \emptyset) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) = 1 + \mathrm{P}(\emptyset)$. *Hence,* $\mathrm{P}(\emptyset) = 0$.

### 4.2.3 Examples

**Example 4.2.5.** *We continue with the classical example of flipping a fair coin.*

Figure 4.2: Flipping a fair coin.

*We write $H$ for heads and $T$ for tail. The sample space is given by $\Omega = \{H, T\}$. The event space can be taken as $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$ which is the collection of all subsets of $\Omega$. Since we are considering a "fair" coin, we have that $\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$, where we typically shorten the notation to:*

$$\mathrm{P}(H) = \mathrm{P}(T) = \frac{1}{2}.$$

*Moreover, we have $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega) = 1$.*

**Exercise 4.2.6.** *Let $\Omega = \mathbb{N} \cup \{0\}$, $\mathcal{F} = \mathcal{P}(\mathbb{N} \cup \{0\})$, $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ with $\mathrm{P}(A) = \sum_{x \in A} \frac{e^{-\lambda}\lambda^x}{x!}$ for $\lambda > 0$. Show that $(\Omega, \mathcal{F}, \mathrm{P})$ is a probability space.*

*Proof.* We know from lectures that the power set is a $\sigma$-algebra. So we only need to show that $\mathrm{P}$ is a probability measure. We note that $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ and hence it remains to check the three axioms of the definition of a probability measure:

**Axiom (ii)** We use the fact that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$, to deduce that

$$\mathrm{P}(\Omega) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$

**Axiom (i)** Since $\lambda > 0$, we have that for any $A \in \mathcal{F}$,

$$0 \leq \mathrm{P}(A) = \sum_{x \in A} \frac{\lambda^x}{x!} e^{-\lambda} \leq \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1.$$

**Axiom (iii)** Let $A_1, A_2, \ldots \in \mathcal{F}$ be disjoint, then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{x \in \cup_{i=1}^{\infty} A_i} \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{i=1}^{\infty} \sum_{x \in A_i} \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{i=1}^{\infty} \mathrm{P}(A_i),$$

where we interchanged the order of summation (to be covered in Analysis!).

□

# Chapter 5

# Conditional probabilities

*The material of this chapter is based on Blitzstein & Hwang (2019), p.45-63, Anderson et al. (2018), p.43-56, Grimmett & Welsh (1986), p.11-12.*

After having introduced the axiomatic definition of a probability measure, we will now turn our attention to so-called conditional probabilities. We will learn how probabilities can be computed based on some given evidence. Conditional probabilities play a key role in almost all subsequent probability and statistics course und you will learn that they constitute a powerful concept for computing unconditional probabilities as well.

## 5.1 Definition

**Definition 5.1.1** (Conditional probability)**.** *Consider a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*. Consider events* $A, B \in \mathcal{F}$ *with* $\mathrm{P}(B) > 0$*. Then the* conditional probability *of* $A$ *given* $B$*, denoted by* $\mathrm{P}(A|B)$*, is defined as*

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

**Remark 5.1.2.** *Interpretation: You could call* $\mathrm{P}(A)$ *the* prior *probability of event* $A$ *and* $\mathrm{P}(A|B)$ *the* posterior *probability of* $A$*. Here we view* $B$ *as additional evidence which becomes available, and the prior probability is formulated without knowledge of the additional evidence and the posterior probability describes the updated probability based on the additional evidence.*

Let us now show that the conditional probability measure does indeed satisfy the axioms of a probability measure:

**Theorem 5.1.3** (Conditional probability)**.** *Let* $B \in \mathcal{F}$ *with* $\mathrm{P}(B) > 0$ *and define* $\mathrm{Q} : \mathcal{F} \to \mathbb{R}$ *by* $\mathrm{Q}(A) = \mathrm{P}(A|B)$*. Then* $(\Omega, \mathcal{F}, \mathrm{Q})$ *is a probability space.*

*Proof.* The only thing we need to show is that $\mathrm{Q}$ satisfies the axioms of a probability measure on $(\Omega, \mathcal{F})$.
**Axiom (i):** Since $\mathrm{P}$ is a probability measure satisfying axiom (i), we deduce that, for any $A \in \mathcal{F}$,

$$\mathrm{Q}(A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \geq 0.$$

**Axiom (ii):**

$$\mathrm{Q}(\Omega) = \frac{\mathrm{P}(\Omega \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B)}{\mathrm{P}(B)} = 1.$$

**Axiom (iii):** Consider disjoint events $A_1, A_2, \cdots \in \mathcal{F}$. Then

$$\mathrm{Q}\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\mathrm{P}((\bigcup_{i=1}^{\infty} A_i) \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(\bigcup_{i=1}^{\infty} (A_i \cap B))}{\mathrm{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathrm{P}(A_i \cap B)}{\mathrm{P}(B)}$$

$$= \sum_{i=1}^{\infty} \frac{\mathrm{P}(A_i \cap B)}{\mathrm{P}(B)} = \sum_{i=1}^{\infty} \mathrm{Q}(A_i),$$

where for the third equality we used the fact that the events $A_i \cap B \in \mathcal{F}$ are disjoint and that $\mathrm{P}$ satisfies axiom (iii). □

## 5.2 Examples

**Example 5.2.1.** *We roll a single fair die. Hence* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *What is the probability that the score is greater than 3 given that the score is even? We define the events*

$$B = \{\omega \in \Omega : \omega \text{ is even}\} = \{2, 4, 6\},$$

$$A = \{\omega \in \Omega : \omega > 3\} = \{4, 5, 6\}.$$

*Then,* $A \cap B = \{4, 6\}$. *We have* $\mathrm{P}(A) = 1/2, \mathrm{P}(B) = 1/2$ *and* $\mathrm{P}(A \cap B) = 2/6 = 1/3$.

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

**Example 5.2.2.** *A family has two children. Assume that Female (F)/Male (M) are equally likely and successive births are independent. We write* $\Omega = \{FF, FM, MF, MM\}$, *where e.g.* $FM$ *stands for the event that the first child is Female and the second child is Male. Note that all four outcomes are equally likely, so* $\mathrm{P}(\omega) = 1/4$ *for all* $\omega \in \Omega$.

1. *If one child is a boy, what is the probability that the other child is a boy?*
2. *If the eldest is a boy, what is the probability that the other child is a boy?*

*Let* $A = \{MM\}$ *both male,* $B = \{MM, MF, FM\}$ *at least one male,* $C = \{MM, MF\}$ *eldest is male. Then*

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1/4}{3/4} = \frac{1}{3},$$

*and*

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

**Remark 5.2.3.** *Consider the case of a finite state space* $\Omega$ *where all events are equally likely and hence the classical interpretation of probability can be used. Then for two events* $A, B \subseteq \Omega$, *we have that*

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

## 5.3 Multiplication rule

Suppose that $A, B \in \mathcal{F}$ with $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$. Then

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

Similarly,

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

Let us extend the above result to three events: Let $C \in \mathcal{F}$ with $\mathrm{P}(C) > 0$. Then

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

Repeating the arguments above, we obtain the following result:

**Theorem 5.3.1** (Multiplication rule)**.** *Let $n \in \mathbb{N}$, then for any events $A_1, \ldots, A_n$ with $\mathrm{P}(A_2 \cap \cdots \cap A_n) > 0$, we have*

$$\begin{aligned}\mathrm{P}(A_1 \cap \cdots \cap A_n) = \mathrm{P}(A_1|A_2 \cap \cdots \cap A_n)\mathrm{P}(A_2|A_3 \cap \cdots \cap A_n) \cdots \\ \mathrm{P}(A_{n-2}|A_{n-1} \cap A_n)\mathrm{P}(A_{n-1}|A_n)\mathrm{P}(A_n),\end{aligned}$$

*where the right hand side is a product of $n$ terms.*

Clearly, the ordering of the events in the theorem above can be changed, so there are $n!$ possible formulations of the above theorem!

**End of lecture 6.** .....

## 5.4 Bayes' rule and law of total probability

### 5.4.1 Bayes' rule

Bayes' rule is a famous and extremely useful result for computing conditional probabilities:

**Theorem 5.4.1.** *Let $A, B \in \mathcal{F}$ with $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$. Then*

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

*Proof.* This is an immediate consequence of the definition of conditional probability and the multiplication rule. □

### 5.4.2 Law of total probability

Next we study the so-called *law of total probability*, which is an extremely useful tool for computing complicated probabilities in terms of simpler pieces based on conditional probabilities.

**Definition 5.4.2** (Partition)**.** *A* partition *of the sample space $\Omega$ is a collection $\{B_i : i \in \mathcal{I}\}$ (for a countable index set $\mathcal{I}$) of disjoint events (meaning that $B_i \in \mathcal{F}$ and $B_i \cap B_j = \emptyset$ for $i \neq j$) such that $\Omega = \bigcup_{i \in \mathcal{I}} B_i$.*

**Remark 5.4.3.** *We note that a partition of the sample space is often not unique and the choice of the particular partition typically very much depends on the problem we want to solve!*

**Theorem 5.4.4** (Law of total probability)**.** *Let $\{B_i : i \in \mathcal{I}\}$ denote a partition of $\Omega$, with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$. Then, for all $A \in \mathcal{F}$,*

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i) = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i)\mathrm{P}(B_i).$$

*Proof.* Using the properties of a partition and the distributivity property, we deduce that

$$A = A \cap \Omega = A \cap \left( \bigcup_{i \in \mathcal{I}} B_i \right) = \bigcup_{i \in \mathcal{I}} (A \cap B_i),$$

where the events $A \cap B_i$ are disjoint. Using axiom (iii) of the definition of the probability measure leads to

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i),$$

and using the multiplication formula leads to

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i) = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i)\mathrm{P}(B_i).$$

□

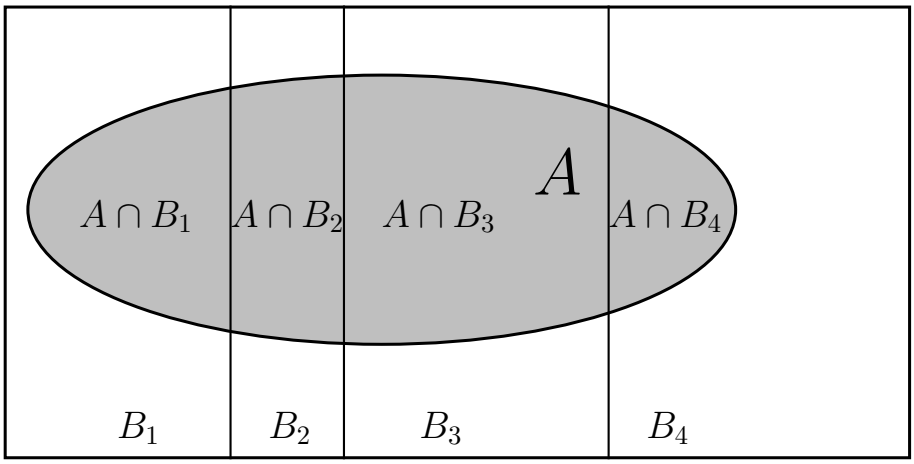



Figure 5.1: Illustration of the law of total probability. Here we have that $\mathrm{P}(A) = \sum_{i=1}^{4} \mathrm{P}(A \cap B_i)$.

### 5.4.3 General Bayes' rule

When we combined Bayes' rule with the law of total probability, we get the following useful result:

**Theorem 5.4.5.** *Consider a partition $\{B_i : i \in \mathcal{I}\}$ of $\Omega$ with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$, then for any event $A \in \mathcal{F}$ with $\mathrm{P}(A) > 0$, we have*

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\sum_{k \in \mathcal{I}} \mathrm{P}(A|B_k)\mathrm{P}(B_k)}.$$

### 5.4.4 Bayes' rule and law of total probability with additional conditioning

The Bayes' rule and the law of total probability also hold with extra conditioning.

**Theorem 5.4.6** ( Bayes' rule with extra conditioning)**.** *For events $A, B, E$ with $\mathrm{P}(A \cap E) > 0, \mathrm{P}(B \cap E) > 0$, we have*

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

**Theorem 5.4.7** (Law of total probability with additional conditioning)**.** *Consider events $A, E$ with $\mathrm{P}(E) > 0$ and let $\{B_i : i \in \mathcal{I}\}$ denote a partition of $\Omega$, with $\mathrm{P}(B_i \cap E) > 0$ for all $i \in \mathcal{I}$. Then,*

$$\mathrm{P}(A|E) = \sum_{i\in\mathcal{I}} \frac{\mathrm{P}(A \cap B_i \cap E)}{\mathrm{P}(E)} = \sum_{i\in\mathcal{I}} \mathrm{P}(A|B_i \cap E)\mathrm{P}(B_i|E).$$

The proof of these two results is left as an exercise, see Exercise 3- 3.

## 5.5 Examples

### 5.5.1 Examples: Cards and marbles

**Example 5.5.1.** *We pick 2 cards at random from a well-shuffled 52-card deck. We consider the event $E$:="2nd card is red". What is $\mathrm{P}(E)$?*

*We use the law of total probability to deduce that*

$$\begin{aligned}\mathrm{P}(E) &= \mathrm{P}(E|\text{ 1st card red})\mathrm{P}(\text{1st card red}) + \mathrm{P}(E|\text{1st card black})\mathrm{P}(\text{1st card black})\\ &= \frac{25}{51}\cdot\frac{1}{2} + \frac{26}{51}\cdot\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

**Example 5.5.2.** *You have three bags that each contain 100 marbles. Bag 1 has 70 red and 30 green marbles. Bag 2 has 60 red and 40 green marbles. Bag 3 has 50 red and 50 green marbles. 1) If you choose one bag at random and then pick a marble at random from the chosen bag. What is the probability that the chosen marble is red? 2) Suppose that the chosen marble was red, what is the probability that bag 1 was chosen?*

1. *We define the events $B_i$:=bag $i$ was chosen, $\mathrm{P}(B_i) = \frac{1}{3}$, $i = 1, 2, 3$. $R :=$ marble is red. Then*

$$\mathrm{P}(R|B_1) = \frac{7}{10}, \qquad \mathrm{P}(R|B_2) = \frac{6}{10}, \qquad \mathrm{P}(R|B_3) = \frac{5}{10},$$

*We use the law of total probability:*

$$\mathrm{P}(R) = \mathrm{P}(R|B_1)\mathrm{P}(B_1) + \mathrm{P}(R|B_2)\mathrm{P}(B_2) + \mathrm{P}(R|B_3)\mathrm{P}(B_3) = \left(\frac{7}{10} + \frac{6}{10} + \frac{5}{10}\right)\frac{1}{3} = 0.6.$$

2. *We use Bayes' rule:*

$$\mathrm{P}(B_1|R) = \frac{\mathrm{P}(R|B_1)\mathrm{P}(B_1)}{\mathrm{P}(R)} = \frac{\frac{7}{10}\cdot\frac{1}{3}}{\frac{6}{10}} = \frac{7}{18} \approx 0.38 > \frac{1}{3}.$$

### 5.5.2 Example: Testing for a rare disease

**Example 5.5.3.** *Let us look in more detail into visualisation of conditional probabilities, which often arise in medical screening tests. We quote the example from the article (Spiegelhalter et al. 2011, p. 1396).*

*Consider a*

> *'mammography test on a population with a 1% prevalence of breast cancer. The test is positive for around 90% of women with cancer, but it is also positive for around 10% of women without cancer'. ((Spiegelhalter et al. 2011, p. 1396))*

*What is the probability that a woman whose mammography test is positive has breast cancer?*

*Define the events: $B$ := breast cancer present; $TP$ = test is positive. We know that $\mathrm{P}(B) = 0.01$ and $\mathrm{P}(TP|B) = 0.9$ and $\mathrm{P}(TP|B^c) = 0.1$. We can derive that $\mathrm{P}(B^c) = 1 - \mathrm{P}(B) = 0.99$. Further, by the law of total probability,*

$$\mathrm{P}(TP) = \mathrm{P}(TP|B)\mathrm{P}(B) + \mathrm{P}(TP|B^c)\mathrm{P}(B^c)$$

$$= 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108.$$

*Using Bayes' Theorem, we get*

$$\mathrm{P}(B|TP) = \frac{\mathrm{P}(B \cap TP)}{\mathrm{P}(TP)} = \frac{\mathrm{P}(TP|B)\mathrm{P}(B)}{\mathrm{P}(TP)} = \frac{0.9 \cdot 0.01}{0.108} \approx 8\%$$

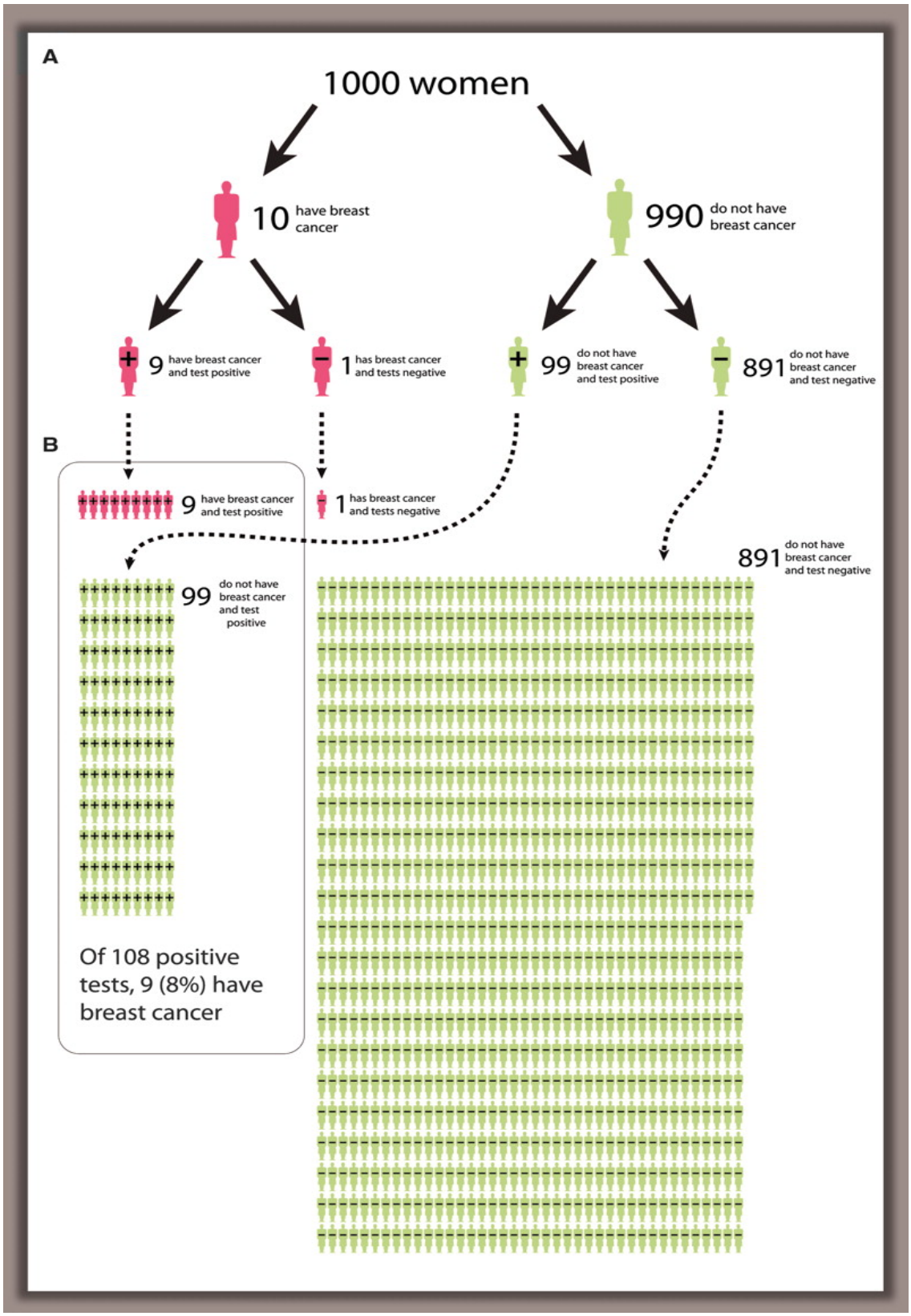



Figure 5.2: Visualising conditional probabilities in the case of testing for a rare disease. This picture is a copy of Figure 4 in the article Spiegelhalter et al. (2011).

### 5.5.3 Example: Monty Hall – Conditioning on the missing information

As a general strategy, we typically use conditioning on the information we wish we had to make our probability computations easier. To illustrate this strategy, we consider a famous example. In the TV Game show *Let's make a deal*, hosted by Monty Hall, a contestant selects one of three doors; behind one of the doors there is a prize (a car), and behind the other two there are no prizes (in fact, there are goats!). After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice. Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH ? (Assume that the host selects a door to open, from those available, with equal probability).

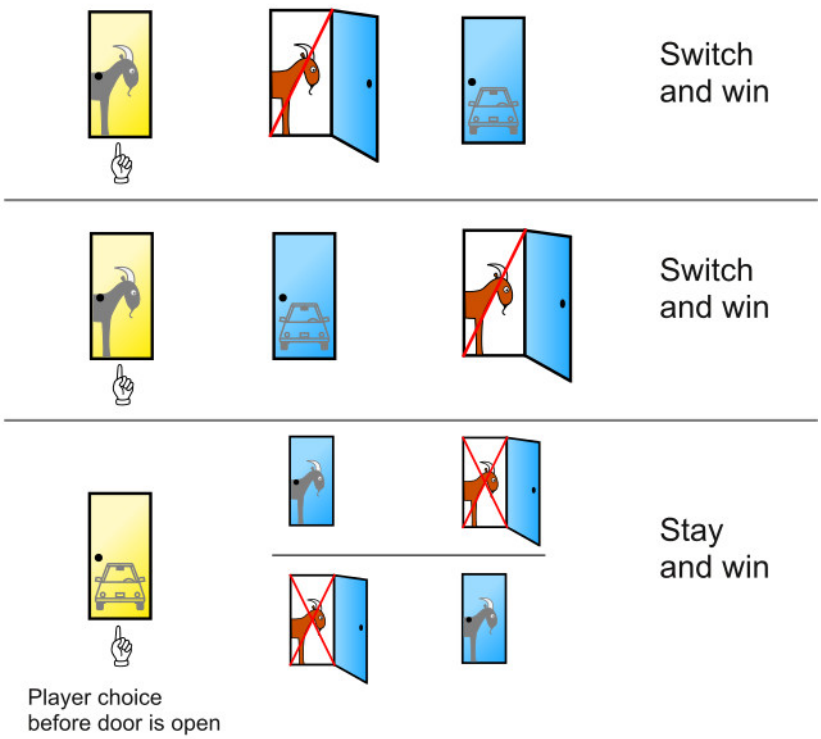



Figure 5.3: In the Monty Hall example it is advantageous to SWITCH!

Label the doors 1, 2 and 3 and assume without loss of generality that the candidate selects door 1. Monty Hall then opens a door and reveals a goat. When deciding whether or not to switch, which information would the contestant like to have? She would like to know the location of the car.

So let us consider the partition: $\{C_i, i = 1, 2, 3\}$ where $C_i$ is the event that the car is behind the $i$th door (for $i = 1, 2, 3$). Also, we denote by $H_2$ the event that Monty Hall opens door 2. Then $\mathrm{P}(C_1) = \mathrm{P}(C_2) = \mathrm{P}(C_3) = 1/3$ and $\mathrm{P}(H_2|C_1) = 1/2$, $\mathrm{P}(H_2|C_2) = 0$ and $\mathrm{P}(H_2|C_3) = 1$. We want to compare the probabilities of $\mathrm{P}(C_1|H_2)$ (STICK) with $\mathrm{P}(C_3|H_2)$ (SWITCH). Using the law of total probability, we have

$$\begin{aligned}\mathrm{P}(H_2) &= \mathrm{P}(H_2|C_1)\mathrm{P}(C_1) + \mathrm{P}(H_2|C_2)\mathrm{P}(C_2) + \mathrm{P}(H_2|C_3)\mathrm{P}(C_3) \\ &= \frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} = \frac{1}{2}.\end{aligned}$$

Then, using the (general) Bayes' rule implies that

$$\mathrm{P}(C_1|H_2) = \frac{\mathrm{P}(H_2|C_1)\mathrm{P}(C_1)}{\mathrm{P}(H_2)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Similarly

$$\mathrm{P}(C_3|H_2) = \frac{\mathrm{P}(H_2|C_3)\mathrm{P}(C_3)}{\mathrm{P}(H_2)} = \frac{1\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

So it is better to SWITCH, see Figure 5.3.

**End of lecture 7.** . . . . . .

# Chapter 6

# Independence

*The material of this chapter is based on Blitzstein & Hwang (2019), p.63-65, Anderson et al. (2018), p.51-56, Grimmett & Welsh (1986), p.12-16.*

## 6.1 Independence of events

We will call two events $A, B \in \mathcal{F}$ *independent* if the occurrence of one of them does not affect the probability that the other one occurs, meaning that, if $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$,

$$\mathrm{P}(A|B) = \mathrm{P}(A), \text{ and } \mathrm{P}(B|A) = \mathrm{P}(B). \tag{6.1.1}$$

Now, recall the definition of the conditional probability as

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

Hence, the following definition appears suitable:

**Definition 6.1.1** (Independent events)**.** *The events $A, B$ are called* independent *if*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B), \tag{6.1.2}$$

*and* dependent *otherwise.*

**Remark 6.1.2.** *The definition given in equation 6.1.2 is more general than the one in equation 6.1.1 since it does not require that $A$ and $B$ have nonzero probabilities.*

**Theorem 6.1.3.** *If the events $A$ and $B$ are independent, then the same is true for each of the pairs $A^c$ and $B$, $A$ and $B^c$, and $A^c$ and $B^c$.*

*Proof.* We only prove that $\mathrm{P}(A^c \cap B) = \mathrm{P}(A^c)\mathrm{P}(B)$ (the proof for the remaining pairs follows the same arguments). Let us start from the left hand side of the equation:

$$\mathrm{P}(A^c)\mathrm{P}(B) = (1 - \mathrm{P}(A))\mathrm{P}(B) = \mathrm{P}(B) - \mathrm{P}(A)\mathrm{P}(B).$$

From the law of total probability, we deduce that

$$\mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c),$$

also

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

since $A$ and $B$ are independent. Overall we get

$$\mathrm{P}(A^c)\mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c) - \mathrm{P}(B \cap A) = \mathrm{P}(B \cap A^c).$$

□

Let us now generalise the definition of independence to more than two events.

**Definition 6.1.4** (Independence of events (general case)). *1. A finite collection of events $A_1, \dots, A_n$ is defined to be independent if*

$$\mathrm{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathrm{P}(A_{i_1})\mathrm{P}(A_{i_2}) \cdots \mathrm{P}(A_{i_k})$$

*for every subcollection $\{i_1, \dots, i_k\}$ of $\{1, \dots, n\}$, $k = 1, \dots, n$.*

*2. A countable or uncountably infinite collection of events is defined to be independent if each finite subcollection is independent.*

**Remark 6.1.5.** *Note that pairwise independence of $(A_i, A_j)$ is in general not sufficient to conclude the independence of $(A_1, \dots, A_n)$.*

**Example 6.1.6.** *The three events $A_1, A_2, A_3$ are independent if and only if*

$$\begin{aligned}
\mathrm{P}(A_1 \cap A_2 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3), \\
\mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_1)\mathrm{P}(A_2), \\
\mathrm{P}(A_1 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_3), \\
\mathrm{P}(A_2 \cap A_3) &= \mathrm{P}(A_2)\mathrm{P}(A_3),
\end{aligned}$$

**Example 6.1.7.** *We roll two fair dice and write the sample space as $\Omega = \{(i, j) : i, j = 1, \dots, 6\}$. We note that $\mathrm{card}(\Omega) = 6^2 = 36$ and all outcomes are equally likely. We define three events: $A_1 =$ first roll is odd, $A_2 =$ second roll is odd, $A_3 =$ sum is odd. Then $A_1, A_2, A_3$ are pairwise independent, but they are not independent since $\mathrm{P}(A_1 \cap A_2 \cap A_3) = 0 \neq \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3)$.*

### 6.1.1 Conditional independence of events

**Definition 6.1.8** (Conditional independence of events). *Consider events $A, B, C \in \mathcal{F}$ with $\mathrm{P}(C) > 0$. Then we say that $A$ and $B$ are* conditionally independent given *$C$ if*

$$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C). \tag{6.1.3}$$

If we, in addition, assume that $\mathrm{P}(B \cap C) > 0$, then equation (6.1.3) is equivalent to the condition

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

The equivalence can be shown as follows:

$$\begin{aligned}
&\mathrm{P}(A|B \cap C) = \frac{\mathrm{P}(A \cap B \cap C)}{\mathrm{P}(B \cap C)} = \frac{\mathrm{P}(A \cap B|C)\mathrm{P}(C)}{\mathrm{P}(B \cap C)} = \mathrm{P}(A|C) \\
&\Leftrightarrow \mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\frac{\mathrm{P}(B \cap C)}{\mathrm{P}(C)} \\
&\Leftrightarrow \mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C).
\end{aligned}$$

**Example 6.1.9.** *You have a fair and an "unfair" coin. The unfair coin lands heads with probability $\frac{3}{4}$. You pick one coin at random and toss it three times. It lands heads three times. Given this information, what is the probability that you picked the fair coin?*

*Let $A$:=the chosen coin lands heads three times, $F$:= you picked the fair coin. We want to find $\mathrm{P}(F|A)$. We use the generalised Bayes rule and the fact that*

$$\mathrm{P}(F) = \frac{1}{2} = \mathrm{P}(F^c),\ \mathrm{P}(A|F) = \left(\frac{1}{2}\right)^3,\ \mathrm{P}(A|F^c) = \left(\frac{3}{4}\right)^3.$$

*Then:*

$$\mathrm{P}(F|A) = \frac{\mathrm{P}(A|F)\mathrm{P}(F)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|F)\mathrm{P}(F)}{\mathrm{P}(A|F)\mathrm{P}(F) + \mathrm{P}(A|F^c)\mathrm{P}(F^c)} = \frac{\left(\frac{1}{2}\right)^3 \cdot \frac{1}{2}}{\left(\frac{1}{2}\right)^3 \cdot \frac{1}{2} + \left(\frac{3}{4}\right)^3 \cdot \frac{1}{2}} \approx 0.23.$$

*After having seen that it landed heads three times, what is the probability that, if we toss the same coin a fourth time, it lands heads a fourth time?*

*Let $H :=$ chosen coin lands heads at 4th toss. We want to find $\mathrm{P}(H|A)$. Here we use the law of total probability with extra conditioning:*

$$\mathrm{P}(H|A) = \mathrm{P}(H|A \cap F)\mathrm{P}(F|A) + \mathrm{P}(H|A \cap F^c)\mathrm{P}(F^c|A).$$

*Note that $H$ and $A$ are conditional independent given $F$ since*

$$\mathrm{P}(H \cap A|F) = \frac{1}{2^4} = \frac{1}{2} \cdot \left(\frac{1}{2}\right)^3 = \mathrm{P}(H|F)\mathrm{P}(A|F).$$

*Hence $\mathrm{P}(H|A \cap F) = \mathrm{P}(H|F) = \frac{1}{2}$. Similarly, $\mathrm{P}(H|A \cap F^c) = \mathrm{P}(H|F^c) = \frac{3}{4}$. Hence*

$$\mathrm{P}(H|A) \approx \frac{1}{2} \cdot 0.23 + \frac{3}{4}(1 - 0.23) = 0.69.$$

### 6.1.2 Continuity of the probability measure and product rule

**Definition 6.1.10.** *The* set difference *between two sets $A, B \subseteq \Omega$, denoted by $A \setminus B$ is defined as*

$$A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\} = A \cap B^c.$$

**Lemma 6.1.11.** *Any countable union can be written as a countable union of disjoint sets. I.e. let $A_1, A_2, \cdots \in \mathcal{F}$ and define $D_1 = A_1, D_2 = A_2 \setminus A_1, D_3 = A_3 \setminus (A_1 \cup A_2), \ldots$. Then $\{D_i\}$ is a collection of disjoint sets and $\cup_{i=1}^n A_i = \cup_{i=1}^n D_i$ for $n$ being any positive integer or $\infty$.*

The proof of the Lemma is left as an exercise, see Exercise 3- 5.

**Definition 6.1.12** (Increasing and decreasing sets)**.** *A sequence of sets $(A_i)_{i=1}^\infty$ is said to* increase *to $A$, i.e. $A_i \uparrow A$, if $A_1 \subseteq A_2 \subseteq \cdots$ and $\cup_{i=1}^\infty A_i = A$. Similarly, a sequence of sets $(A_i)_{i=1}^\infty$ is said to* decrease *to $A$, i.e. $A_i \downarrow A$, if $A_1 \supseteq A_2 \supseteq \cdots$ and $\cap_{i=1}^\infty A_i = A$.*

Note that $A_i \uparrow A$ if and only if $A_i^c \downarrow A^c$.

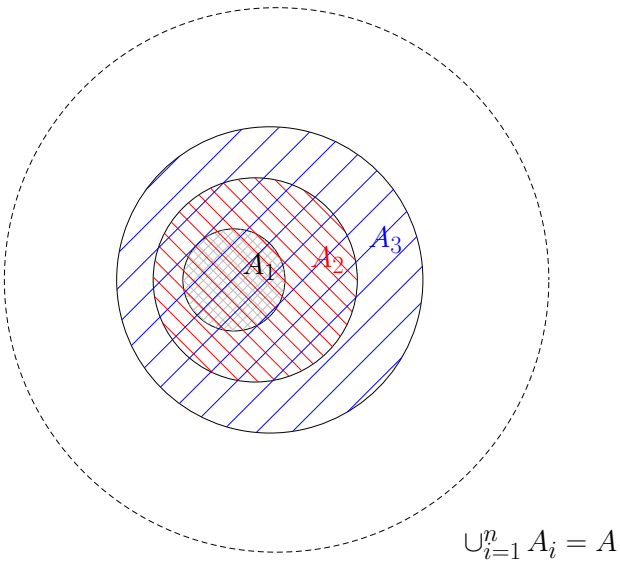



Figure 6.1: Example of increasing sets in $\mathbb{R}^2$.

**Exercise 6.1.13.** *Give an example of an increasing and a decreasing sequence of sets.*

Many examples are possible. We will be discussing two examples here.

**Example 6.1.14.** *Example of an increasing sequence of sets: Let* $\Omega = (0, \infty)$ *and define* $A_n := \left(0, 1 - \frac{1}{n}\right]$. *Then*

$$A_1 = (0, 0] = \emptyset \subseteq A_2 = \left(0, \frac{1}{2}\right] \subseteq A_3 = \left(0, \frac{2}{3}\right] \subseteq \cdots$$

*with*

$$A_n \uparrow A := \bigcup_{n=1}^{\infty} A_n = (0, 1).$$

*We note that the sequence of complements is decreasing:* $A_n^c \downarrow A^c$*: We have* $A_n^c = \left(1 - \frac{1}{n}, \infty\right)$ *and hence*

$$A_1^c = (0, \infty) = \Omega \supseteq A_2^c = \left(\frac{1}{2}, \infty\right) \supseteq A_3^c = \left(\frac{2}{3}, \infty\right) \supseteq \cdots$$

*with*

$$A_n^c \downarrow A^c = \bigcap_{n=1}^{\infty} A_n^c = [1, \infty).$$

**Example 6.1.15.** *Example of an increasing sequence of sets: Let* $\Omega = \mathbb{R}$ *and define* $A_n := (-\infty, n]$. *Then*

$$A_1 = (-\infty, 1] \subseteq A_2 = (-\infty, 2] \subseteq A_3 = (-\infty, 3] \subseteq \cdots$$

*with*

$$A_n \uparrow A := \bigcup_{n=1}^{\infty} A_n = \Omega = \mathbb{R}.$$

*We note that the sequence of complements is decreasing:* $A_n^c \downarrow A^c$*: We have* $A_n^c = (n, \infty)$ *and hence*

$$A_1^c = (1, \infty) = \Omega \supseteq A_2^c = (2, \infty) \supseteq A_3^c = (3, \infty) \supseteq \cdots$$

*with*

$$A_n^c \downarrow A^c = \bigcap_{n=1}^{\infty} A_n^c = \emptyset.$$

Next we will state and prove the continuity property of the probability measure[1].

**Theorem 6.1.16.** *If* $A_1, A_2, \cdots \in \mathcal{F}$ *and* $A_i \uparrow A$ *or* $A_i \downarrow A$*, then*

$$\lim_{i \to \infty} \mathrm{P}(A_i) = \mathrm{P}(A).$$

The above theorem states that, for increasing or decreasing sets, we can interchange the limit operation and the probability measure, i.e. we have

$$\lim_{i \to \infty} \mathrm{P}(A_i) = \mathrm{P}(\lim_{i \to \infty} A_i),$$

where the set limit on the right hand side needs to be understood as taking an infinite union or intersection for increasing and decreasing sequences, respectively.

[1] Recall that a sequence of real numbers $(x_n)$ is said to converge to a real number $x$ if for all $\epsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $|x_n - x| < \epsilon$.

*Proof.* Suppose that $A_i \uparrow A$. Then using Lemma 6.1.11, we write $A = \cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} D_i$, where the $D_i = A_i \setminus (\cup_{k=1}^{i-1} A_k)$ are disjoint. By axiom (iii) of the definition of the probability measure (applied twice), we deduce that

$$\begin{aligned} \mathrm{P}(A) &= \sum_{i=1}^{\infty} \mathrm{P}(D_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(D_i) \\ &= \lim_{n\to\infty} \mathrm{P}(\cup_{i=1}^{n} D_i) = \lim_{n\to\infty} \mathrm{P}(\cup_{i=1}^{n} A_i) = \lim_{n\to\infty} \mathrm{P}(A_n). \end{aligned}$$

Now let $A_i \downarrow A$. Then $F_i = A_i^c \uparrow F = A^c$. Then, we deduce from the first part of the proof that $\lim_{i\to\infty} \mathrm{P}(F_i) = \mathrm{P}(F)$. Using the properties of a probability measure, since $\mathrm{P}(F_i) = 1 - \mathrm{P}(A_i)$ and $\mathrm{P}(F) = 1 - \mathrm{P}(A)$, we deduce that $\lim_{i\to\infty} \mathrm{P}(A_i) = \mathrm{P}(A)$. □

**End of lecture 8.** .....

We can now formulate the so-called product rule for countable number of independent sets:

**Theorem 6.1.17.** *If $A_1, A_2, \ldots$ is a countably infinite set of independent events, then*

$$\mathrm{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} \mathrm{P}(A_i).$$

The proof of the above theorem relies on the *continuity property* of the probability measure.

*Proof of Theorem 6.1.17.* Let $B_n = \cap_{i=1}^{n} A_i$. Then $B_n \downarrow B = \cap_{i=1}^{\infty} A_i$, so by the continuity property of the probability measure, see Theorem 6.1.16, we deduce that

$$\begin{aligned} \mathrm{P}(\cap_{i=1}^{\infty} A_i) = \mathrm{P}(B) &= \lim_{n\to\infty} \mathrm{P}(B_n) \\ &= \lim_{n\to\infty} \mathrm{P}(\cap_{i=1}^{n} A_i) = \lim_{n\to\infty} \prod_{i=1}^{n} \mathrm{P}(A_i) = \prod_{i=1}^{\infty} \mathrm{P}(A_i). \end{aligned}$$

□

# Chapter 7

# Discrete random variables

*The material of this chapter is based on Blitzstein & Hwang (2019), p.103-120, Grimmett & Welsh (1986), p.24-28.*

In this and the following chapter, we will be introducing discrete and continuous random variables and their distributions.

## 7.1 Pre-images and their properties

We will be defining random variables using the notion of so-called pre-images which you will study in Analysis. Here we will briefly review their definition and some key properties which we will need in probability.

**Definition 7.1.1.** *Consider a function with domain $\mathcal{X}$ and co-domain $\mathcal{Y}$, i.e. $f : \mathcal{X} \to \mathcal{Y}$.*

- *For any subset $A \subseteq \mathcal{X}$, we define the* image of $A$ under $f$ *as*

$$f(A) = \{y \in \mathcal{Y} : \exists x \in A : f(x) = y\} = \{f(x) : x \in A\}.$$

  *If $A = \mathcal{X}$, then we call $f(\mathcal{X}) = \mathrm{Im} f$ the* image of $f$.

- *For any subset $B \subseteq \mathcal{Y}$, we define the* pre-image of $B$ under $f$ *as*

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

Please not that the pre-image should not be confused with the inverse function (despite the fact that we are using the same notation). The pre-image is well-defined for any function, whereas the inverse function obviously only exists when the function $f$ is invertible.

The definition of the pre-image implies that

$$x \in f^{-1}(B) \Leftrightarrow f(x) \in B.$$

Note that in the case when $B$ is a singleton, i.e. $B = \{b\}$ for an element $b \in \mathcal{Y}$, then we often simplify the notation to $f^{-1}(\{b\}) = f^{-1}(b)$.

**Lemma 7.1.2.** *For any collection of subsets $B_i \subseteq \mathcal{Y}$, $i \in \mathcal{I}$ where $\mathcal{I}$ denotes an (arbitrary) index set, we have that*

$$f^{-1}\left(\bigcup_{i \in \mathcal{I}} B_i\right) = \bigcup_{i \in \mathcal{I}} f^{-1}(B_i).$$

*Proof.* We have that

$$x \in f^{-1}\left(\bigcup_{i\in\mathcal{I}} B_i\right) \Leftrightarrow f(x) \in \bigcup_{i\in\mathcal{I}} B_i \Leftrightarrow \exists\, i \in \mathcal{I} \text{ such that } f(x) \in B_i$$
$$\Leftrightarrow \exists\, i \in \mathcal{I} \text{ such that } x \in f^{-1}(B_i) \Leftrightarrow x \in \bigcup_{i\in\mathcal{I}} f^{-1}(B_i).$$

□

## 7.2 Random variables

We consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and we will think of a *random variable* (r.v.) as a function from the sample space to the real numbers $\mathbb{R}$, i.e.

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

The function needs to satisfy some properties, which we introduce in the formal definition below. Note that

- Despite the name, a random variable is a *function* and not a variable.
- We typically use capital letters such as $X, Y, Z$ to denote random variables.
- The value of the random variable $X$ at the sample point $\omega$ is given by $X(\omega)$ and is called a *realisation* of $X$.
- The randomness stems from $\omega \in \Omega$ (we don't know which outcome $\omega$ appears in the random experiment), the mapping itself given by $X$ is deterministic.

## 7.3 Discrete random variables and probability distributions

**Definition 7.3.1** (Discrete random variable)**.** *A* discrete random variable *on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is defined as a mapping $X : \Omega \to \mathbb{R}$ such that*

*(i) the* image/range of $\Omega$ under $X$ *denoted by $\mathrm{Im}X = \{X(\omega) : \omega \in \Omega\}$ is a countable subset of $\mathbb{R}$,*

*(ii) $X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.*

**Remark 7.3.2.** *The name* discrete *stems from the first condition in the above definition, which says that the random variable can only take countably many values in $\mathbb{R}$. In most applications, we deal with discrete random variables taking values in (a subset of) $\mathbb{N}$ or $\mathbb{Z}$.*

**Remark 7.3.3.** *Let us clarify the second condition in the above definition: We note that the set appearing there is the so-called pre-image of $x$ defined as*

$$X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\},$$

*i.e. the set of all $\omega$ which $X$ maps to $x$. We require that this set is an event in $\mathcal{F}$ (for all possible $x$) so that we can later assign probabilities to these events.*

**Definition 7.3.4** (Probability mass function)**.** *The* probability mass function *(pmf) of the discrete random variable $X$ is defined as the function $p_X : \mathbb{R} \to [0, 1]$ given by*

$$p_X(x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) = \mathrm{P}(X^{-1}(x)). \tag{7.3.1}$$

We typically shorten the notation significantly and write $p_X(x) = \mathrm{P}(X = x)$. Keep in mind, that this is short hand notation for equation (7.3.1).

Note that the definition of the pmf implies the following properties:

$$p_X(x) = 0 \quad \text{if } x \notin \mathrm{Im}X.$$

Note that for $x_1, x_2 \in \mathrm{Im}X$ with $x_1 \neq x_2$, then

$$X^{-1}(x_1) \cap X^{-1}(x_2) = \{\omega \in \Omega : X(\omega) = x_1 \text{ and } X(\omega) = x_2\} = \emptyset.$$

Using axiom (iii) in the definition of the probability measure, we have

$$\begin{aligned}\sum_{x \in \mathrm{Im}X} p_X(x) &= \sum_{x \in \mathrm{Im}X} \mathrm{P}(X^{-1}(x)) = \mathrm{P}\left( \bigcup_{x \in \mathrm{Im}X} X^{-1}(x) \right) \\ &= \mathrm{P}\left( \bigcup_{x \in \mathrm{Im}X} \{\omega \in \Omega : X(\omega) = x\} \right) = \mathrm{P}(\Omega) = 1.\end{aligned}$$

The above equation is often written as

$$\sum_{x \in \mathbb{R}} p_X(x) = 1,$$

since only countably many values of $x$ result in non-zero values for the pmf and hence non-zero contributions to the sum.

**Example 7.3.5.** *Consider the experiment where we toss a fair coin twice. Write $H$ for heads and $T$ for tails. Then $\Omega = \{HH, HT, TH, TT\}$. Define random variables on $\Omega$!*

- *$X$ = number of heads:*

$$X(HH) = 2, \quad X(HT) = H(TH) = 1, \quad X(TT) = 0.$$

- *$Y$ = number of tails: $Y = 2 - X$.*
- *$I = 1$ if first toss lands heads and $0$ otherwise.*

$$I(HH) = I(HT) = 1, \quad I(TH) = I(TT) = 0.$$

*This is a so-called* indicator random variable *indicating whether or not the first toss lands heads (1="yes", 0="no").*

*We can write 1 for H and 0 for T, then $\Omega = \{(1,1), (1,0), (0,1), (0,0)\}$. I.e. $\omega = (\omega_1, \omega_2) \in \Omega$ if $\omega_i \in \{0, 1\}, i = 1, 2$. Then we can express the three random variables defined above as follows:*

$$\begin{aligned}X(\omega_1, \omega_2) &= \omega_1 + \omega_2, \\ Y(\omega_1, \omega_2)) &= 2 - \omega_1 - \omega_2, \\ I(\omega_1, \omega_2) &= \omega_1.\end{aligned}$$

*Note that we can define the event $A := \{(1,1), (1,0)\}$, then $I = \mathbb{I}_A$. Also, note that*

$$\mathrm{P}(\mathbb{I}_A = 1) = \mathrm{P}(\{\omega \in \Omega : \mathbb{I}_A(\omega) = 1\}) = \mathrm{P}(\{\omega \in \Omega : \omega \in A\}) = \mathrm{P}(A).$$

**End of lecture 9.** .....

**Theorem 7.3.6.** *Let $\mathcal{I}$ denote a countable (index) set. Suppose that $S = \{s_i : i \in \mathcal{I}\}$ is a countable set of distinct real numbers and $\{\pi_i : i \in \mathcal{I}\}$ is a collection of numbers satisfying*

$$\pi_i \geq 0 \text{ for all } i \in \mathcal{I}, \text{and } \sum_{i \in \mathcal{I}} \pi_i = 1,$$

*then there exists a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a discrete random variable $X$ on that probability space such that its probability mass function is given by*

$$p_X(s_i) = \pi_i, \qquad \text{for all } i \in \mathcal{I}$$
$$p_X(s) = 0, \qquad \text{if } s \notin S.$$

*Proof.* This is a constructive proof: Take $\Omega = S$, let $\mathcal{F} = \mathcal{P}(\Omega)$ be the power set (i.e. the set of all subsets of $\Omega$) and set

$$\mathrm{P}(A) = \sum_{i : s_i \in A} \pi_i \qquad \text{for all } A \in \mathcal{F}.$$

The discrete random variable $X : \Omega \to \mathbb{R}$ is then defined as $X(\omega) = \omega$ for all $\omega \in \Omega$. □

The above theorem is incredibly useful for our further study of discrete random variables. It implies that we do not need to worry about sample spaces, event spaces and probability measures too much. Instead, we can just say that we study a random variable $X$ taking the value $s_i$ with probability $\pi_i$ for $i \in \mathcal{I}$ and we know that such a random variable actually exists!

**Exercise 7.3.7.** *Can you think of an example of a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and one function $X : \Omega \to \mathbb{R}$ which is a random variable and one function $Y : \Omega \to \mathbb{R}$ which is not a random variable on that probability space?*

Please try solving the exercise before consulting the model solutions below. There are many possible examples, one is stated below and you might find other ones which are equally correct!

*Proof.* Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and define the event space $\mathcal{F} = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$. Define the probability measure $\mathrm{P}$ to be the naive probability measure, i.e. $\mathrm{P}(A) = \mathrm{card}(A)/\mathrm{card}(\Omega)$ for $A \in \mathcal{F}$.

(i) Define $X : \Omega \to \mathbb{R}$ such that $X(\omega) = 1$ if $\omega$ is even and $X(\omega) = -1$ if $\omega$ is odd. Then $\mathrm{Im}X = \{-1, 1\}$ is finite and $X^{-1}(\{-1\}) = \{1, 3, 5\} \in \mathcal{F}$ and $X^{-1}(\{1\}) = \{2, 4, 6\} \in \mathcal{F}$. For all $x \notin \mathrm{Im}X$ we have that $X^{-1}(\{x\}) = \emptyset \in \mathcal{F}$. Hence $X$ is a discrete random variable.

(ii) Define $Y : \Omega \to \mathbb{R}$ such that $Y(\omega) = \omega$. Then e.g. $Y^{-1}(\{1\}) = \{1\} \notin \mathcal{F}$, hence $Y$ is not a (discrete) random variable with respect to the given sigma-algebra $\mathcal{F}$. (It would be one if we had chosen $\mathcal{F} = \mathcal{P}(\Omega)$ to be the power sigma-algebra of $\Omega$ as in the proof of Theorem 7.3.6!)

□

## 7.4 Common discrete distributions

In this section, we will introduce some widely used discrete distributions.

### 7.4.1 Bernoulli distribution

**Definition 7.4.1** (Bernoulli distribution)**.** *A discrete random variable $X$ is said to have* Bernoulli distribution *with parameter $p \in (0, 1)$, if $X$ can only take two possible values, 0 and 1, i.e. $\mathrm{Im}X = \{0, 1\}$ and*

$$p_X(1) = \mathrm{P}(X = 1) = p, \quad p_X(0) = \mathrm{P}(X = 0) = 1 - p, \quad p_X(x) = 0 \text{ if } x \notin 0, 1.$$

*We write $X \sim \mathrm{Bern}(p)$.*

Note that for *any* event there is a natural way of associating a Bernoulli random variable with it: We can define the so-called indicator variable of the event:

**Definition 7.4.2** (Indicator variable)**.** *Consider an event $A \in \mathcal{F}$, we denote by*

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases}$$

*the* indicator variable of the event $A$.

Note that the random variable $\mathbb{I}_A \sim \text{Bern}(p)$ with $p = \mathrm{P}(A)$, since

$$\mathrm{P}(\mathbb{I}_A = 1) = \mathrm{P}(A), \quad \mathrm{P}(\mathbb{I}_A = 0) = \mathrm{P}(A^c) = 1 - \mathrm{P}(A), \quad \mathrm{P}(\mathbb{I}_A = x) = 0 \text{ for } x \notin \{0, 1\}.$$

Background: Think of an experiment with two possible outcomes "success" or "failure" (but not both). We call such an experiment a *Bernoulli trial*. We can think of a Bernoulli random variable as an indicator of success, where an outcome of 1 represents success and an outcome of 0 represents failure. Hence we often call the parameter $p$ in the Bernoulli distribution the *success probability*.

### 7.4.2 Binomial distribution

Consider a sequence of $n \in \mathbb{N}$ independent and identical Bernoulli trials with success probability $p \in (0, 1)$ and count the number of successes and denote it by the random variable $X$ [e.g. count the number of heads when tossing a coin repeatedly]. Then $X$ can take the values $\mathrm{Im}X = \{0, 1, \ldots, n\}$. Let $x \in \mathrm{Im}X$, and suppose we have $x$ successes and $n - x$ failures. Since the trials are independent, the probability of any sequence with $x$ successes and $n - x$ failures is $p^x(1 - p)^{n-x}$. In total, there are $\binom{n}{x}$ possible sequences with $x$ successes and $n - x$ failures, hence

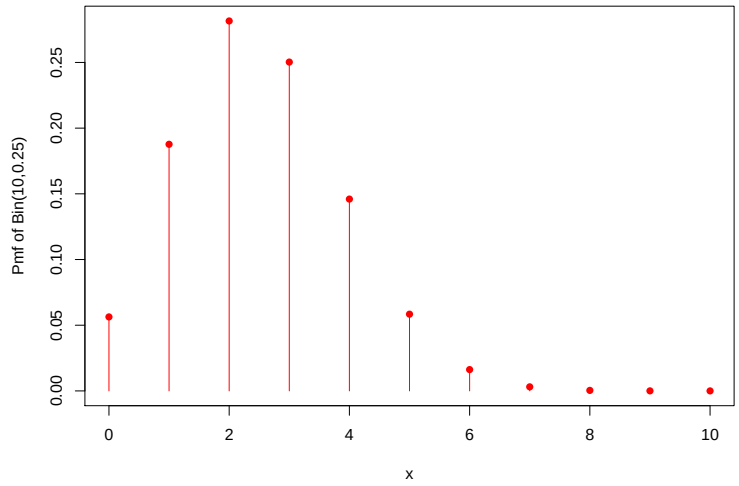
$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

**Definition 7.4.3** (Binomial distribution)**.** *A discrete random variable $X$ is said to follow the* binomial distribution *with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$ if $\mathrm{Im}X = \{0, 1, \ldots, n\}$ and*

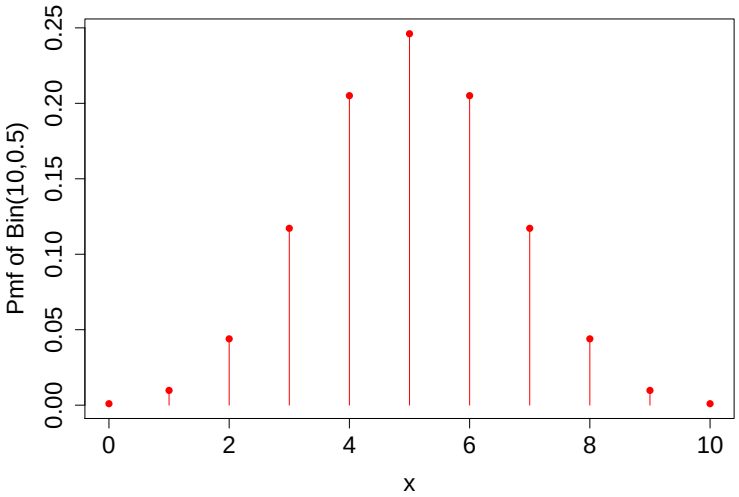
$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \text{ for } x \in \{0, 1, \ldots, n\},$$

*and $\mathrm{P}(X = x) = 0$ otherwise. We write $X \sim \mathrm{Bin}(n, p)$.*

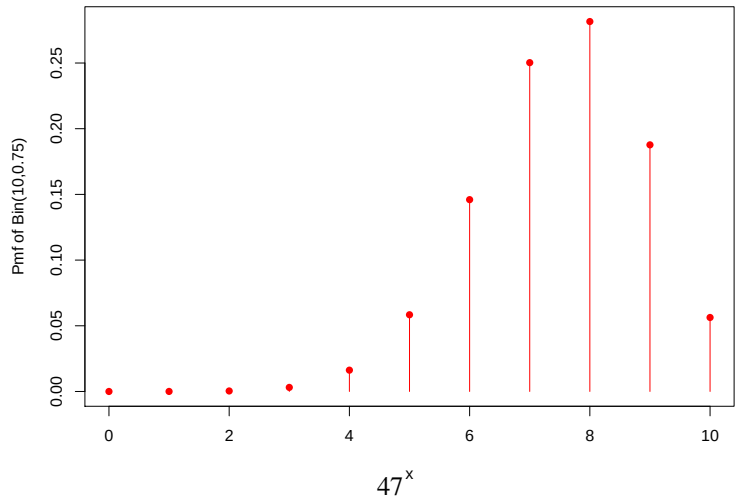
We depict the probability mass function for three random variables with binomial distribution and parameters $n = 10$ and $p \in \{0.25, 0.5, 0.75\}$ in Figure 7.1. We observe that for $p = 0.5$ the pmf is symmetric about $n/2 = 5$ and skewed when $p \neq 0.5$. We will show prove this finding on the problem sheet.




(a) P.m.f. of $X \sim \text{Bin}(10, 0.25)$




(b) P.m.f. of $X \sim \text{Bin}(10, 0.5)$




(c) P.m.f. of $X \sim \text{Bin}(10, 0.75)$

Figure 7.1: We depict the probability mass function for three random variables with binomial distribution and parameters $n = 10$ and $p \in \{0.25, 0.5, 0.75\}$. Note that for $p = 0.5$ the pmf is symmetric about 5 and skewed when $p \neq 0.5$.

### 7.4.3 Hypergeometric distribution

Consider an urn filled with $N$ balls, with $K \in \mathbb{N}$ being white balls and $N - K$ being black. When we draw $n \in N$ balls *with replacement*, we obtain a $\text{Bin}(n, K/N)$ distribution for the number of white balls drawn. Suppose now we draw *without replacement*, then the number of white balls follows the so-called *hypergeometric distribution*.

**Definition 7.4.4** (Hypergeometric distribution)**.** *A discrete random variable $X$ is said to follow the* hypergeometric distribution *with the three parameters $N \in \mathbb{N} \cup \{0\}$, $K, n \in \{0, 1, \ldots, N\}$ if $\text{Im}X = \{0, 1, \ldots, \min(n, K)\}$ and*

$$\text{P}(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \text{ for } x \in \{0, 1, \ldots, K\} \text{ and } n - x \in \{0, 1, \ldots, N - K\},$$

*and $\text{P}(X = x) = 0$ otherwise. We write $X \sim \text{HGeom}(N, K, n)$.*

**Remark 7.4.5.** *We think of $N$ as the size of the population, $K$ the number of success states in the population (e.g. number of white balls), $n$ the number of draws and $x$ is the number of observed successes.*

In Figure 7.2 we show how the pmf of the hypergeometric distributions with $N = 500$ and $K = 200$ shifts when we increase the number of draws from $n = 10$ to $30$ and then $50$.

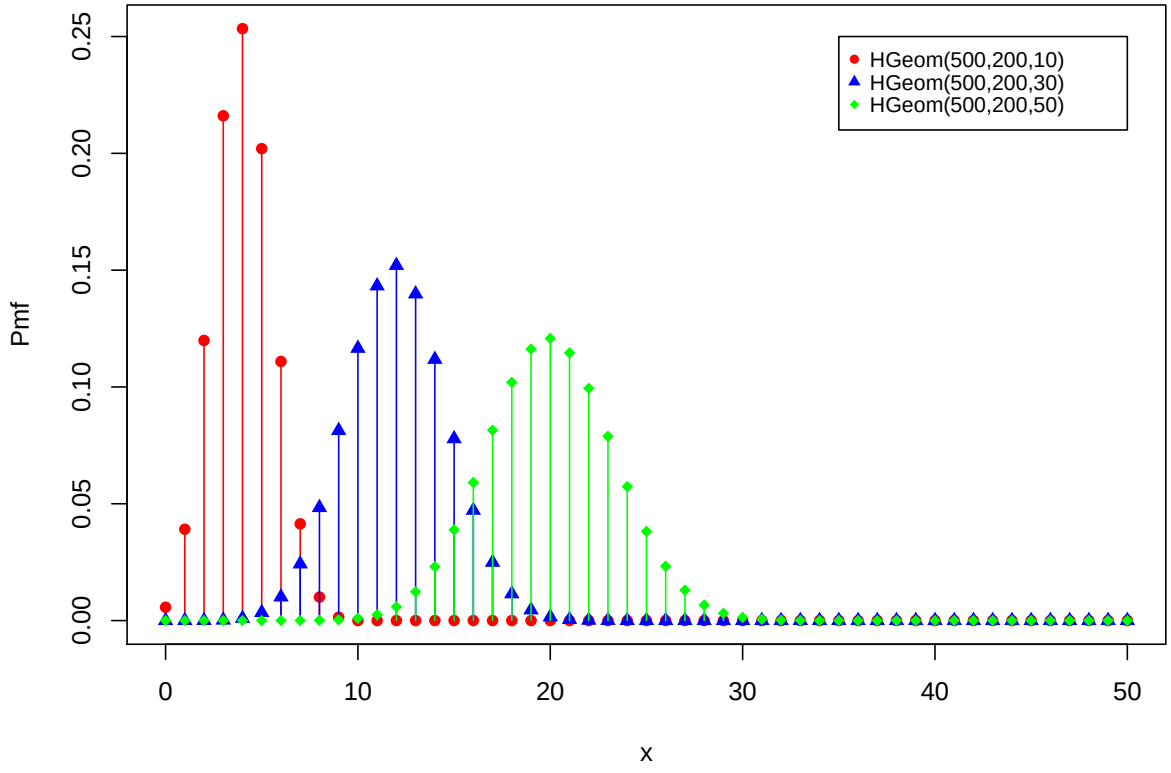



Figure 7.2: This graph shows the probability mass function of the hypergeometric distribution with parameters $N = 500$, $K = 200$ and $n \in \{10, 30, 50\}$.

If we would like to show that the probability mass function of the hypergeometric distribution is indeed a valid probability mass function, we typically use the Vandermonde's identity which we will study next.

**Lemma 7.4.6** (Vandermonde's identity). *For $k, n, m \in \mathbb{N} \cup \{0\}, k \leq n+m$, we have*

$$\binom{m+n}{k} = \sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i}.$$

**Remark 7.4.7.** *Note that we use the convention that, for $n, m \in \mathbb{N}_0$, we set $\binom{m}{n} = 0$ if $n > m$.*

*Proof.* **Story proof/Combinatorial proof:**
Consider selecting a committee of $k$ people from a group of people consisting of $m$ men and $n$ women. The left hand side describes the number of possibilities of selecting $k$ from $m+n$ people (without replacement, order irrelevant). On the right hand side we consider all possible combinations when we choose $i$ men out of $m$ men, then we need to choose $k-i$ women out of $n$ women to obtain a committee of $k$ people. We then need to sum of all possible values of $i$ which gives us the right hand side.
**Algebraic proof:**
Using the binomial theorem, we get

$$(1+x)^{m+n} = \sum_{k=0}^{m+n} \binom{m+n}{k} x^k,$$

and also $(1+x)^{m+n} = (1+x)^m(1+x)^n$ with

$$(1+x)^m(1+x)^n = \sum_{i=0}^{m}\binom{m}{i}x^i \sum_{j=0}^{n}\binom{n}{j}x^j = \sum_{i=0}^{m}\sum_{j=0}^{n}\binom{m}{i}\binom{n}{j}x^{i+j}.$$

Next, we change the summation indices: We have that $0 \leq i \leq m, 0 \leq j \leq n$. We set $k = i+j$, then $0 \leq k \leq m+n, 0 \leq i \leq k$. Hence

$$(1+x)^{m+n} = \sum_{i=0}^{m}\sum_{j=0}^{n}\binom{m}{i}\binom{n}{j}x^{i+j} = \sum_{k=0}^{m+n}\sum_{i=0}^{k}\binom{m}{i}\binom{n}{k-i}x^k$$

So, overall, we found that

$$(1+x)^{m+n} = \sum_{k=0}^{m+n}\binom{m+n}{k}x^k = \sum_{k=0}^{m+n}\sum_{i=0}^{k}\binom{m}{i}\binom{n}{k-i}x^k.$$

We note that two polynomials are identical if they have the same degree and the corresponding coefficients are identical, which implies that for all $0 \leq k \leq n+m$, we have

$$\binom{m+n}{k} = \sum_{i=0}^{k}\binom{m}{i}\binom{n}{k-i}.$$

□

### 7.4.4 Discrete uniform distribution

**Definition 7.4.8** (Discrete uniform distribution). *Let $C$ denote a finite nonempty set of numbers. We say that a discrete random variable $X$ follows the* discrete uniform distribution on $C$, *i.e. $X \sim \mathrm{DUnif}(C)$, if $\mathrm{Im}X = C$ and*

$$\mathrm{P}(X = x) = \frac{1}{\mathrm{card}(C)},$$

*for $x \in C$ and $\mathrm{P}(X = x) = 0$ otherwise.*

**Example 7.4.9.** *Let* $C = \{1, \ldots, n\}$. *If* $X \sim \mathrm{DUnif}(C)$, *then* $\mathrm{P}(X = x) = 1/n$ *for all* $x \in \{1, \ldots, n\}$ *and* $0$ *otherwise.*

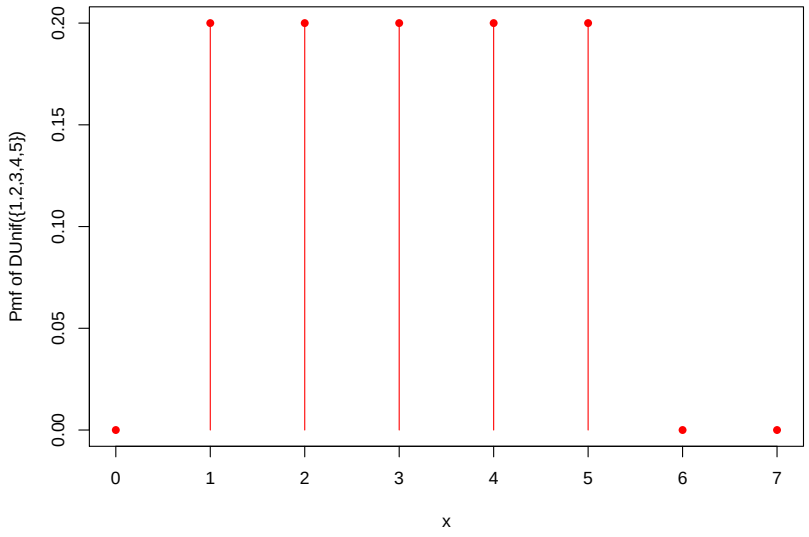



Figure 7.3: This graph shows the probability mass function of the discrete uniform distribution on the set $C = \{1, 2, 3, 4, 5\}$.

### 7.4.5 Poisson distribution

We will now introduce the Poisson distribution which is widely used for counting the number of events/-successes in a certain time period, e.g. the number of earthquakes in some region in the world.

**Definition 7.4.10** (Poisson distribution)**.** *A discrete random variable* $X$ *is said to follow the* Poisson distribution *with parameter* $\lambda > 0$, *i.e.* $X \sim \mathrm{Poi}(\lambda)$, *if* $\mathrm{Im}X = \{0, 1, 2, \ldots\} = \mathbb{N} \cup \{0\}$ *and*

$$\mathrm{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad \textit{for } x = 0, 1, 2, \ldots.$$

We typically call the parameter $\lambda$ in the Poisson distribution the *rate* or *intensity* [of the occurrence of (rare) events].

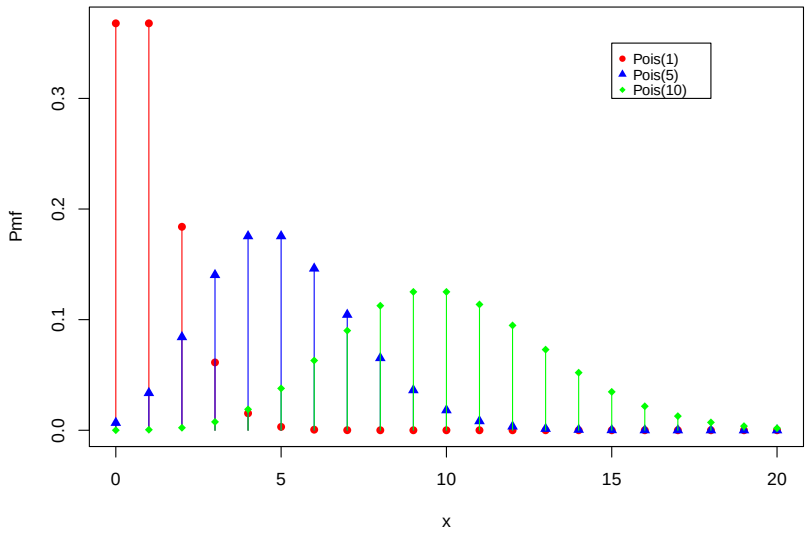



Figure 7.4: This graph shows the probability mass function of the Poisson distribution with three different rate parameters: $\lambda \in \{1, 5, 10\}$.

### 7.4.6 Geometric distribution

**Definition 7.4.11** (Geometric distribution)**.** *A discrete random variable $X$ is said to follow the* geometric distribution *with parameter $p \in (0, 1)$, i.e. $X \sim \text{Geom}(p)$ if $\text{Im}X = \mathbb{N}$ and*

$$\text{P}(X = x) = (1 - p)^{x-1}p, \quad \textit{for } x = 1, 2, \ldots.$$

We can think of an experiment where we carry out repeated (independent) Bernoulli trials with success probability $p$. We stop the experiment after the first success. We denote by $X$ the number of trials to obtain the first success. Then we obtain that $X \sim \text{Geom}(p)$. **Warning:** If we set $Y$ to be the *number of failures until first success* we obtain a slightly different definition of the geometric distribution. Here we have that $\text{Im}Y = \mathbb{N} \cup \{0\}$ and

$$\text{P}(Y = x) = (1 - p)^{x}p, \quad \text{for } x = 0, 1, 2, \ldots.$$

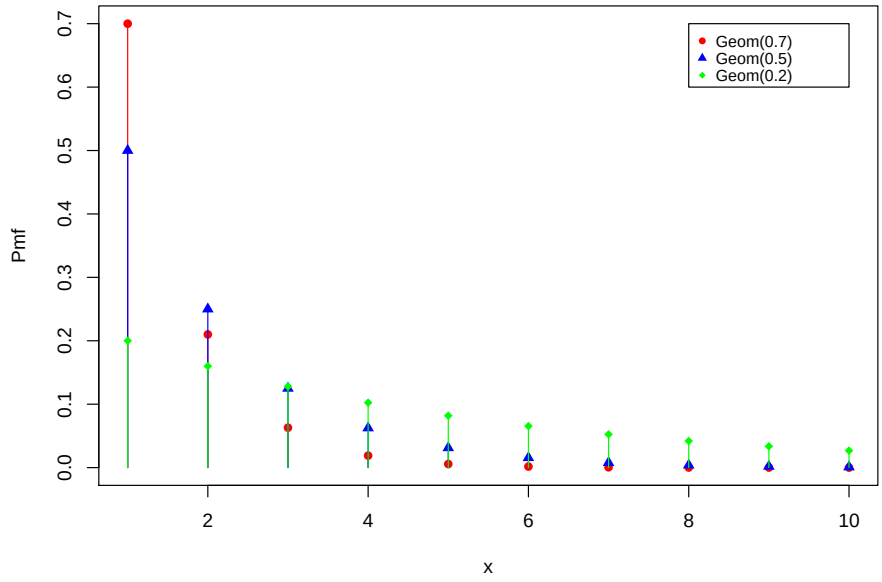



Figure 7.5: This graph shows the probability mass function of the Geometric distribution with three different success probabilities: $p \in \{0.7, 0.5, 0.2\}$.

### 7.4.7 Negative binomial distribution

**Definition 7.4.12** (Negative binomial distribution)**.** *A discrete random variable $X$ is said to follow the* negative binomial distribution *with parameters $r \in \mathbb{N}$ and $p \in (0,1)$, written $X \sim \mathrm{NBin}(r,p)$, if $\mathrm{Im}X = \mathbb{N} \cup \{0\}$ and*

$$\mathrm{P}(X = x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad \text{for } x = 0, 1, \ldots. \tag{7.4.1}$$

The negative binomial distribution arises as the distribution of the number of failures in a sequence of independent Bernoulli trials with success parameter $p$ before $r$ successes have occurred. To see this, let us consider strings of "0" (for failure) and "1" (for success). Each string of $r$ "1"s and $x$ "0"s has probability $p^r(1-p)^x$. Now we need to find the number of such strings: We stop when we reach the $r$th success, so the last element in the string will always be a "1". This leaves us with $r+x-1$ positions, to which we need to assign the remaining $r-1$ "1"s. Hence we obtain equation (7.4.1).

**Remark 7.4.13.** *Recall that in the case of a $\mathrm{Bin}(n,p)$ distribution, we also consider a sequence of independent Bernoulli trials, but we fix the number of trials $n$ and count the number of successes.*

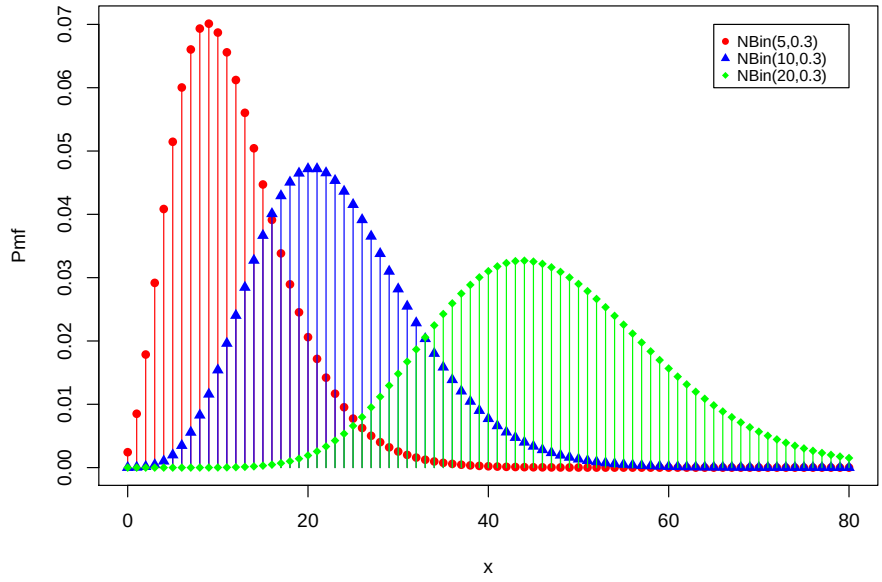



Figure 7.6: This graph shows the probability mass function of the negative binomial distribution with parameter $p = 0.3$ and $r \in \{5, 10, 20\}$.

In order to show that the pmf of the negative binomial distribution is a valid pmf, we study the generalisation of the binomial coefficient:

**Definition 7.4.14.** *For $\alpha \in \mathbb{C}, k \in \mathbb{N}$, we define*

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}.$$

The generalised binomial formula is then given by

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k, \text{ for } |x| < 1.$$

**Lemma 7.4.15.** *For $x \in \mathbb{N} \cup \{0\}, r \in \mathbb{N}$ we have the following identity*

$$\binom{x+r-1}{r-1} = (-1)^x \binom{-r}{x}.$$

*Proof.* To see this, note that

$$\begin{aligned}\binom{x+r-1}{r-1} &= \frac{(x+r-1)!}{x!(r-1)!} = \frac{(x+r-1)(x+r-2)\cdots r}{x!} \\ &= (-1)^x \frac{(-r)(-r-1)\cdots(-r-x+1)}{x!} = (-1)^x \binom{-r}{x}.\end{aligned}$$

□

Lemma 7.4.15 together with the generalised Binomial formula stated above can be used to show that if $X \sim \mathrm{NBin}(r,p)$ for $p \in (0,1)$, then

$$\sum_{x=0}^{\infty} \mathrm{P}(X=x) = \sum_{x=0}^{\infty} \binom{x+r-1}{r-1} p^r (1-p)^x = p^r \sum_{x=0}^{\infty} (-1)^x \binom{-r}{x} (1-p)^x$$
$$= p^r \sum_{x=0}^{\infty} \binom{-r}{x} (p-1)^x = p^r (1+(p-1))^{-r} = p^r p^{-r} = 1,$$

where the generalised Binomial theorem was applicable since $|1-p| = (1-p) < 1$.

### 7.4.8 Exercise

**Exercise 7.4.16.** *Verify that all the probability mass functions listed above are valid in the sense that* $p_X(x) \geq 0$ *for all* $x$ *and* $\sum_x p_X(x) = 1$.

**End of lecture 10.**

# Chapter 8

# Continuous random variables

*The material of this chapter is based on Blitzstein & Hwang (2019), p.121-123, 213-244, Grimmett & Welsh (1986), p.56-65.*

## 8.1 Random variables and their distributions

So far, we have only considered discrete random variables which can take at most countably many values. Now we will give a more general definition which is also suitable for broader applications.

**Definition 8.1.1** (Random variable)**.** *A* random variable *on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is defined as the mapping $X : \Omega \to \mathbb{R}$ which satisfies*

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \qquad \textit{for all } x \in \mathbb{R}. \tag{8.1.1}$$

Note that a discrete random variable (see Definition 7.3.1) satisfies the above definition of a random variable. To see this, we can write

$$\{\omega \in \Omega : X(\omega) \leq x\} = \bigcup_{y \in \mathrm{Im}X : y \leq x} \{\omega : X(\omega) = y\}.$$

The right hand side is a countable union of elements of $\mathcal{F}$ and (according the definition of the sigma-algebra) hence also an element of $\mathcal{F}$.

**Remark 8.1.2.** *Note that, similarly to our previous definition, we call the set $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\}$ the pre-image of $(-\infty, x]$. We can only make probability statements about the set $X^{-1}((-\infty, x])$ if it is an element of the event space $\mathcal{F}$ which motivates our definition of a random variable.*

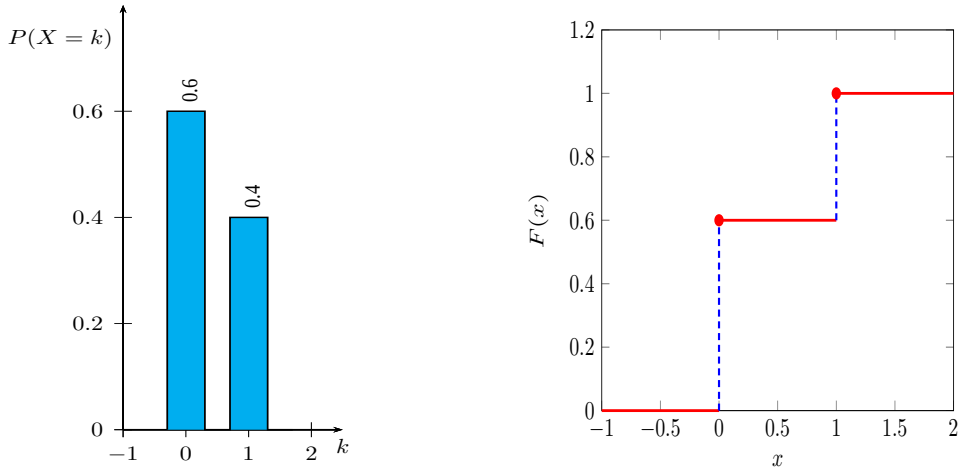
**Definition 8.1.3** (Cumulative distribution function (c.d.f.))**.** *Suppose that $X$ is a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$, then the* cumulative distribution function *(c.d.f.) of $X$ is defined as the mapping $F_X : \mathbb{R} \to [0, 1]$ given by*

$$F_X(x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathrm{P}(X^{-1}((-\infty, x])),$$

*which is typically abbreviated to $F_X(x) = \mathrm{P}(X \leq x)$.*

**Example 8.1.4.** *Consider a Bernoulli random variable $X \sim \mathrm{Bern}(p)$. The c.d.f. of a Bernoulli random variable is given by*

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{k \leq x} \mathrm{P}(X = k) = \begin{cases} 0, & \textit{for } x < 0, \\ 1 - p, & \textit{for } x \in [0, 1), \\ 1, & \textit{for } x \geq 1. \end{cases}$$




(a) P.m.f. of $X \sim \text{Bern}(0.4)$

(b) C.d.f. of $X \sim \text{Bern}(0.4)$

Figure 8.1: Consider a Bernoulli random variable $X$ with parameter $p = 0.4$. Its probability mass function is depicted in Figure 8.1a and its cumulative distribution function in Figure 8.1b.

Let us now derive important properties of the c.d.f..

**Theorem 8.1.5** (Properties of the c.d.f.). *1. $F_X$ is monotonically non-decreasing, i.e. for all $x \leq y$ we have $F_X(x) \leq F_X(y)$.*

*2. $F_X$ is right-continuous, i.e. if $x_n \downarrow x$ (which is short-hand notation for a sequence $(x_n)_{n\in\mathbb{N}}$, which is monotonically non-increasing, i.e. $x_1 \geq \cdots \geq x_n \geq x_{n+1} \geq \cdots \geq x$ and converging to $x$, i.e. $\lim_{n\to\infty} x_n = x$), then $F_X(x_n) \to F_X(x)$ as $n \to \infty$.*

*3. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.*

*Proof.* 1. Monotonicity: Let $x \leq y$. Then

$$\{\omega \in \Omega : X(\omega) \leq x\} \subseteq \{\omega \in \Omega : X(\omega) \leq y\}.$$

Then the result follows from the monotonicity of the probability measure, see the second statement in Theorem 4.2.3.

2. Right continuity: We prove that if $x_n \downarrow x$ (which is short-hand notation for a sequence $(x_n)_{n\in\mathbb{N}}$, which is monotonically non-increasing, i.e. $x_1 \geq \cdots \geq x_n \geq x_{n+1} \geq \cdots \geq x$ and converging to $x$, i.e. $\lim_{n\to\infty} x_n = x$), then $F_X(x_n) \to F_X(x)$ as $n \to \infty$. Define events

$$E_n := \{\omega : X(\omega) \leq x_n\} \downarrow \bigcap_{n=1}^{\infty} E_n = \{\omega : X(\omega) \leq x\} =: E.$$

Using the continuity of the probability measure, see Theorem 6.1.16, $\mathrm{P}(E_n) \to \mathrm{P}(E)$. Since $\mathrm{P}(E_n) = F_X(x_n), \mathrm{P}(E) = F_X(x)$, we have that $F_X(x_n) \to F_X(x)$ as $n \to \infty$.

3. Limit behaviour at $\pm\infty$: Define for an $x_n$

$$E_n = \{\omega : X(\omega) \leq x_n\}.$$

Then $E_n \downarrow \emptyset$ as $x_n \downarrow -\infty$ and $E_n \uparrow \Omega$ as $x_n \uparrow \infty$. Using the continuity property of $\mathrm{P}$ again, we deduce that $F_X(x_n) = \mathrm{P}(E_n) \to \mathrm{P}(\emptyset) = 0$ as $x_n \to -\infty$, and $F_X(x_n) = \mathrm{P}(E_n) \to \mathrm{P}(\Omega) = 1$ as $x_n \to \infty$.

□

**Remark 8.1.6.** *One can show that for any function $F$ which satisfies the three conditions stated in Theorem 8.1.5, there exists a probability space and a random variable on that space which has $F$ as its c.d.f..*

Note that in applications, we often use the following result:

**Theorem 8.1.7.** *For $a < b$, we have $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.*

*Proof.* Note that for $a < b$, we have

$$\{\omega \in \Omega : X(\omega) \leq b\} = \{\omega \in \Omega : X(\omega) \leq a\} \cup \{\omega \in \Omega : a < X(\omega) \leq b\},$$

where the two events on the right hand side are disjoint. Hence

$$\mathrm{P}(\{\omega \in \Omega : X(\omega) \leq b\}) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq a\}) + \mathrm{P}(\{\omega \in \Omega : a < X(\omega) \leq b\}),$$

which implies that

$$\mathrm{P}(\{\omega \in \Omega : a < X(\omega) \leq b\}) = F_X(b) - F_X(a).$$

□

**Remark 8.1.8.** *A cumulative distribution function (c.d.f) of a random variable $X$, say, is right continuous, but not in general left continuous. To see the latter, consider a point $x \in \mathbb{R}$ and an arbitrary sequence $(x_n)_{n\in\mathbb{N}}$ approaching $x$ from the left, i.e. $x_1 \leq x_2 \leq \cdots \leq x_n \leq \cdots \leq x$ and $\lim_{n\to\infty} x_n = x$. Then*

$$E_n := \{\omega \in \Omega : X(\omega) \leq x_n\} \uparrow \bigcup_{n=1}^{\infty} E_n = \{\omega \in \Omega : X(\omega) < x\} =: E.$$

*Hence, by the continuity of the probability measure*

$$\begin{aligned}\lim_{n\to\infty} F_X(x_n) &= \lim_{n\to\infty} \mathrm{P}(E_n) = \mathrm{P}(E)\\ &= \mathrm{P}(\{\omega \in \Omega : X(\omega) < x\})\\ &= \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\}) - \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\})\\ &\leq \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\} = F_X(x).\end{aligned}$$

*So, we observe that $F_X$ is (left) continuous in $x$ if and only if $\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) = 0$.*

**End of lecture 11.** .....

## 8.2 Continuous random variables and probability density function

When looking at the c.d.f. of a Bernoulli random variable, see Figure 8.1b, we noted that the c.d.f. looks like a step function. Indeed, all discrete random variables have c.d.f.s which are right-continuous step functions (with possibly (many) more steps than in the Bernoulli case). In the remainder of this chapter we will now focus on random variables with a smooth c.d.f.:

**Definition 8.2.1** (Continuous random variable and probability density function)**.** *A random variable $X$ is called* continuous *if its c.d.f. can be written as*

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du, \qquad \text{for all } x \in \mathbb{R}, \tag{8.2.1}$$

*where the function $f_X : \mathbb{R} \to \mathbb{R}$ satisfies*

*(i) $f_X(u) \geq 0$ for all $u \in \mathbb{R}$,*

*(ii)* $\int_{-\infty}^{\infty} f_X(u)du = 1$.

*We call* $f_X$ *the* probability density function *(p.d.f.) of* $X$ *(or just the* density*).*[1]

The so-called Fundamental Theorem of Calculus guarantees that a function $F_X$ given as in Definition 8.2.1 is differentiable at every point $x$ where $f$ is continuous with $F_X'(x) = f_X(x)$.

**Remark 8.2.2.** *Note that* $f_X(x)$ *is* not *a probability and while* $f_X$ *is non-negative it is not restricted to be smaller than 1.*

We compare properties of the p.m.f. and the p.d.f. in the following table:

| Discrete random variable | Continuous random variable |
|---|---|
| $p_X(x) \geq 0$, for all $x \in \mathbb{R}$ | $f_X(x) \geq 0$, for all $x \in \mathbb{R}$ |
| $\sum_{x \in \mathrm{Im}X} p_X(x) = 1$ | $\int_{-\infty}^{\infty} f_X(x)dx = 1$ |
| $F_X(x) = \sum_{u \in \mathrm{Im}X : u \leq x} p_X(u)$ | $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ |

Table 8.1: Comparing discrete and random variables with p.m.f. $p_X$ and p.d.f. $f_X$, respectively.

It turns out that, although $f_X(x)$ is not a probability, it can be linked to a probability when we scale it appropriately. Consider a small quantity which we shall denote by $dx > 0$. Then the probability that $X$ is close to $x$ can be written as

$$\mathrm{P}(x < X \leq x + dx) = F_X(x + dx) - F_X(x) = \int_x^{x+dx} f_X(u)du \approx f_X(x)dx.$$

So, we can view the quantity $f_X(x)dx$ as the continuous analogue to a probability mass function $p_X(x)$.

The reason why we typically do not consider point probabilities for continuous random variables becomes clear in the next theorem.

**Theorem 8.2.3.** *For a continuous random variable* $X$ *with density* $f_X$*, we have*

$$\mathrm{P}(X = x) = 0, \qquad \textit{for all } x \in \mathbb{R}, \tag{8.2.2}$$

*and*

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(u)du, \qquad \textit{for all } a, b \in \mathbb{R} \textit{ with } a \leq b. \tag{8.2.3}$$

*Proof.* Consider any $x \in \mathbb{R}$ with a sequence $x_n \uparrow x$ (i.e.$(x_n)_{n\in\mathbb{N}}$, with $\lim_{n\to\infty} x_n = x$ and $x_1 \leq x_2 \leq \cdots \leq x_n \leq \cdots \leq x$), and define events

$$E_n = \{\omega : x_n < X(\omega) \leq x\} \downarrow E = \{\omega : X(\omega) = x\}.$$

Using the continuity of the probability measure, see Theorem 6.1.16, $\mathrm{P}(E_n) \to \mathrm{P}(E)$. Hence we can write

$$\begin{aligned}
\mathrm{P}(X = x) &= \lim_{n\to\infty} \mathrm{P}(E_n) = \lim_{n\to\infty} \mathrm{P}(\{\omega : x_n < X(\omega) \leq x\}) \\
&= \lim_{n\to\infty} (F_X(x) - F_X(x_n)) \\
&= \lim_{n\to\infty} \int_{x_n}^{x} f_X(u)du = 0.
\end{aligned}$$

Now, let $a \leq b$, then we know from the above that $\mathrm{P}(X = a) = 0$, hence

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = F_X(b) - F_X(a),$$

where we used Theorem 8.1.7.

□

[1] In a later analysis/measure course we will say that equation 8.2.1 means that the "c.d.f. of a continuous random variable is *absolutely continuous* with respect to the Lebesgue measure".

**Remark 8.2.4.** *Combining the results from Remark 8.1.8 and Theorem 8.2.3 we conclude that the c.d.f. of a continuous random variable is continuous. It is important to remember that the definition of the continuous random variable guarantees the existence of the density and then the continuity of the associated c.d.f. follows from the properties of the (improper) Riemann integral. Note that if you only assumed that a random variable $X$ has a continuous c.d.f. with, in particular, $\mathrm{P}(X = x) = 0$ for all $x$, then the existence of a density function is not guaranteed. A notorious example of such a case is the so-called* Cantor function *which you might study in a later analysis/measure course and another example we will discuss later, see Example 8.4.1.*

## 8.3 Common continuous distributions [Reading material]

### 8.3.1 Uniform

**Definition 8.3.1** (Uniform distribution)**.** *A continuous random variable $X$ is said to have the* uniform distribution *on the interval $(a, b)$ for $a < b$, i.e. $X \sim \mathrm{U}(a, b)$, if its density function is given by*

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

*Its cumulative distribution function is given by*

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq a, \\ \frac{x-a}{b-a}, & \text{if } a < x < b, \\ 1, & \text{if } x \geq b. \end{cases}$$

### 8.3.2 Exponential

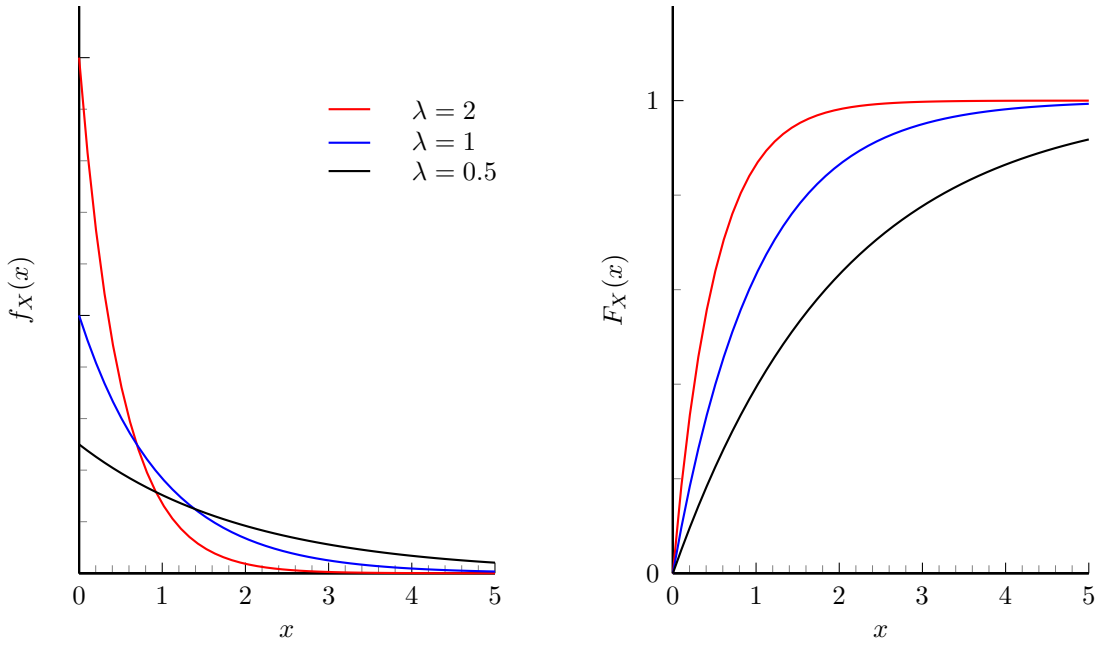



Figure 8.2: Plot of the p.d.f. (left) and the c.d.f. (right) of an $\mathrm{Exp}(\lambda)$ random variable for $\lambda \in \{0.5, 1, 2\}$.

**Definition 8.3.2** (Exponential distribution)**.** *A continuous random variable $X$ is said to have the* exponential distribution *with parameter $\lambda > 0$, i.e. $X \sim \text{Exp}(\lambda)$, if its density function is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

*Its cumulative distribution function is given by*

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - e^{-\lambda x}, & \text{if } x > 0. \end{cases}$$

### 8.3.3 Gamma distribution

The Gamma distribution is – as the exponential distribution – also supported on the positive real line only and extends the exponential distribution discussed above.

For $t > 0$ we define the *Gamma function* by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx,$$

which has the following properties:

$$\Gamma(t) = (t-1)\Gamma(t-1), \text{ for } t > 1,$$

and in the case when $t \in \mathbb{N}$ we have $\Gamma(t) = (t-1)!$.

**Definition 8.3.3** (Gamma distribution)**.** *A continuous random variable $X$ is said to have the* Gamma distribution *with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, i.e. $X \sim \text{Gamma}(\alpha, \beta)$, if its density function is given by*

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

*Its cumulative distribution function is not available in closed form.*

In the case when $\alpha = n \in \mathbb{N}$, we often call the Gamma distribution the *Erlang distribution* which has density

$$f_X(x) = \begin{cases} \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

### 8.3.4 Chi-squared distribution

**Definition 8.3.4** (Chi-squared distribution)**.** *A continuous random variable $X$ is said to have the* chi-squared distribution *with $n \in \mathbb{N}$ degrees of freedom, i.e. $X \sim \chi^2(n)$ (or also $X \sim \chi^2_n$) , if its density function is given by*

$$f_X(x) = \begin{cases} \frac{1}{2\Gamma(n/2)} \left(\frac{x}{2}\right)^{n/2-1} e^{-x/2}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

*Its cumulative distribution function is not available in closed form.*

We note that the $\chi^2(n)$ distribution is the same as the $\text{Gamma}(n/2, 1/2)$ distribution.

### 8.3.5 F-distribution

**Definition 8.3.5** (F-distribution)**.** *A continuous random variable $X$ is said to have the* F-distribution *with $d_1, d_2 > 0$ degrees of freedom, i.e. $X \sim F(d_1, d_2)$ (or also $X \sim F_{d_1,d_2}$) , if its density function is given by*

$$f_X(x) = \begin{cases} \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

*Its cumulative distribution function is not available in closed form.*

We note that the positive parameters $d_1, d_2$ are not restricted to be integer-valued.
Note that if we have independent random variables $X_1 \sim \chi^2_n$ and $X_2 \sim \chi^2_m$, then the random variable

$$X = \frac{X_1/n}{X_2/m} \sim F_{n,m}.$$

### 8.3.6 Beta distribution

For $\alpha, \beta > 0$ denote by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

the so-called Beta function.

**Definition 8.3.6** (Beta distribution)**.** *A continuous random variable $X$ is said to have the* Beta distribution *with parameters $\alpha, \beta > 0$, i.e. $X \sim \text{Beta}(\alpha, \beta)$, if its density function is given by*

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

*Its cumulative distribution function is not available in closed form.*

### 8.3.7 Normal distribution

**Definition 8.3.7** (Standard normal distribution)**.** *A random variable $X$ has the* standard normal/standard Gaussian *distribution if it has density function $f(x) = \phi(x)$ with*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim \text{N}(0, 1)$ since a standard normal random variable has mean zero and variance one. The c.d.f. is then denoted by $F(x) = \Phi(x)$ with

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, \quad \text{for } x \in \mathbb{R}.$$

Unfortunately there is no explicit formula for the integral appearing in the c.d.f.!

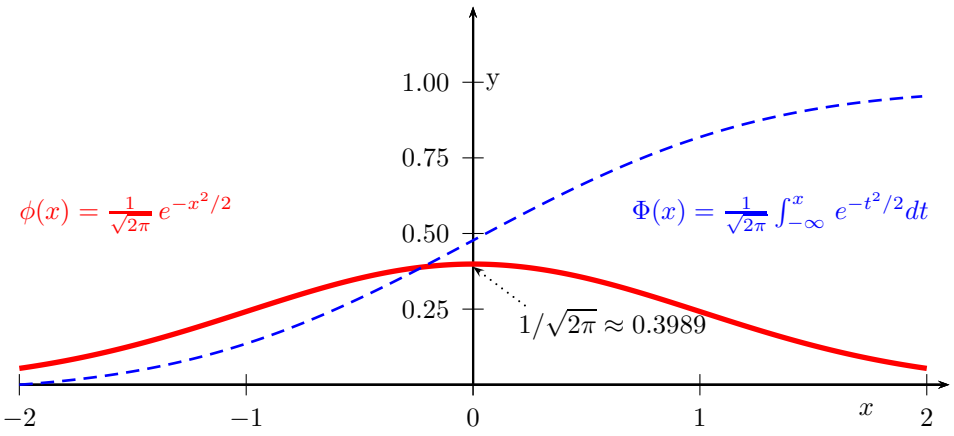



Figure 8.3: The red solid line depicts the standard Gaussian probability density function and the blue dashed line the corresponding cumulative distribution function.

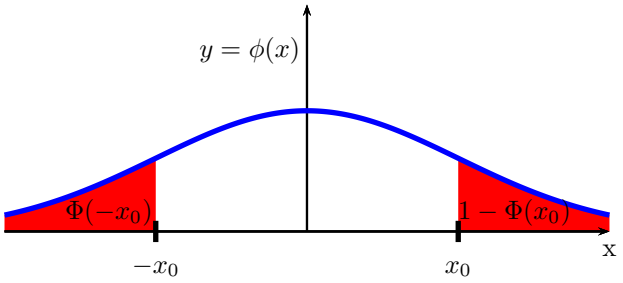



Figure 8.4: Note that the standard normal density is symmetric around $0$, i.e. $\phi(x) = \phi(-x)$ for all $x$. This also implies that $\Phi(-x) = 1 - \Phi(x)$.

**Definition 8.3.8** (Normal distribution)**.** *Let $\mu$ denote a real number and let $\sigma > 0$. A random variable $X$ has the* normal/ Gaussian *distribution with mean $\mu$ and variance $\sigma^2$ if it has density function*

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

Note that we typically write $X \sim \mathrm{N}(\mu, \sigma^2)$.

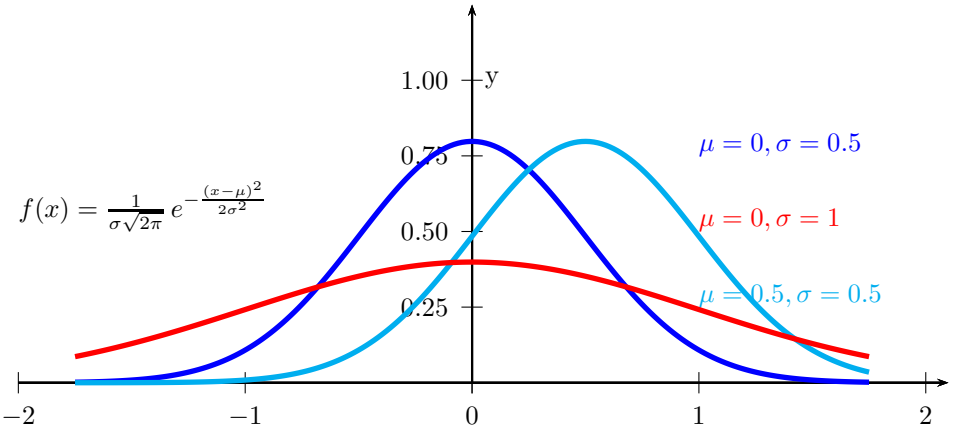



Figure 8.5: The red line depicts the standard Gaussian probability density function and the two blue lines show non-standard Gaussian probability density functions.

### 8.3.8 Cauchy distribution

**Definition 8.3.9** (Cauchy distribution)**.** *A continuous random variable $X$ is said to have the* Cauchy distribution*, if its density function is given by*

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \qquad \text{for } x \in \mathbb{R}.$$

*Its cumulative distribution function is given by*

$$F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}, \qquad \text{for } x \in \mathbb{R}.$$

We note that if we have two independent standard normal random variables $X, Y \sim \mathrm{N}(0,1)$, then their ratio $Z = X/Y$ follows the Cauchy distribution.

### 8.3.9 Student t-distribution

**Definition 8.3.10** ((Student's) t-distribution)**.** *A continuous random variable $X$ is said to have the* (Student's) t-distribution *with $\nu > 0$ degrees of freedom, if its density function is given by*

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad \text{for } x \in \mathbb{R}.$$

*Its cumulative distribution function is not available in closed form.*

## 8.4 Example of a random variable which is neither discrete nor continuous

Throughout this course we will typically focus on random variable which are either discrete or continuous. However, it is important to note that random variables exist which are neither discrete nor continuous! Let us study such an example in the following.

**Example 8.4.1.** *We flip an unfair coin infinitely many times and assume that we obtain heads with probability $p \in (0,1)$. We denote the outcomes by $X_1, X_2, \ldots$ with $X_i = 0$ if we obtain tails in the $i$th flip and $X_i = 1$ if we obtain heads in the $i$th flip. Define a random number*

$$Y = 0.X_1X_2X_3\ldots \qquad \text{(in base 2), i.e.}$$
$$Y = X_1 \cdot \frac{1}{2} + X_2 \left(\frac{1}{2}\right)^2 + X_3 \cdot \left(\frac{1}{2}\right)^3 + \cdots .$$

*Then*

- *$X_1$ determines whether $Y$ is in the first half $[0, \frac{1}{2})$ (if $X_1 = 0$) or in the second half $[\frac{1}{2}, 1]$ (if $X_1 = 1$).*
- *$X_2$ determines whether $Y$ is in the first half (if $X_2 = 0$) (either in $[0, \frac{1}{4})$ or in $[\frac{1}{2}, \frac{3}{4})$) or in the second half (if $X_2 = 1$) (either in $[\frac{1}{4}, \frac{1}{2})$ or in $[\frac{3}{4}, 1]$) of the previous half.*
- *etc.*

*We note that for any $y = 0.x_1x_2x_3\ldots$ in base 2 representation, we have*

$$\mathrm{P}(Y = y) = \mathrm{P}(X_1 = x_1)\mathrm{P}(X_2 = x_2)\cdots = 0$$

*since each term in the product is either equal to $p$ or $1-p$ which are both smaller than 1, so their infinite product will converge to 0. Hence $Y$ cannot be a discrete random variable.*

*One (not we!) can show that for $p \neq 0.5$, $Y$ does not have a density, whereas if $p = 0.5$, then $Y$ is uniformly distributed on $[0,1]$ and hence has a density.*

**End of lecture 12.** . . . . . .

# Chapter 9

# Transformations of random variables

*The material of this chapter is based on Blitzstein & Hwang (2019), p.123-129, Grimmett & Welsh (1986), p.28-29, 65-67.*

Let us consider a random variable $X$ on $(\Omega, \mathcal{F}, \mathrm{P})$ and a (deterministic) function $g : \mathbb{R} \to \mathbb{R}$. Clearly $Y = g(X)$ is a mapping from $\Omega$ to $\mathbb{R}$ with $Y(\omega) = g(X(\omega))$. In this chapter, we would like to study under which conditions $Y$ is itself a random variable and we would like to study its distribution

## 9.1 The discrete case

Let us first consider the case when $X$ is a discrete random variable.

**Theorem 9.1.1.** *Let $X$ be a discrete random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and let $g : \mathbb{R} \to \mathbb{R}$ denote a deterministic function. Then $Y = g(X)$ is a discrete random variable with probability mass function given by*

$$p_Y(y) = \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x), \tag{9.1.1}$$

*for all $y \in \mathrm{Im}Y$ and 0 otherwise.*

*Proof.* We observe that $Y : \Omega \to \mathbb{R}$ and that $\mathrm{Im}(Y) = \{g(X(\omega)) : \omega \in \Omega\}$ is countable since $\mathrm{Im}(X) = \{X(\omega) : \omega \in \Omega\}$ is countable. Moreover, for all $y \in \mathbb{R}$, we have

$$\begin{aligned}
Y^{-1}(\{y\}) &= \{\omega \in \Omega : Y(\omega) = y\} = \{\omega \in \Omega : g(X(\omega)) = y\} \\
&= \{\omega \in \Omega : X(\omega) \in \{x \in \mathrm{Im}X : g(x) = y\}\} \\
&= \{\omega \in \Omega : X(\omega) \in \bigcup_{x \in \mathrm{Im}X : g(x) = y} \{x\}\} \\
&= X^{-1}\left( \bigcup_{x \in \mathrm{Im}X : g(x) = y} \{x\} \right) \\
&\overset{Lemma 7.1.2}{=} \bigcup_{x \in \mathrm{Im}X : g(x) = y} X^{-1}(\{x\}) \\
&= \bigcup_{x \in \mathrm{Im}X : g(x) = y} \{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F},
\end{aligned} \tag{9.1.2}$$

since each event $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$ and, by the definition of a $\sigma$-algebra, a countable union of elements of $\mathcal{F}$ is in $\mathcal{F}$, too. Since $X$ is discrete, we indeed have that $\{x \in \mathrm{Im}X : g(x) = y\} \subseteq \mathrm{Im}X$ is (at most) countably infinite. Hence we can conclude that $Y = g(X)$ is indeed a discrete random variable.

We can compute the p.m.f. of $Y$ as follows:

$$p_Y(y) = \mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x),$$

so we are just summing up the probabilities for all values $x$ for which $g(x) = y$. Here we used the fact that the union in (9.1.2) is a countable union of disjoint events, hence Axiom (iii) of the definition of the probability measure applies, see also the discussion before Example 7.3.5 for more details.

□

In the special case when $g$ is invertible (i.e. bijective), then the pmf of $Y$ can be expressed as

$$p_Y(y) = p_X(g^{-1}(y)) \qquad \text{for all } y \in \mathrm{Im}Y.$$

## 9.2 The continuous case

For the continuous (or more general case) recall that $Y = g(X)$ is only a random variable if $Y$ satisfies condition (8.1.1), i.e.

$$\{\omega \in \Omega : Y(\omega) \leq y\} \in \mathcal{F} \qquad \text{for all } y \in \mathbb{R}.$$

This condition is only satisfied if $g$ satisfies some additional properties (e.g. if it is continuous or monotone[1]).

**Example 9.2.1.** *Consider a linear transformation of the random variable $X$. I.e. let $a > 0, b \in \mathbb{R}$ and define $g(x) = ax + b$. Then $Y = g(X) = aX + b$ is indeed a random variable and, for any $y \in \mathbb{R}$ its c.d.f. is given by*

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

*Assuming that $X$ is a continuous random variable, we can now differentiate (with respect to $y$– using the chain rule) and obtain*

$$f_Y(y) = F_Y'(y) = F_X'\left(\frac{y-b}{a}\right)\frac{1}{a} = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

In the previous example, we have seen that in the case that the function $g$ can be inverted, we can find an explicit formula for the corresponding density of the transformed random variable. We can now state and prove this result in a more general form.

**Theorem 9.2.2.** *Suppose that $X$ is a continuous random variable with density $f_X$ and $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing/decreasing and differentiable with inverse function denoted by $g^{-1}$, then $Y = g(X)$ has density*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}[g^{-1}(y)] \right|, \qquad \textit{for all } y \in \mathbb{R}. \tag{9.2.1}$$

*Proof.* First, suppose that $g$ is strictly increasing. Then, for any $y \in \mathbb{R}$, we have

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Differentiating w.r.t. $y$ and an application of the chain rule leads to

$$f_Y(y) = F_Y'(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy}[g^{-1}(y)].$$

[1] More generally, we will need that $g$ is Borel-measurable, but this concept is beyond the scope of this course.

Second, suppose that $g$ is strictly decreasing. Then, for any $y \in \mathbb{R}$, we have

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)),$$

and hence

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)] = f_X(g^{-1}(y))\left|\frac{d}{dy}[g^{-1}(y)]\right|,$$

since in this case $\frac{d}{dy}[g^{-1}(y)] < 0$.

□

**Remark 9.2.3.** *In the proof above, we used the following result from Analysis: If $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing (decreasing) function, then its inverse function denoted by $g^{-1}$ is also strictly increasing (decreasing).*

**Remark 9.2.4.** *We typically call the term $\left|\frac{d}{dy}[g^{-1}(y)]\right|$ the* Jacobian of the transformation.

**Remark 9.2.5.** *You might remember equation* (9.2.1) *more easily when you write $x = g^{-1}(y)$ and note that*

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

*In the case when $g$ is strictly increasing, you can remove the absolute value signs and you get the pretty symmetric formula:*

$$f_Y(y)dy = f_X(x)dx.$$

If $X$ is continuous and we want to find the c.d.f. and/or the p.d.f. of $Y = g(X)$ in the case when $g$ is not necessarily strictly increasing/decreasing, then we compute

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \in \{x \in \mathrm{Im}X : g(x) \leq y\}) \\ &= \int_{\{x \in \mathrm{Im}X : g(x) \leq y\}} f_X(x)dx. \end{aligned}$$

If $Y$ is continuous, we would then differentiate its c.d.f. to obtain its p.d.f..

**Example 9.2.6.** *Let $X \sim \mathrm{N}(0,1)$, $g(x) = x^2$ and set $Y = g(X) = X^2$. We would like to find the c.d.f. and the p.d.f. of $Y$. First we compute the c.d.f. of $Y$. Clearly, for $y < 0$, we have $F_Y(y) = \mathrm{P}(Y \leq y) = 0$ and hence $f_Y(y) = 0$. Now let $y \geq 0$, then*

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(X^2 \leq y) = \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}). \end{aligned}$$

*Differentiating leads to*

$$\begin{aligned} f_Y(y) &= \frac{1}{2}y^{-1/2}f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-1/2}\right)f_X(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}}\left[\phi(\sqrt{y}) + \phi(-\sqrt{y})\right] \\ &= \frac{1}{2\sqrt{y}}\left[2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y\right)\right] = \frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{1}{2}y\right), \end{aligned}$$

*which is the density of a $\chi^2_1$-random variable.*

**Example 9.2.7.** *Consider a random variable $X : \Omega \to \mathbb{R}$ with c.d.f. $F_X$.*

1. *Find the c.d.f. of* $Y = \max\{X, 3\}$*: Let* $y \in \mathbb{R}$*, then*

$$F_Y(y) = \mathrm{P}(\max\{X,3\} \leq y) = \mathrm{P}(X \leq y, 3 \leq y) = \mathrm{P}(X \leq y)\mathbb{I}_{[3,\infty)}(y)$$
$$= \begin{cases} F_X(y), & \text{for } y \geq 3, \\ 0, & \text{for } y < 3. \end{cases}$$

2. *Find the c.d.f. of* $Y = |X|$*. Let* $y \in \mathbb{R}$*, then*

$$F_Y(y) = \mathrm{P}(|X| \leq y) = \begin{cases} 0, & \text{for } y < 0, \\ \mathrm{P}(-y \leq |X| \leq y), & \text{for } y \geq 0. \end{cases}$$
$$= \begin{cases} 0, & \text{for } y < 0, \\ F_X(y) - F_X((-y)-), & \text{for } y \geq 0. \end{cases}$$

*Recall that* $F_X((-y)-)$ *is the left limit of* $F_X$ *at the point* $-y$*.*

**End of lecture 13.** . . . . . .

# Chapter 10

# Expectation of random variables

*The material of this chapter is based on Blitzstein & Hwang (2019), p.149-174, Grimmett & Welsh (1986), p.29-32, 67-70, 90-92.*

This chapter introduces the *expectation* of a random variable. We will distinguish the two cases of a discrete and continuous random variable and study the so-called *law of the unconscious statistician* (LOTUS). We will also learn that the expectation is a *linear* operator and we will introduce the concept of a *variance* and other (higher) *moments*.

## 10.1 Definition of the expectation

Next we define the expectation of a discrete random variable.

**Definition 10.1.1** (Expectation of discrete random variable)**.** *Let $X$ denote a discrete random variable, then the* expectation *of $X$ is defined as*

$$\mathrm{E}(X) = \sum_{x \in \mathrm{Im} X} x \mathrm{P}(X = x)$$

*whenever the sum on the right hand side converges absolutely, i.e. when we have $\sum_{x \in \mathrm{Im} X} |x| \mathrm{P}(X = x) < \infty$.*[1]

The expectation of $X$ is also called *expected value* or *mean*. Note that we typically simplify the notation and write

$$\mathrm{E}(X) = \sum_{x} x \mathrm{P}(X = x) = \sum_{x} x p_X(x).$$

**Definition 10.1.2** (Expectation of a continuous random variable)**.** *For a continuous random variable $X$ with density $f_X$, we define the* expectation *of $X$ as*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx,$$

*provided that $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.*

As in the discrete case, we often refer to the expectation as *mean* or *expected value*.

[1] This assumption matters in the case when $\mathrm{Im} X$ is infinite. If the sum converges absolutely, then the sum takes the same value irrespectively of the order of summation.

**Remark 10.1.3.** *Recall that we said that $p_X(x)$ for a discrete random variable is comparable to $f_X(x)dx$ for a continuous random variable. Also, in the discrete case, we deal with sums, whereas in the continuous case we have integrals. Using these analogies it makes sense to use the definition*

$$\mathrm{E}(X) = \begin{cases} \sum_x x p_X(x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

**Remark 10.1.4.** *Note that the above definition of the expectation ensures that the expectation is finite. We can relax that definition slightly and allow for infinite expectations as well. In doing that, we need to be a bit more careful with the precise definition. We can proceed as follows.*

*Suppose $X$ is a non-negative discrete/continuous random variable with pmf $p_X$/pdf $f_X$. Then, define*

$$\mathrm{E}(X) = \begin{cases} \sum_{x \geq 0} x p_X(x) & \text{if } X \text{ is discrete,} \\ \int_0^{\infty} x f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases} \tag{10.1.1}$$

*Here the sum/integral is allowed to take the value $+\infty$.*

*For a general discrete/continuous random variable, which is not necessarily restricted to be non-negative, we can define its positive and negative parts as follows:*

$$X^+ = \max\{0, X\}, \qquad X^- = \max\{0, -X\}.$$

*Since both $X^+$ and $X^-$ are non-negative, we can now define $\mathrm{E}(X^+)$ and $\mathrm{E}(X^-)$ as in* (10.1.1)*. Here the expectations can take the value $+\infty$.*

*Since*

$$X = X^+ - X^-,$$

*we can then define $\mathrm{E}(X)$ as*

$$\mathrm{E}(X) = \mathrm{E}(X^+) - \mathrm{E}(X^-),$$

*provided the right hand side is not of the form $\infty - \infty$, in which case we would say that the corresponding expectation is undefined.*

## 10.2 Law of the unconscious statistician (LOTUS)

Consider the situation that we have a transformation of a random variable $Y = g(X)$ and we would like to find its expectation. The law of the unconscious statistician will tell us that we do not need to find the p.m.f./p.d.f. of the transformed variable but rather use the following formula:

**Theorem 10.2.1** (LOTUS: Discrete case)**.** *Let $X$ be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathrm{E}(g(X)) = \sum_{x \in \mathrm{Im}X} g(x)\mathrm{P}(X = x),$$

*whenever the sum on the right hand side converges absolutely.*

*Proof.* We note that if $Y = g(X)$, then according to equation (9.1.1) the p.m.f. of $Y$ is given by

$$\mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x).$$

Hence

$$\begin{aligned} \mathrm{E}(Y) &= \sum_{y \in \mathrm{Im}Y} y\mathrm{P}(Y = y) = \sum_{y \in \mathrm{Im}Y} y\mathrm{P}(g(X) = y) \\ &= \sum_{y \in \mathrm{Im}Y} y \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x) \end{aligned}$$

$$= \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y : y = g(x)} y \mathrm{P}(X = x)$$
$$= \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y : y = g(x)} g(x) \mathrm{P}(X = x)$$
$$= \sum_{x \in \mathrm{Im}X} g(x) \mathrm{P}(X = x),$$

where we were allowed to interchange the order of summation since the sum converges absolutely. □

**Example 10.2.2.** *Consider a Bernoulli random variable $X \sim \mathrm{Bern}(p)$. We compute its mean as follows:*

$$\mathrm{E}(X) = \sum_x x \mathrm{P}(X = x) = 0 \cdot \mathrm{P}(X = 0) + 1 \cdot \mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

*Next, we want to find $\mathrm{E}(X^2)$. Using Theorem 10.2.1, we find*

$$\mathrm{E}(X^2) = \sum_x x^2 \mathrm{P}(X = x) = 0^2 \cdot \mathrm{P}(X = 0) + 1^2 \cdot \mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

**Example 10.2.3.** *Let $X \sim \mathrm{Poi}(\lambda)$ for $\lambda > 0$. Find $\mathrm{E}(X!)$.*

$$\mathrm{E}(X!) \overset{\text{LOTUS}}{=} \sum_{n=0}^{\infty} n! \mathrm{P}(X = n) = \sum_{n=0}^{\infty} n! \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_{n=0}^{\infty} \lambda^n \overset{\text{geom. series}, |\lambda|<1}{=} e^{-\lambda} \frac{1}{1-\lambda},$$

*if $|\lambda| = \lambda < 1$ and $\mathrm{E}(X!) = \infty$ for $\lambda \geq 1$.*

We will now state (without proof) the LOTUS for the continuous case:

**Theorem 10.2.4** (LOTUS: Continuous case)**.** *Let $X$ be a continuous random variable with density $f_X$, consider a function[2] $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

*provided that $\int_{-\infty}^{\infty} |g(x)| f_X(x) dx < \infty$.*

**Example 10.2.5.** *Take $g(x) = x^k$ for $k \in \mathbb{N}$. Then $\mathrm{E}(X^k)$ is called the $k$th moment of $X$ (provided it exists).*

The LOTUS theorems imply the linearity of the expectation in the following sense:

**Theorem 10.2.6.** *Consider a discrete/continuous random variable $X$ with finite expectation.*

1. *If $X$ is non-negative, then $\mathrm{E}(X) \geq 0$.*
2. *If $a, b \in \mathbb{R}$, then $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.*

*Proof.* The proof is left as an exercise, see Exercise 6- 2. □

**Example 10.2.7.** *Let $X$ be a continuous random variable with density $f_X(x) = cx^2$ for $x \in [0, 2]$ and $f_X(x) = 0$ otherwise. Find $c$ and $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$. The probability density function needs to be nonnegative and integrate to 1, hence we set*

$$1 = \int_{-\infty}^{\infty} f_X(x) dx = c \int_0^2 x^2 dx = c \frac{1}{3} x^3 \Big|_0^2 = c \frac{8}{3},$$

*which implies that $c = \frac{3}{8} (\geq 0)$.*

[2] We assume that the function $g$ is such that $g(X)$ is also a continuous random variable.

*Then*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^2 \frac{3}{8}x^3 dx = \frac{3}{8}\frac{1}{4}x^4\Big|_0^2 = \frac{3}{2}.$$

*Using the LOTUS, we get*

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^2 \frac{3}{8}x^4 dx = \frac{3}{8}\frac{1}{5}x^5\Big|_0^2 = \frac{12}{5}.$$

## 10.3 Variance

While the expectation tells you something about the centre of the distribution, in many applications we also want to know about the dispersion of $X$ about its mean value. Hence we introduce the so-called *variance*

**Definition 10.3.1** (Variance)**.** *Let $X$ be a discrete/continuous random variable. Then its variance is defined as*

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2],$$

*provided that it exists. Often we write $\sigma^2 = \mathrm{Var}(X)$.*

From Theorem 10.2.6 we deduce that the variance of a random variable is always non-negative.

If we are considering a random variable which is just given by a deterministic constant, e.g. for some $c \in \mathbb{R}$ we have $\mathrm{P}(X = c) = 1$, then $\mathrm{E}(X) = \sum_x x\mathrm{P}(X = x) = c \cdot \mathrm{P}(X = c) = c$ and $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}[(c-c)^2] = 0$. That means that only true randomness generates a non-zero variance.

In practice, it is often easier to work with a slightly different expression for the variance which we shall derive next.

**Theorem 10.3.2.** *For a discrete/continuous random variable with finite variance we have that*

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

*Proof.* In order to simplify the notation we write $\mu = \mathrm{E}(X)$. In the discrete case we have, using Theorem 10.2.1,

$$\begin{aligned}
\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \sum_x (x^2 - 2\mu x + \mu^2) p_X(x) \\
&= \sum_x x^2 p_X(x) + \sum_x (-2\mu x) p_X(x) + \sum_x \mu^2 p_X(x) \\
&= \sum_x x^2 p_X(x) - 2\mu \sum_x x p_X(x) + \mu^2 \sum_x p_X(x) \\
&= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.
\end{aligned}$$

In the continuous case, we have after applying Theorem 10.2.4,

$$\begin{aligned}
\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx + \int_{-\infty}^{\infty} (-2\mu x) f_X(x)dx + \int_{-\infty}^{\infty} \mu^2 f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)dx \\
&= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.
\end{aligned}$$

□

A very useful property of the variance is that it is not affected by deterministic additions and a multiplicative constant can be taken out of the variance provided we square it:

**Theorem 10.3.3.** *Let $X$ be a discrete/continuous random variable with finite variance and consider deterministic constants $a, b \in \mathbb{R}$. Then*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

*Proof.* This is left as an exercise, see Exercise 6- 3. □

**End of lecture 14.** .....

# Chapter 11

# Bridging lecture: Multivariate calculus

*The material of this chapter is based on Blitzstein & Hwang (2019), p.594–596.*

In this bridging lecture we will state the main concepts from multivariate calculus which we will need in the Y1 Probability and Statistics course in order to be able to study multivariate random variables. More details of this material will be provided in the Analysis and Calculus courses at a later point in time.

In this lecture, we will focus on bi-variable calculus only, but the concepts will extend to the general multivariate case.

## 11.1 Partial derivatives

Suppose you have a bivariate function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) = xy^2 + x^3y.$$

We would like to find the *partial derivative* with respect to $x$. Then we treat the variable $y$ as a constant and differentiate the function $g(x) := f(x, y)$ in the usual way with respect to $x$. This leads to

$$\frac{\partial f(x, y)}{\partial x} = \frac{d}{dx}g(x) = y^2 + 3x^2y.$$

Similarly, if we would like to find the *partial derivative* with respect to $y$, then we treat the variable $x$ as a constant and differentiate the function $h(y) := f(x, y)$ in the usual way with respect to $y$. This leads to

$$\frac{\partial f(x, y)}{\partial y} = \frac{d}{dy}h(y) = 2xy + x^3.$$

The partial derivatives above are so-called *first order* partial derivatives. Repeating the steps above, by taking partial derivatives of the partial derivatives, we get *second order* partial derivatives. In our example, this leads to

$$\frac{\partial^2 f(x, y)}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f(x, y)}{\partial x}\right) = 2y + 3x^2.$$

We note that it does not matter in which order we differentiate and we get the same result when we compute

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f(x, y)}{\partial y}\right) = 2y + 3x^2.$$

Under mild technical assumptions, we have for a general function $f$ that

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} = \frac{\partial^2 f(x, y)}{\partial y \partial x}$$

Consider a transformation which maps $\mathbf{x} = (x_1, x_2)$ to $\mathbf{y} = (y_1, y_2)$. Then the *Jacobian* of the transformation is defined as the 2x2 matrix of all possible first order partial derivatives:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix}$$

## 11.2 Bivariate integrals

Let $A \subseteq \mathbb{R}^2$ and $f : \mathbb{R}^2 \to \mathbb{R}$. Then (under mild conditions), we have that the order of integration does not matter and that

$$\int\int_A f(x,y)dxdy = \int\int_A f(x,y)dydx,$$

and the joint integral can be computed by iteratively computing the univariate integrals. I.e., similar to the concept of partial derivatives, we can compute multiple integrals by treating the variable which is not the integration variable constant and integrate out one variable at a time. We illustrate this idea in an example. Let $A = \{(x,y) : 0 \leq x \leq y \leq 1\} \subseteq \mathbb{R}^2$ and $f(x,y) = xy^2 + x^3y$. First we will be integrating with respect to $x$ and then with respect to $y$:

$$\begin{aligned}
\int\int_A f(x,y)dxdy &= \int_0^1 \int_0^y (xy^2 + x^3y)dxdy = \int_0^1 \left( \frac{1}{2}x^2y^2 + \frac{1}{4}x^4y \Big|_0^y \right) dy \\
&= \int_0^1 \left( \frac{1}{2}y^4 + \frac{1}{4}y^5 \right) dy \\
&= \frac{1}{2}\frac{1}{5}y^5 + \frac{1}{4}\frac{1}{6}y^6 \Big|_0^1 = \frac{1}{10} + \frac{1}{24} = \frac{17}{120}.
\end{aligned}$$

Alternatively, we can also integrate with respect to $y$ first and then with respect to $x$. When switching the order of integration, we need to carefully check the area of integration defined by $A$. Then we have

$$\begin{aligned}
\int\int_A f(x,y)dydx &= \int_0^1 \int_x^1 (xy^2 + x^3y)dydx = \int_0^1 \left( \frac{1}{3}xy^3 + \frac{1}{2}x^3y^2 \Big|_x^1 \right) dx \\
&= \int_0^1 \left( \frac{1}{3}x + \frac{1}{2}x^3 - \frac{1}{3}x^4 - \frac{1}{2}x^5 \right) dx \\
&= \frac{1}{2}\frac{1}{3}x^2 + \frac{1}{2}\frac{1}{4}x^4 - \frac{1}{3}\frac{1}{5}x^5 - \frac{1}{2}\frac{1}{6}x^6 \Big|_0^1 \\
&= \frac{1}{6} + \frac{1}{8} - \frac{1}{15} - \frac{1}{12} = \frac{17}{120}.
\end{aligned}$$

## 11.3 Change of variables formula

As in the univariate case, a change of variables formula also exist for the multivariate case. In that formula, a so-called Jacobian appears. We will explain the key ideas again in the bivariate setting.

Let $f : \mathbb{R}^2 \to \mathbb{R}$. We define the mapping $T : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$T(x,y) = (u(x,y), v(x,y)),$$

and assume that $T$ is a bijection from the domain $D \subseteq \mathbb{R}^2$ to some range $S \subseteq \mathbb{R}^2$. Then we can write $T^{-1} : S \to D$ for the inverse mapping of $T$, i.e. $(x, y) = T^{-1}(u, v)$. For the first component we write $x = x(u, v)$ and for the second $y = y(u, v)$. The *Jacobian determinant* of $T^{-1}$ is defined as the determinant

$$J(u,v) = \det\left(\frac{\partial(x,y)}{\partial(u,v)}\right) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

The change of variable formula states that (under mild conditions[1])

$$\int\int_D f(x,y)dxdy = \int\int_S f(x(u,v), y(u,v))|J(u,v)|dudv. \tag{11.3.1}$$

### 11.3.1 Example using polar coordinates

Suppose we want to compute the integral

$$I := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2+y^2)\right) dxdy.$$

Now we consider the (invertible) transformation to polar coordinates:

$$x = r\cos(\theta), \qquad y = r\sin(\theta),$$

for $r > 0$ and $\theta \in [0, 2\pi)$. Then, we compute the Jacobian determinant $J(r, \theta)$ of the transformation as follows:

$$J(r,\theta) = \det\left(\frac{\partial(x,y)}{\partial(r,\theta)}\right) = \det\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} = r(\cos^2(\theta) + \sin^2(\theta)) = r.$$

Then,

$$\begin{aligned} I &= \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{1}{2}(r^2\cos^2(\theta) + r^2\sin^2(\theta))\right)|J(r,\theta)|drd\theta \\ &= \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{1}{2}r^2\right) rdrd\theta. \end{aligned}$$

You can now do another variable transformation (or integrate directly): We set $u = r^2/2$, the $du = rdr$ and

$$I = \int_0^{2\pi}\left(\int_0^{\infty} e^{-u}du\right)d\theta = \int_0^{2\pi}\left(-e^{-u}|_{u=0}^{\infty}\right)d\theta = \int_0^{2\pi} 1d\theta = 2\pi.$$

Hence, we have that $\sqrt{I} = \sqrt{2\pi}$. Now you might wonder why we do these kind of computations in the probability course...

Suppose you would like to check that the standard normal density is indeed a valid density and, apart from being nonnegative, satisfies

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1. \tag{11.3.2}$$

It turns out that this integral can be computed using the transformation to polar coordinates used above. To this end, note that showing equation (11.3.2) is equivalent to showing that

[1] For instance, we need that the partial derivatives in the Jacobian exist and are continuous and that the Jacobian determinant is never 0.

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi}.$$

Squaring both sides leads to

$$\left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx\right)^2 = 2\pi.$$

Now we expand the left hand side as follows:

$$\begin{aligned}\left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx\right)^2 &= \left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx\right) \cdot \left(\int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2}\right) dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2+y^2)\right) dxdy = I.\end{aligned}$$

Since we have already shown that $I = 2\pi$, we can conclude that $\int_{-\infty}^{\infty} \phi(x)dx = 1$.

In the next chapter, we will also show how the change of variables formula can be applied in the context of transformations of multivariate random variables.

**End of bridging lecture** .....

# Chapter 12

# Multivariate random variables

*The material of this chapter is based on Blitzstein & Hwang (2019), p.129-133, 303-306, 312-313, Grimmett & Welsh (1986), p.36-43, 75-88.*

## 12.1 Multivariate distributions

### 12.1.1 The bivariate case

Let us now consider two (arbitrary, i.e. not restricted to discrete or continuous) random variables $X$ and $Y$ on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We would like to understand how they relate to each other and whether or not they are independent. We will write them as a random vector $(X, Y)$ taking values in $\mathbb{R}^2$.

**Definition 12.1.1** (Joint distribution function)**.** *The* joint distribution function *of the random vector $(X, Y)$ is defined as the mapping $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by*

$$F_{X,Y}(x, y) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x, Y(\omega) \leq y\}), \qquad \textit{for any } x, y \in \mathbb{R}.$$

Using our shortened notation, we typically write

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y), \qquad \text{for any } x, y \in \mathbb{R}.$$

We can now list some of the key properties of joint distribution functions:

- $F_{X,Y}$ is non-decreasing in each variable, meaning that
$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \qquad \text{if } x_1 \leq x_2 \text{ and } y_1 \leq y_2.$$
- $F_{X,Y}$ is *continuous from above* (the multivariate version of right-continuity), i.e., for two sequences $(x_n), (y_n)$ which approach $x$ and $y$ from the right as $n \to \infty$ we get that $F_{X,Y}(x_n, y_n) \to F_{X,Y}(x, y)$.
- We have the following two limits:
$$\lim_{x \to -\infty, y \to -\infty} F_{X,Y}(x, y) = 0, \qquad \lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1.$$
- They determine the *marginal distributions* uniquely, i.e.
$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y), \qquad F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y).$$

**Example 12.1.2.** *Let $F_{X,Y}$ denote the joint c.d.f. of $(X, Y)$. For $x, y \in \mathbb{R}$, find an expression for $\mathrm{P}(X \leq x, Y \geq y)$ in terms of $F_{X,Y}$. We note that*

$$\mathrm{P}(X \leq x, Y \geq y) + \mathrm{P}(X \leq x, Y < y) \overset{\text{Law of total prob.}}{=} \mathrm{P}(X \leq x) = F_{X,Y}(x, \infty).$$

*Also, $\mathrm{P}(X \leq x, Y < y) = F_{X,Y}(x, y-)$. Hence,*

$$\mathrm{P}(X \leq x, Y \geq y) = F_{X,Y}(x, \infty) - F_{X,Y}(x, y-).$$

### 12.1.2 The $n$-dimensional case

The extension to the $n$-dimensional case (for $n \in \mathbb{N}$) is now straightforward: We consider random variables $X_1, \ldots, X_n$ on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We write $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. Then the joint distribution function of $\mathbf{X}$ is given by $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$:

$$F_{\mathbf{X}}(\mathbf{x}) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

## 12.2 Independence

We have now all the tools to define what we mean by independence of (general) random variables: We call random variables $X$ and $Y$ independent if the events $\{\omega \in \Omega : X(\omega) \leq x\}$ and $\{\omega \in \Omega : Y(\omega) \leq y\}$ are independent for all $x, y \in \mathbb{R}$. I.e. we define:

**Definition 12.2.1** (Independence of random variables)**.** *The random variables $X$ and $Y$ are independent if and only if*

$$\mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\mathrm{P}(Y \leq y), \qquad \textit{for all } x, y \in \mathbb{R},$$

*which is equivalent to saying that the joint distribution function factorises as the product of the two marginal distribution functions:*

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \qquad \textit{for all } x, y \in \mathbb{R}.$$

We call the random variables $X_1, \ldots, X_n$ *independent* if

$$\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathrm{P}(X_1 \leq x_1) \cdots \mathrm{P}(X_n \leq x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

or equivalently if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

**Definition 12.2.2** (Pairwise independence for $n \in \mathbb{N}, n > 2$ random variables)**.** *We call the random variables $X_1, \ldots, X_n$* pairwise independent *if*

$$F_{X_i,X_j}(x_i, x_j) = F_{X_i}(x_i)F_{X_j}(x_j), \quad \textit{for all } x_i, x_j \in \mathbb{R} \textit{ whenever } i \neq j.$$

**Remark 12.2.3.** *Independence of random variables implies pairwise independence, the reverse statement, however, is not true in general.*

Finally, we define what we mean by independence of a family of (infinitely many) random variables.

**Definition 12.2.4** (Independence of a family of random variables)**.** *Let $\mathcal{I} \subset \mathbb{R}$ denote an index set. A family of random variables $\{X_i : i \in \mathcal{I}\}$ is said to be* independent *if for all finite subsets $\mathcal{J} \subseteq \mathcal{I}$ and all $x_j \in \mathbb{R}, j \in \mathcal{J}$, the following product rule holds:*

$$\mathrm{P}(\cap_{j \in \mathcal{J}} \{X_j \leq x_j\}) = \prod_{j \in \mathcal{J}} \mathrm{P}(X_j \leq x_j).$$

**Remark 12.2.5.** *Note that for independent random variables $X_1, \ldots X_n$, and continuous functions $f_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \ldots, n$ the transformed random variables $Y_1 = f_1(X_1), \ldots, Y_n = f_n(X_n)$ are also independent. [It would be sufficient to assume that functions $f_i$ are Borel-measurable, but this concept is beyond the scope of this course.]*

## 12.3 Multivariate discrete distributions and independence

**Definition 12.3.1** (Joint probability mass function)**.** *Let $X, Y$ denote discrete random variables on $(\Omega, \mathcal{F}, \mathrm{P})$. Their* joint probability mass function *denoted by $p_{X,Y}$ is defined as the function $p_{X,Y} : \mathbb{R}^2 \to [0,1]$ given by*

$$p_{X,Y}(x,y) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}),$$

*which is typically shortened to*

$$p_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y).$$

*We have that $p_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$ and $\sum_x \sum_y p_{X,Y}(x,y) = 1$.*

*The* marginal probability mass functions *of $X$ and $Y$ are then given by*

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad \text{and } p_Y(y) = \sum_x p_{X,Y}(x,y).$$

It turns out that for any "nice" set $A \subseteq \mathbb{R}^2$, we obtain that

$$\mathrm{P}((X,Y) \in A) = \sum \sum_{(x,y) \in A} \mathrm{P}(X = x, Y = y).$$

### 12.3.1 Independence

Definition 12.2.1 covers the case of general random variables. In the discrete (or continuous) case, we can formulate equivalent independence conditions:

**Definition 12.3.2** (Independence of discrete random variables)**.** *Suppose that $X$ and $Y$ are discrete random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. $X$ and $Y$ are said to be* independent *if the pair of events $\{\omega \in \Omega : X(\omega) = x\}$ and $\{\omega \in \Omega : Y(\omega) = y\}$ are independent for all $x, y \in \mathbb{R}$, i.e. if*

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y), \quad \text{for all } x, y \in \mathbb{R}. \tag{12.3.1}$$

Condition (12.3.1) is equivalent to saying that

$$p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \text{for all } x, y \in \mathbb{R}.$$

Random variables which are not independent are called *dependent*.

## 12.4 Multivariate continuous distributions and independence

We can also extend the concept of a continuous random variable to random vectors. Again, we will be focussing on the bivariate case, but the $n$-dimensional case works in exactly the same way.

**Definition 12.4.1** (Continuous random vector)**.** *We call the random vector $(X,Y)$ on $(\Omega, \mathcal{F}, \mathrm{P})$* (jointly) continuous *if*

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v) dv du,$$

*for a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ satisfying*

*(i)* $f_{X,Y}(u,v) \geq 0$ *for all* $u, v \in \mathbb{R}$,

*(ii)* $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v) dv du = 1$.

*We call $f_{X,Y}$ the* (joint) density *function of $(X,Y)$.*

Similar to the univariate case, we typically obtain the joint density by differentiating the joint distribution function. I.e. we take

$$f_{X,Y}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y), & \text{if this derivative exists at } (x,y), \\ 0, & \text{otherwise.} \end{cases}$$

It turns out that for any "nice" set $A \subseteq \mathbb{R}^2$, we obtain that

$$\mathrm{P}((X,Y) \in A) = \int \int_{(x,y)\in A} f_{X,Y}(x,y) dx dy.$$

We do not prove this result here formally.

We note that the *marginal densities* can be obtained from the joint density as follows:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f_{X,Y}(u,v) dv du \\ &= \int_{v=-\infty}^{\infty} f_{X,Y}(x,v) dv, \end{aligned}$$

and also

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u,y) du.$$

### 12.4.1 Independence

From our definition of independence of random variables, see Definition 12.2.1, we can immediately deduce by differentiating/integrating that jointly continuous random variables $X$ and $Y$ are independent if and only if their joint density factorises:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \qquad \text{for all } x, y \in \mathbb{R}.$$

### 12.4.2 Examples

**Example 12.4.2.** *Suppose the joint density of $(X,Y)$ is given by*

$$f_{X,Y}(x,y) = \begin{cases} \frac{7}{\sqrt{2\pi}} e^{-x^2/2 - 7y}, & \textit{if } -\infty < x < \infty, y > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

*We want to check whether or not $X$ and $Y$ are independent and compute $\mathrm{P}(X > 2, Y < 1)$. We note that for $x \in \mathbb{R}, y > 0$, we can write*

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot 7 e^{-7y},$$

*which is in fact the product of a standard normal random variable and an $\mathrm{Exp}(7)$ random variable. Hence, we have $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all $x, y \in \mathbb{R}$ which implies independence. Hence*

$$\begin{aligned} \mathrm{P}(X > 2, Y < 1) &= \mathrm{P}(X > 2)\mathrm{P}(Y < 1) = (1 - \Phi(2)) F_Y(1) \\ &= (1 - \Phi(2))(1 - e^{-7}). \end{aligned}$$

The following is a worked example where you can practice doing computations involving bivariate p.d.f.s and c.d.f.s.

**Example 12.4.3 (Reading material: Worked example).** *Consider jointly continuous random variables $X, Y$ with joint density given by*

$$f_{X,Y}(x,y) = \begin{cases} c(x^2 + y^2), & \text{for } 0 < x < 2, 0 < y < 2, \\ 0, & \text{otherwise.} \end{cases}$$

1. *Find $c$ such that $f_{X,Y}$ is a p.d.f.: We need that $c \geq 0$ and*

$$1 = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx = \int_{x=0}^{2}\left(\int_{y=0}^{2} c(x^2+y^2)dy\right)dx = c\int_{x=0}^{2}\left(x^2y + \frac{1}{3}y^3\Big|_{y=0}^{2}\right)dx$$
$$= c\int_{x=0}^{2}\left(2x^2+\frac{8}{3}\right)dx = c\left(\frac{2}{3}x^3+\frac{8}{3}x\Big|_{x=0}^{2}\right) = c\frac{32}{3} \Leftrightarrow c = \frac{3}{32}.$$

2. *Find the joint c.d.f. of $(X,Y)$.*

**Case 1: Let $x \leq 0$ or $y \leq 0$:** $F_{X,Y}(x,y) = 0$.

**Case 2: Let $x,y \in (0,2)$:**

$$F_{X,Y}(x,y) = \int_{u=0}^{x}\left(\int_{v=0}^{y} c(u^2+v^2)dv\right)du = c\int_{u=0}^{x}\left(u^2v + \frac{1}{3}v^3\Big|_{v=0}^{y}\right)du$$
$$= c\int_{u=0}^{x}\left(u^2y+\frac{1}{3}y^3\right)du = c\left(\frac{1}{3}u^3y + \frac{1}{3}y^3u\Big|_{u=0}^{x}\right) = c\left(\frac{1}{3}x^3y+\frac{1}{3}y^3x\right)$$
$$= \frac{1}{32}(x^3y+xy^3).$$

**Case 3: Let $x \in (0,2), y \geq 2$:** $F_{X,Y}(x,y) = \frac{1}{32}\left(2x^3+8x\right) = \frac{1}{16}x^3+\frac{1}{4}x = F_X(x)$.

**Case 4: Let $y \in (0,2), x \geq 2$:** $F_{X,Y}(x,y) = \frac{1}{32}\left(2y^3+8y\right) = \frac{1}{16}y^3+\frac{1}{4}y = F_Y(y)$.

**Case 5: Let $x,y \geq 2$:** $F_{X,Y}(x,y) = 1$.

3. *Differentiate the c.d.f. to obtain the p.d.f.:*

**Case 1: Let $x \leq 0$ or $y \leq 0$:** $F_{X,Y}(x,y) = 0$. *Hence*

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = 0.$$

**Case 2: Let $x,y \in (0,2)$:** $F_{X,Y}(x,y) = \frac{1}{32}(x^3y+xy^3)$. *Hence*

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = \frac{\partial}{\partial y}\left[\frac{\partial}{\partial x}\frac{1}{32}(x^3y+xy^3)\right] = \frac{\partial}{\partial y}\left[\frac{1}{32}(3x^2y+y^3)\right]$$
$$= \frac{1}{32}(3x^2+3y^2) = \frac{3}{32}(x^2+y^2).$$

**Case 3: Let $x \in (0,2), y \geq 2$:** $F_{X,Y}(x,y) = \frac{1}{16}x^3+\frac{1}{4}x$. *Hence* $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = 0$.

**Case 4: Let $y \in (0,2), x \geq 2$:** $F_{X,Y}(x,y) = \frac{1}{16}y^3+\frac{1}{4}y$. *Hence* $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = 0$.

**Case 5: Let $x,y \geq 2$:** $F_{X,Y}(x,y) = 1$. *Hence* $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = 0$.

4. *Find the marginal densities of $X$ and $Y$.*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^2 c(x^2+y^2)dy = c\left(x^2y+\frac{1}{3}y^3\Big|_{y=0}^{2}\right)$$
$$= c\left(2x^2+\frac{8}{3}\right) = \frac{3}{16}x^2+\frac{1}{4},$$

*for $x \in (0,2)$ and $f_X(x) = 0$ otherwise. By symmetry, $f_Y(y) = \frac{3}{16}y^2+\frac{1}{4}$, for $y \in (0,2)$ and $f_Y(y) = 0$ otherwise.*

5. *Show that $X$ and $Y$ are not independent.*
   *We have*
$$f_X(x)f_Y(y) = \left(\frac{3}{16}x^2 + \frac{1}{4}\right)\left(\frac{3}{16}y^2 + \frac{1}{4}\right) \neq f_{X,Y}(x,y),$$
   *for $x, y \in (0,2)$, hence $X$ and $Y$ are not independent.*

6. *Find the marginal c.d.f.s of $X$ and $Y$.*
   *$F_X(x) = F_{X,Y}(x,\infty) = \frac{1}{16}x^3 + \frac{1}{4}x$ for $x \in (0,2)$, $F_X(x) = 0$ for $x \leq 0$ and $F_X(x) = 1$ for $x \geq 2$. Also, $F_Y(x) = F_{X,Y}(\infty, y) = \frac{1}{16}y^3 + \frac{1}{4}y$ for $y \in (0,2)$, $F_Y(y) = 0$ for $y \leq 0$ and $F_Y(y) = 1$ for $y \geq 2$.*

## 12.5 Transformations of random vectors: The bivariate case

In Chapter 9 we discussed how we can compute the p.d.f. of a transformed random variable. Here we will illustrate how the methodology works in bivariate setting, where the change of variable formula, see equation (11.3.1), discussed in the bridging lecture plays a key role.

Consider the case of jointly continuous random variables $(X, Y)$ with density $f_{X,Y}$. Let $u, v : \mathbb{R}^2 \to \mathbb{R}$ denote deterministic functions and define a new pair of random variables by

$$U = u(X,Y), \qquad V = v(X,Y).$$

We would like to find the joint density of $(U, V)$.

We define the mapping $T : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$T(x,y) = (u(x,y), v(x,y)),$$

and assume that $T$ is a bijection from the domain $D = \{(x,y) : f_{X,Y}(x,y) > 0\} \subseteq \mathbb{R}^2$ to some range $S \subseteq \mathbb{R}^2$. Then we can write $T^{-1} : S \to D$ for the inverse mapping of $T$, i.e. $(x,y) = T^{-1}(u,v)$. For the first component we write $x = x(u,v)$ and for the second $y = y(u,v)$. The *Jacobian determinant* of $T^{-1}$ is defined as the determinant

$$J(u,v) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Then the joint density of $(U, V)$ is given by

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}(x(u,v), y(u,v))|J(u,v)|, & \text{if } (u,v) \in S, \\ 0, & \text{otherwise.} \end{cases}$$

**Example 12.5.1.** *Let us demonstrate how the methodology works in practice, see Grimmett & Welsh (1986, p. 87).*

*Suppose that $X, Y \sim \text{Exp}(1)$ are independent. Define*

$$U := X + Y, \qquad V := \frac{X}{X+Y}.$$

*We want to find the joint density of $(U, V)$ and the marginal densities of $U$ and $V$.*

*First, we note that the joint density of $(X, Y)$ is – due to independence – given by*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = e^{-(x+y)}, \text{ if } x, y > 0,$$

*and zero otherwise.*

*In our case, the mapping $T$ is given by*

$$T(x,y) = (u,v) = \left(x + y, \frac{x}{x+y}\right),$$

*where $T$ maps the set $D = \{(x, y) : x, y > 0\}$, onto $S = \{(u, v) : 0 < u < \infty, 0 < v < 1\}$.*

*Next, we find the inverse function of $T$:*

$$T^{-1}(u, v) = (x, y) = (uv, (1 - v)u).$$

*The Jacobian of $T^{-1}$ is given by*

$$J(u, v) = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \det \begin{pmatrix} v & u \\ (1 - v) & -u \end{pmatrix} = -uv - (1 - v)u = -u.$$

*Then, for $(u, v) \in S$, we have*

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}(x(u, v), y(u, v))|J(u, v)| \\ &= \exp(-(uv + (1 - v)u)) \, | - u| = u \exp(-u), \end{aligned}$$

*and zero otherwise. I.e.*

$$f_{U,V}(u, v) = \begin{cases} ue^{-u}, & \text{for } u > 0, 0 < v < 1, \\ 0 & \text{otherwise.} \end{cases}$$

*The marginal density of $U$ is given by*

$$f_U(u) = \int_0^1 f_{U,V}(u, v)dv = \int_0^1 u \exp(-u)dv = u \exp(-u), \text{ for } u > 0,$$

*and zero otherwise. Hence, $U \sim \text{Gamma}(2, 1)$. The marginal density of $V$ is given by*

$$f_V(v) = \int_0^\infty f_{U,V}(u, v)du = \int_0^\infty u \exp(-u)du = \Gamma(2) = 1, \text{ for } 0 < v < 1,$$

*and zero otherwise. Hence $V \sim \text{U}(0, 1)$.*

*We notice that $f_{U,V}(u, v) = f_U(u)f_V(v)$ for all $u, v$, which implies that $U$ and $V$ are independent.*

**End of lecture 15.** . . . . . .

## 12.6 Two dimensional law of the unconscious statistician (2D LOTUS)

Using exactly the same arguments as in the univariate case, we can also formulate a law of the unconscious statistician applied to a function of a (bivariate) random vector. We will only state the result here.

**Theorem 12.6.1** (2D LOTUS: discrete case)**.** *Let $X, Y$ denote discrete random variables on $(\Omega, \mathcal{F}, \text{P})$ and let $g : \mathbb{R}^2 \to \mathbb{R}$. Then $Z = g(X, Y)$ is also a discrete random variable on $(\Omega, \mathcal{F}, \text{P})$ and its expectation is given by*

$$\text{E}(g(X, Y)) = \sum_{x \in \text{Im}X} \sum_{y \in \text{Im}Y} g(x, y)\text{P}(X = x, Y = y).$$

**Theorem 12.6.2** (2D LOTUS: continuous case)**.** *Let $X, Y$ be jointly continuous random variables with density $f_{X,Y}$ and let $h : \mathbb{R}^2 \to \mathbb{R}$. Then*

$$\text{E}[h(X, Y)] = \int_{-\infty}^\infty \int_{-\infty}^\infty h(x, y)f_{X,Y}(x, y)dxdy.$$

Theorems 12.6.1,12.6.2 can be used to prove the linearity of the expectation:

**Theorem 12.6.3** (Linearity of expectation)**.** *Let $X, Y$ denote jointly discrete/continuous random variables on $(\Omega, \mathcal{F}, \mathrm{P})$, and $a, b \in \mathbb{R}$, then*

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y),$$

*provided that $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ exist.*

*Proof.* In the discrete case, we apply Theorem 12.6.1 with $g(x, y) = ax + by$. Then

$$\begin{aligned}
\mathrm{E}(aX + bY) &= \sum_x \sum_y (ax + by)\mathrm{P}(X = x, Y = y) \\
&= a \sum_x x \sum_y \mathrm{P}(X = x, Y = y) + b \sum_y y \sum_x \mathrm{P}(X = x, Y = y) \\
&= a \sum_x x\mathrm{P}(X = x) + b \sum_y y\mathrm{P}(Y = y) \\
&= a\mathrm{E}(X) + b\mathrm{E}(Y).
\end{aligned}$$

In the continuous case, we apply Theorem 12.6.2 with $g(x, y) = ax + by$. Then

$$\begin{aligned}
\mathrm{E}(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{X,Y}(x, y) dx dy \\
&= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y) dx dy \\
&= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= a\mathrm{E}(X) + b\mathrm{E}(Y).
\end{aligned}$$

□

Using induction, one can easily deduce that for $n \in \mathbb{N}$ and random variables $X_1, \ldots, X_n$ with finite expectations and constants $a_1, \ldots, a_n \in \mathbb{R}$ we have

$$\mathrm{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathrm{E}(X_1) + \cdots + a_n \mathrm{E}(X_n). \tag{12.6.1}$$

**Remark 12.6.4.** *It is important to remember that the linearity of the expectation* (12.6.1) *holds in general without assuming any independence between the random variables.*

## 12.7 Covariance and correlation between random variables

**Definition 12.7.1.** *Consider two (one-dimensional) random variables $X$ and $Y$ on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The* covariance *of $X$ and $Y$ is defined as*

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

*if the expectation on the right hand side takes a finite value. Also, we define the* correlation *of $X$ and $Y$ as*

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

When we set $X = Y$, then the covariance simplifies to the variance:

$$\mathrm{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2] = \mathrm{Var}(X).$$

For concrete computations it is often useful to work with the following alternative expression for the covariance.

**Theorem 12.7.2** (Covariance)**.** *For jointly discrete/continuous random variables $X, Y$ with finite expectations, we have*

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

*Proof.* See Exercise 6- 9. □

The following example follows on from Example 12.4.3. Here you can practice deriving the covariance and correlation of two random variable.

**Example 12.7.3** (**Reading material: Worked example**)**.** *Consider the same jointly continuous random variables $X, Y$ as in Example 12.4.3. Their joint density given by*

$$f_{X,Y}(x,y) = \begin{cases} c(x^2+y^2), & \text{for } 0<x<2, 0<y<2, \\ 0, & \text{otherwise.} \end{cases}$$

*for $c = \frac{3}{32}$.*

1. *Find $\mathrm{Cov}(X,Y)$.*
   *We note that $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.*
   *Recall that*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^2 c(x^2+y^2)dy = c\left(x^2y + \frac{1}{3}y^3\Big|_{y=0}^{2}\right) \\ &= c\left(2x^2 + \frac{8}{3}\right) = \frac{3}{16}x^2 + \frac{1}{4}, \end{aligned}$$

   *for $x \in (0,2)$ and $f_X(x) = 0$ otherwise. Then*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^2 \left(\frac{3}{16}x^3 + \frac{1}{4}x\right)dx = \frac{3}{16}\cdot\frac{1}{4}x^4 + \frac{1}{4}\cdot\frac{1}{2}x^2\Big|_{x=0}^{2} = \frac{3}{4} + \frac{1}{2} = \frac{5}{4}.$$

   *By symmetry, we also have that $\mathrm{E}(Y) = \frac{5}{4}$. Also,*

$$\begin{aligned} \mathrm{E}(XY) &\overset{\text{LOTUS}}{=} \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} xy f_{X,Y}(x,y)dydx = c\int_{x=0}^{2}\left(\int_{y=0}^{2}(x^3y + xy^3)dy\right)dx \\ &= c\int_{x=0}^{2}\left(\frac{1}{2}x^3y^2 + \frac{1}{4}xy^4\Big|_{y=0}^{2}\right)dx = c\int_{x=0}^{2}\left(2x^3+4x\right)dx = c\left(\frac{2}{4}x^4 + \frac{4}{2}x^2\right)\Big|_{x=0}^{2} \\ &= \frac{3}{32}(8+8) = \frac{3}{2}. \end{aligned}$$

   *Hence*

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \frac{3}{2} - \frac{5^2}{4^2} = -\frac{1}{16}.$$

2. *Find $\mathrm{Cor}(X,Y)$.*
   *Recall that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$. Here we have*

$$\begin{aligned} \mathrm{E}(X^2) &\overset{\text{LOTUS}}{=} \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^2\left(\frac{3}{16}x^4 + \frac{1}{4}x^2\right)dx = \frac{3}{16}\cdot\frac{1}{5}x^5 + \frac{1}{4}\cdot\frac{1}{3}x^3\Big|_{x=0}^{2} \\ &= \frac{6}{5} + \frac{2}{3} = \frac{28}{15}. \end{aligned}$$

   *Hence*

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{28}{15} - \frac{5^2}{4^2} = \frac{73}{240},$$

*and by symmetry* $\mathrm{Var}(Y) = \frac{73}{240}$. *Hence*

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{-\frac{1}{16}}{\frac{73}{240}} = -\frac{15}{73} \approx -0.2.$$

**Remark 12.7.4.** *It is important to note that independent random variables always have zero covariance, but the converse does not hold in general!*

A very important property of independent random variables is the fact that the expectation of their product can be written as the product of their expectation (a property which does not hold in general!):

**Theorem 12.7.5.** *Let* $X, Y$ *denote independent and jointly discrete/continuous random variables with finite expectation, then*

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y). \tag{12.7.1}$$

*Proof.* In the case when $X, Y$ are jointly discrete, we use Theorem 12.6.1 with $g(x,y) = xy$. Then

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_x \sum_y xy \mathrm{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy \mathrm{P}(X = x)\mathrm{P}(Y = y) \quad \text{(by independence)} \\
&= \sum_x x\mathrm{P}(X = x) \sum_y y\mathrm{P}(Y = y) \quad \text{(using the existence of } \mathrm{E}(X), \mathrm{E}(Y)) \\
&= \mathrm{E}(X)\mathrm{E}(Y).
\end{aligned}$$

Using Theorem 12.6.2 and similar computations as above gives us the result for the jointly continuous case. □

**Remark 12.7.6.** *It is important to note that if* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$*, then this does not in general imply that* $X$ *and* $Y$ *are independent, see Exercise 6- 6.*

The results stated in Theorem 12.7.5 can be extended to the $n \in \mathbb{N}$ dimensional case by induction: If $X_1, \ldots, X_n$ are independent, then

$$\mathrm{E}(X_1 \cdots X_n) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_n).$$

In statistics, we often deal with sums of random variables. How can we compute their variance? The following theorem gives an answer.

**Theorem 12.7.7** (Variance of a sum of random variables)**.** *Let* $X, Y$ *denote two jointly discrete/continuous random variables with finite variances. Then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

*Proof.* See Exercise 6- 10. □

**End of lecture 16.** .....

# Chapter 13

# Generating functions

*The material of this chapter is based on Blitzstein & Hwang (2019), p.279-293, Grimmett & Welsh (1986), p.45-52.*

In probability theory we often use so-called generating functions to derive/prove statements regarding the distribution of random variables/vectors or to compute moments. In this course, we study so-called *probability generating functions* and *moment generating functions* and we will give an outlook on what *characteristic functions* are.

## 13.1 Probability generating functions

First of all, we introduce so-called *probability generating functions* and explain why they are extremely useful!

Throughout this section, we will only consider **discrete random variables** taking values in the **non-negative integers**, i.e. $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\} = \{0, 1, 2, \dots\}$.

**Definition 13.1.1** (Probability generating function (p.g.f.))**.** *Let $X$ denote a discrete random variable with $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\}$. We denote by*

$$\mathcal{S}_X = \left\{ s \in \mathbb{R} : \sum_{x=0}^{\infty} |s|^x \mathrm{P}(X = x) < \infty \right\}.$$

*Then the* probability generating function *(pgf) of $X$ is defined as the function $G_X : \mathcal{S}_X \to \mathbb{R}$ given by*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \mathrm{P}(X = x).$$

We observe that the pgf is well-defined for $|s| \leq 1$ since

$$\sum_{x=0}^{\infty} |s|^x \mathrm{P}(X = x) \leq \sum_{x=0}^{\infty} \mathrm{P}(X = x) = 1 < \infty.$$

Also, $G_X(0) = \mathrm{P}(X = 0)$ and $G_X(1) = 1$.

The reason why probability generating functions are extremely useful is that they uniquely determine the probability mass function (i.e. the distribution) of a discrete random variable:

**Theorem 13.1.2.** *Let $X, Y$ denote discrete random variables with $\mathrm{Im}X, \mathrm{Im}Y \subseteq \mathbb{N} \cup \{0\}$. Their p.g.f.s are denoted by $G_X$ and $G_Y$, respectively. Then*

$$G_X(s) = G_Y(s), \text{ for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y, \tag{13.1.1}$$

*if and only if*

$$\mathrm{P}(X = x) = \mathrm{P}(Y = x), \text{ for all } x = 0, 1, 2, \dots. \tag{13.1.2}$$

*Proof.* Assume that (13.1.1) holds. First we note that $G_X(0) = G_Y(0)$ implies $\mathrm{P}(X = 0) = \mathrm{P}(Y = 0)$. When we differentiate[1] the pgfs we get

$$G'_X(s) = \sum_{x=1}^{\infty} x s^{x-1} \mathrm{P}(X = x).$$

When we plug in $s = 0$ in the first derivative, we get $G'_X(0) = \mathrm{P}(X = 1)$. Hence, $G'_X(0) = \mathrm{P}(X = 1) = \mathrm{P}(Y = 1) = G'_Y(0)$. This procedure can be repeated and we obtain

$$\frac{d^n}{ds^n} G_X(s)\Big|_{s=0} = n!\mathrm{P}(X = n),$$

the same can be done for $G_Y$, and the identity of the two pgfs implies the result.

The other direction of the proof is trivial. □

**Example 13.1.3.** *Let $X$ be a discrete random variable with $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\}$. Suppose that*

$$G_X(s) = \frac{1}{3} + \frac{1}{5}s^5 + \frac{1}{5}s^{10} + \frac{4}{15}s^{12}.$$

*Find the p.m.f. of $X$. Recall that*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \mathrm{P}(X = x) = \mathrm{P}(X = 0) + s\mathrm{P}(X = 1) + s^2\mathrm{P}(X = 2) + \cdots.$$

*Hence, in the example above, we can just read off the probabilities:*

$$\mathrm{P}(X = 0) = \frac{1}{3}, \mathrm{P}(X = 5) = \frac{1}{5}, \mathrm{P}(X = 10) = \frac{1}{5}, \mathrm{P}(X = 12) = \frac{4}{15},$$

*and $\mathrm{P}(X = x) = 0$ for $x \notin \{0, 5, 10, 12\}$.*

### 13.1.1 Common probability generating functions

We will now list the p.g.f.s of some common discrete distributions.

**Example 13.1.4** (Bernoulli distribution)**.** *Let $X \sim \mathrm{Bern}(p)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = s^0\mathrm{P}(X = 0) + s^1\mathrm{P}(X = 1) = 1 - p + sp$$

*for all $s \in \mathbb{R}$.*

**Example 13.1.5** (Binomial distribution)**.** *Let $X \sim \mathrm{Bin}(n, p)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{n} \binom{n}{x} s^x p^x (1-p)^{n-x} = (1 - p + sp)^n,$$

*for all $s \in \mathbb{R}$, by an application of the binomial theorem.*

**Example 13.1.6** (Poisson distribution)**.** *Let $X \sim \mathrm{Poi}(\lambda)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(s\lambda)^x}{x!} = e^{-\lambda} e^{s\lambda} = \exp(\lambda(s-1)),$$

*for all $s \in \mathbb{R}$. Here we used the series expansion of the exponential function.*

[1] You will learn in the real analysis course under which conditions we are allowed to interchange the infinite sum and the derivative. For the purpose of this course, we will just assume that the above computation is valid.

### 13.1.2 Probability generating function of a sum of independent discrete random variables

**Theorem 13.1.7.** *Let $X, Y$ be independent discrete random variables with $\mathrm{Im}X, \mathrm{Im}Y \subseteq \mathbb{N} \cup \{0\}$. Then*

$$G_{X+Y}(s) = G_X(s)G_Y(s), \text{ for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y.$$

*Proof.* Let $s \in \mathcal{S}_X \cap \mathcal{S}_Y$. Since $X$ and $Y$ are independent, Exercise 6- 7 implies that $s^X$ and $s^Y$ satisfy the product formula for expectations (in fact, they are also independent and Theorem 12.7.5 applies). Hence using we conclude that

$$G_{X+Y}(s) = \mathrm{E}(s^{X+Y}) = \mathrm{E}(s^X s^Y) = \mathrm{E}(s^X)\mathrm{E}(s^Y) = G_X(s)G_Y(s).$$

□

An immediate consequence of the above results is, that for independent non-negative integer-valued random variables $X_1, \ldots, X_n$ ($n \in \mathbb{N}$), we have

$$G_{\sum_{i=1}^n X_i}(s) = \prod_{i=1}^n G_{X_i}(s),$$

for all $s \in \cap_{i=1}^n \mathcal{S}_{X_i}$.

### 13.1.3 Moments

We have already introduced, the mean and variance of a (discrete) random variable $X$. More generally, for $k \in \mathbb{N}$, we call $\mathrm{E}(X^k)$ the $k$th moment of $X$ provided it exists. It turns out that we can use the probability generating function for deriving moments of random variables. More precisely, we differentiate the pgf $k$ times and plug in $s = 1$:

**Theorem 13.1.8.** *Let $X$ be a discrete random variable with $\mathrm{Im}X \in \mathbb{N} \cup \{0\}$. Let $k \in \mathbb{N}$. Then the $k$th derivative of the pgf is given by*

$$\frac{d^k}{ds^k} G_X(s)\Big|_{s=1} = G^{(k)}(1) = \mathrm{E}[X(X-1)\cdots(X-k+1)].$$

*Proof.* As before, we assume that we are allowed to interchange derivatives and summation under suitable conditions. Then

$$\frac{d}{ds} G_X(s) = \frac{d}{ds}\mathrm{E}(s^X) = \mathrm{E}(Xs^{X-1}).$$

Hence

$$\frac{d}{ds} G_X(s)\Big|_{s=1} = \mathrm{E}(X).$$

Similarly,

$$\frac{d^k}{ds^k} G_X(s) = \frac{d}{ds}\mathrm{E}(s^X) = \mathrm{E}[X(X-1)\cdots(X-k+1)s^{X-k}].$$

Hence

$$\frac{d^k}{ds^k} G_X(s)\Big|_{s=1} = \mathrm{E}[X(X-1)\cdots(X-k+1)].$$

□

**Example 13.1.9** (Computing the variance using pgfs)**.** *The above theorem can be used for computing the variance of a discrete non-negative integer-valued random variable $X$. We note that*

$$G_X''(1) = \mathrm{E}[X(X-1)] = \mathrm{E}(X^2) - \mathrm{E}(X).$$

*Hence,*

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = G_X''(1) + G_X'(1) - (G_X'(1))^2.$$

**Example 13.1.10.** *Compute the mean and variance of the Bernoulli, Binomial and Poisson distributions using the probability generating functions.*

- *For $X \sim \mathrm{Bern}(p)$, we have, for $s \in \mathbb{R}$,*

$$G_X(s) = \mathrm{E}(s^X) \overset{\text{LOTUS}}{=} \sum_x s^x \mathrm{P}(X=x) = s^0 \mathrm{P}(X=0) + s^1 \mathrm{P}(X=1) = 1 - p + sp.$$

*Then*

$$\frac{d}{ds} G_X(s)\Big|_{s=1} = p, \qquad \frac{d^2}{ds^2} G_X(s)\Big|_{s=1} = 0.$$

*Hence*

$$\mathrm{E}(X) = G_X'(1) = p, \qquad \mathrm{Var}(X) = G_X''(1) + G_X'(1) - (G_X'(1))^2 = p - p^2 = p(1-p).$$

- *For $X \sim \mathrm{Bin}(n,p)$, we have, for $s \in \mathbb{R}$,*

$$G_X(s) = \mathrm{E}(s^X) \overset{\text{LOTUS}}{=} \sum_x s^x \mathrm{P}(X=x) = \sum_{x=0}^n s^x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^n \binom{n}{x} (sp)^x (1-p)^{n-x}$$
$$\overset{\text{Binomial theorem}}{=} (sp + 1 - p)^n.$$

*Then*

$$\frac{d}{ds} G_X(s)\Big|_{s=1} = n(sp+1-p)^{n-1} p\big|_{s=1} = np,$$
$$\frac{d^2}{ds^2} G_X(s)\Big|_{s=1} = n(n-1)(sp+1-p)^{n-2} p^2\big|_{s=1} = n(n-1)p^2.$$

*Hence*

$$\mathrm{E}(X) = G_X'(1) = np,$$
$$\mathrm{Var}(X) = G_X''(1) + G_X'(1) - (G_X'(1))^2 = n^2p^2 - np^2 + np - (np)^2 = np(1-p).$$

- *For $X \sim \mathrm{Poi}(\lambda)$, we have, for $s \in \mathbb{R}$,*

$$G_X(s) = \mathrm{E}(s^X) \overset{\text{LOTUS}}{=} \sum_x s^x \mathrm{P}(X=x) = \sum_{x=0}^\infty s^x \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=0}^\infty \frac{(s\lambda)^x}{x!} e^{-\lambda} = e^{\lambda s} e^{-\lambda} = \exp(\lambda(s-1)).$$

*Then*

$$\frac{d}{ds} G_X(s)\Big|_{s=1} = \exp(\lambda(s-1))\lambda\big|_{s=1} = \lambda,$$
$$\frac{d^2}{ds^2} G_X(s)\Big|_{s=1} = \exp(\lambda(s-1))\lambda^2\big|_{s=1} = \lambda^2.$$

*Hence*

$$\mathrm{E}(X) = G_X'(1) = \lambda,$$
$$\mathrm{Var}(X) = G_X''(1) + G_X'(1) - (G_X'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

**End of lecture 17.** .....

## 13.2 Moment generating functions

**Definition 13.2.1** (Moment generating functions)**.** *Let $X$ be a random variable. Then its* moment generating function *(m.g.f.) is defined as*

$$M_X(t) = \mathrm{E}(e^{tX}),$$

*provided the expectation exists in some neighbourhood of zero, i.e. the expectation exists for all $|t| < \epsilon$ for some $\epsilon > 0$.*

We compute the m.g.f. as follows:

$$M_X(t) = \mathrm{E}(e^{tX}) = \begin{cases} \sum_x e^{tx} p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, & \text{if } X \text{ is continuous,} \end{cases}$$

whenever the sum/integral is (absolutely) convergent.

**Remark 13.2.2.** *Relation between the p.g.f. and the m.g.f. for discrete random variables: Let $X$ be a discrete random variable with $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\}$. Then*

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}((e^t)^X) = G_X(e^t).$$

**Remark 13.2.3.** *Why is $M_X$ called the* moment generating function*?*
*Let $X$ be a random variable. Then its* moment generating function *(m.g.f.) is defined as*

$$M_X(t) = \mathrm{E}(e^{tX}),$$

*provided the expectation exists in some neighbourhood of zero, i.e. the expectation exists for all $|t| < \epsilon$ for some $\epsilon > 0$. Then*

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right) \overset{(*)}{=} \sum_{n=0}^{\infty} \mathrm{E}(X^n) \frac{t^n}{n!}.$$

*Note that, in general, we are not allowed to interchange an infinite sum with the expectation. However, here the equality in (*) holds, since we assume the existence of the moment generating function in a neighbourhood of zero.*

*Also, we can do a Taylor series expansion of $M_X(t)$ around $0$, which leads to*

$$M_X(t) = \sum_{n=0}^{\infty} M^n(0) \frac{t^n}{n!}.$$

*Clearly, for the two infinite series to be the same, we need that $M^{(n)}(0) = \mathrm{E}(X^n)$.*

**Example 13.2.4.** *Let $X \sim \mathrm{N}(0,1)$. Then we use the trick of "completing the square" in the second line:*

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 + tx} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx + t^2)} e^{\frac{t^2}{2}} dx \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx = e^{\frac{t^2}{2}}, \end{aligned}$$

*since the latter integral is equal to 1 since it is the integral of a $\mathrm{N}(t,1)$ density function. We note that the m.g.f. exists for all $t \in \mathbb{R}$ in this case.*

Let us now consider an example where the m.g.f. does not exist for all $t \in \mathbb{R}$.

**Example 13.2.5.** *Let $X \sim \mathrm{Exp}(\lambda)$, then*

$$\begin{aligned} M_X(t) = \mathrm{E}(e^{tX}) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \int_0^{\infty} e^{(t-\lambda)x} \lambda dx \\ &= \begin{cases} \frac{\lambda}{\lambda - t}, & \text{if } t < \lambda, \\ \infty, & \text{if } t \geq \lambda. \end{cases} \end{aligned}$$

### 13.2.1 Properties

**Theorem 13.2.6.** *If $X$ has a m.g.f., then for $k \in \mathbb{N}$, the $k$th moment of $X$ is given by*

$$\mathrm{E}(X^k) = M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}.$$

*Proof.* We give a sketch proof of the theorem. Assuming we can interchange expectation and differentiation, we write

$$\frac{d^k}{dt^k} M_X(t) = \frac{d^k}{dt^k} \mathrm{E}(e^{tX}) = \mathrm{E}\left(\frac{d^k}{dt^k} e^{tX}\right) = \mathrm{E}\left(X^k e^{tX}\right),$$

and then plug in $t = 0$. □

**Theorem 13.2.7.** *For $a, b \in \mathbb{R}$, we have $M_{aX+b}(t) = e^{bt} M_X(at)$.*

*Proof.*

$$M_{aX+b}(t) = \mathrm{E}\{\exp[t(aX+b)]\} = \mathrm{E}[e^{taX+tb}] = e^{tb}\mathrm{E}[\exp(taX)] = e^{tb} M_X(at).$$

□

**Example 13.2.8.** *Let $Z \sim \mathrm{N}(0,1)$. Let $\mu \in \mathbb{R}, \sigma > 0$, $X := \mu + \sigma Z$, then*

$$M_X(t) = e^{t\mu} M_Z(\sigma t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2},$$

*which is the m.g.f. of an $\mathrm{N}(\mu, \sigma^2)$ distributed random variable. We can find the mean of $X$, by*

$$\mathrm{E}(X) = M_X'(0) = e^{\mu t + \sigma^2 t^2/2}(\mu + \sigma^2 t)\Big|_{t=0} = \mu.$$

**Theorem 13.2.9.** *Let $X_1, \ldots, X_n$ denote a sequence of independent random variables with m.g.f.s $M_{X_1}, \ldots, M_{X_n}$. Then*

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

*Proof.* Since the expectation of a product of independent random variables is the product of their corresponding expectation, see Theorem 12.7.5, we have

$$M_{\sum_{i=1}^n X_i}(t) = \mathrm{E}\left[\exp\left(t\sum_{i=1}^n X_i\right)\right] = \mathrm{E}\left[\prod_{i=1}^n \exp(tX_i)\right] = \prod_{i=1}^n \mathrm{E}(e^{tX_i}) = \prod_{i=1}^n M_{X_i}(t).$$

Here we used that continuous transformations (with $f(x) = e^{tx}$) of independent random variables are independent, too, see Remark 12.2.5. □

We will now state without proof the famous characterisation theorem:

**Theorem 13.2.10** (Characterisation)**.** *If the m.g.f.s of the random variables $X$ and $Y$ exist and $M_X(t) = M_Y(t)$ in a neighbourhood of zero, then*

$$F_X(u) = F_Y(u) \qquad \textit{for all } u.$$

The above theorem states that m.g.f. characterise the distribution of a random variable uniquely.

## 13.3 Using m.g.f.s for finding all moments of the exponential and the standard normal distributions

**Example 13.3.1.** *We would like to find all the moments of the exponential distribution:*

- *Let* $X \sim \mathrm{Exp}(1)$. *Then* $M_X(t) = (1-t)^{-1}$ *for all* $t < 1$. *Using the geometric series for* $|t| < 1$, *we have*

$$M_X(t) = \frac{1}{1-t} = \sum_{n=0}^{\infty} t^n = \sum_{n=0}^{\infty} n! \frac{t^n}{n!} = \sum_{n=0}^{\infty} \mathrm{E}(X^n) \frac{t^n}{n!},$$

  *hence* $\mathrm{E}(X^n) = n!$ *for all* $n \in \mathbb{N}$.

- *Now consider the general case, when* $Y \sim \mathrm{Exp}(\lambda)$. *Then* $X := \lambda Y \sim \mathrm{Exp}(1)$. *To see this, note that, for* $x > 0$,

$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(\lambda Y \leq x) = \mathrm{P}(Y \leq x/\lambda) = F_Y(x/\lambda) = 1 - \exp(-x),$$

  *and* $F_X(x) = 0$ *for* $x < 0$, *which is the c.d.f. of an* $\mathrm{Exp}(1)$*-distributed random variable. Then, we can deduce that, for all* $n \in \mathbb{N}$, *we have*

$$\mathrm{E}(X^n) = n! = \mathrm{E}(\lambda^n Y^n) = \lambda^n \mathrm{E}(Y^n) \Leftrightarrow \mathrm{E}(Y^n) = \frac{n!}{\lambda^n}.$$

  *In particular,* $\mathrm{E}(X) = \lambda^{-1}, \mathrm{Var}(X) = \lambda^{-2}$.

**Example 13.3.2.** *We would like to find all the moments of the standard normal distribution: Let* $X \sim \mathrm{N}(0,1)$. *Then*

$$M_X(t) = e^{t^2/2} = \sum_{n=0}^{\infty} \frac{(t^2/2)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!} = \sum_{n=0}^{\infty} \frac{(2n)!}{2^n n!} \cdot \frac{t^{2n}}{(2n)!} = \sum_{n=0}^{\infty} \mathrm{E}(X^{2n}) \frac{t^{2n}}{(2n)!}.$$

*I.e.* $\mathrm{E}(X^{2n}) = \frac{(2n)!}{2^n n!}$ *and* $\mathrm{E}(X^{2n-1}) = 0$ *for all* $n \in \mathbb{N}$.

The even moments can be computed using the following identity:

**Lemma 13.3.3.** $\frac{(2n)!}{2^n n!} = (2n-1)(2n-3)\cdots 3 \cdot 1$, *for* $n \in \mathbb{N}$.

*Proof.* We can give a story proof/proof by interpretation. Both sides count how many ways there are to break a group of $2n$ people into $n$ pairs: Left hand side: Take $2n$ people and label them $1$ to $2n$. We can line up the $2n$ people (there are $(2n)!$ possible permutations) and say that the first two are a pair, the next two are a pair etc. Here we overcount by a factor of $n!$ since the order of the pairs does not matter and by a factor of $2^n$ since the order within each pair does not matter. Right hand side: There are $2n-1$ ways to choose a partner for the first person, then there are $2n-3$ choices for person 2 (or 3 if 2 was already paired to person 1) etc. □

## 13.4 Outlook: Characteristic function and Laplace transform

Note that moment generating functions do not exist for all distributions.

Hence we often work with the *characteristic function* of a random variable $X$ instead which is defined as

$$\phi_X(t) = \mathrm{E}(e^{itX}) = \mathrm{E}[\cos(tX) + i\sin(tX)], \qquad \text{for all } t \in \mathbb{R},$$

where $i = \sqrt{-1}$. It turns out that characteristic functions exist indeed for all distributions and hence they are a useful tool for general proofs in probability theory. However, we will defer the detailed discussion of complex-valued objects to a later probability/statistics (and analysis!) course.

For a non-negative random variable $X$ we sometimes work with the *Laplace transform* instead which is defined as

$$\mathcal{L}_X(t) = \mathrm{E}(e^{-tX}), \qquad \text{for all } t \geq 0.$$

We note that $\mathcal{L}_X(t) = M_X(-t)$ for $t \geq 0$.

**Example 13.4.1.** *Let $X \sim \mathrm{Exp}(\lambda)$, then, for $t \geq 0$,*

$$\begin{aligned}\mathcal{L}_X(t) = \mathrm{E}(e^{-tX}) &= \int_0^\infty e^{-tx}\lambda e^{-\lambda x}dx = \int_0^\infty e^{-(t+\lambda)x}\lambda dx \\ &= \frac{\lambda}{\lambda + t}.\end{aligned}$$

**End of lecture 18.**

# Chapter 14

# Conditional distribution and conditional expectation

*The material of this chapter is based on Blitzstein & Hwang (2019), p.306-311, 313-321, Grimmett & Welsh (1986), p.32-33, 88-89, 92-95.*

Let us now study *conditional distributions* both for discrete and continuous random variables. They allow us to define the *conditional expectation*, which is a really useful concept as we shall see when stating the *law of total expectation*.

## 14.1 Discrete case: Conditional expectation and the law of total expectation

We have already introduced the notation of conditional probabilities. Now we are going to define the conditional distribution of a discrete random variable.

**Definition 14.1.1** (Conditional distribution and conditional expectation)**.** *Let $X$ denote a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Consider an event $B \in \mathcal{F}$ such that $\mathrm{P}(B) > 0$. The* conditional distribution of $X$ given $B$ *is defined as*

$$\mathrm{P}(X = x|B) = \frac{\mathrm{P}(\{X = x\} \cap B)}{\mathrm{P}(B)}, \text{ for } x \in \mathbb{R}.$$

*Further, the* conditional expectation of $X$ given $B$ *is defined as*

$$\mathrm{E}(X|B) = \sum_{x \in \mathrm{Im}X} x\mathrm{P}(X = x|B),$$

*provided the sum is absolutely convergent.*

Similarly to the ideas presented in the law of total probability, it can often be useful to consider a partition of the probability space to compute an (unconditional) expectation via conditional expectations as we describe in the following theorem.

**Theorem 14.1.2.** *[Law of total expectation] Consider a partition $\{B_i : i \in \mathcal{I}\}$ of $\Omega$ with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$. Let $X$ denote a discrete random variable with finite expectation. Then*

$$\mathrm{E}(X) = \sum_{i \in \mathcal{I}} \mathrm{E}(X|B_i)\mathrm{P}(B_i),$$

*whenever the sum converges absolutely.*

*Proof.* First we use the definition of the expectation, followed by the law of total probability (Theorem 5.4.4):

$$\begin{aligned}\mathrm{E}(X) &= \sum_x x\mathrm{P}(X=x) = \sum_x x \sum_{i\in\mathcal{I}} \mathrm{P}(X=x|B_i)\mathrm{P}(B_i)\\ &= \sum_{i\in\mathcal{I}} \mathrm{P}(B_i) \sum_x x\mathrm{P}(X=x|B_i) = \sum_{i\in\mathcal{I}} \mathrm{P}(B_i)\mathrm{E}(X|B_i).\end{aligned}$$

We use the fact that the series is absolutely convergent to justify that we are allowed to change the order of summation. □

### 14.1.1 Conditioning on a random variable

Suppose $(X, Y)$ are jointly discrete random variables. In the above definition, consider the event $B = \{X = x\}$ for some $x \in \mathbb{R}$ such that $p_X(x) = \mathrm{P}(X = x) > 0$. Then the *conditional distribution/probability mass function of* $Y$ *given* $X = x$ is given by

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \quad \text{for } y \in \mathbb{R}.$$

Also, the *conditional expectation of* $Y$ *given* $X = x$ is given by

$$\mathrm{E}(Y|X = x) = \sum_y y p_{Y|X}(y|x),$$

provided the sum is absolutely convergent.

Also, the LOTUS for conditional expectations says that

$$\mathrm{E}(g(Y)|X = x) = \sum_y g(y) p_{Y|X}(y|x).$$

Note that we can also formulate an independence condition in terms of conditional p.m.f.s: Discrete $X$ and $Y$ are independent if and only if

$$\mathrm{P}(Y = y|X = x) = \mathrm{P}(Y = y)$$

for all $x, y$ such that $\mathrm{P}(X = x) > 0$. Also, we get a Bayes' type result of the form

$$p_{Y|X}(y|x)p_X(x) = p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y),$$

for all $x, y$ for which $p_X(x), p_Y(y) > 0$.

### 14.1.2 Example

Let us study an example:

**Example 14.1.3.** *Suppose you sit in Heathrow waiting for your flight to go on your well deserved holiday. You denote by* $N$ *the total (random) number of planes arriving while you wait and you assume that, for some* $\lambda > 0$, $N \sim \mathrm{Poi}(\lambda)$. *Each plane, independently, turns out to be a British Airways plane with probability* $p \in (0, 1)$, *hence with probability* $1 - p$ *it will be a plane from another airline. We write* $N = X + Y$ *where* $X$ *represents the number of British Airways planes and* $Y$ *the number of planes from other airlines. You are wondering what might be the joint probability mass function of* $X$ *and* $Y$.

*You recall that the Bernoulli distribution describes binary outcomes with success probability* $p$. *So, every time a plane you observe turns out to be a British Airways plane, you view this as a success and a failure otherwise.*

*You recall that the number of success given the total number of trials follows a Binomial distribution, so more precisely, in your case you have for* $n \in \mathbb{N}$

$$X|N = n \sim \text{Bin}(n, p), \text{ and } Y|N = n \sim \text{Bin}(n, 1-p).$$

*Given this information, you try to compute* $\text{P}(X = x, Y = y)$*. For this, it would be really useful to know* $N$*, so let us apply the law of total probability given information on* $N$*. For* $x, y \in \mathbb{N} \cup \{0\}$*:*

$$\text{P}(X = x, Y = y) = \sum_{n=0}^{\infty} \text{P}(X = x, Y = y|N = n)\text{P}(N = n).$$

*Clearly* $\text{P}(X = x, Y = y|N = n) > 0 \Leftrightarrow x + y = n$*. So, in the sum, we can get rid off all the terms which result in conditional probabilities being equal to 0.*

$$\begin{aligned}\text{P}(X = x, Y = y) &= \sum_{n=0}^{\infty} \text{P}(X = x, Y = y|N = n)\text{P}(N = n) \\ &= \sum_{n:x+y=n} \text{P}(X = x, Y = y|N = n)\text{P}(N = n) \\ &= \text{P}(X = x, Y = y|N = x + y)\text{P}(N = x + y).\end{aligned}$$

*Conditional on the event that* $\{N = x + y\}$*, the events* $\{X = x\}$ *and* $\{Y = y\}$ *contain exactly the same information, hence we get*

$$\text{P}(X = x, Y = y|N = x + y)\text{P}(N = x + y) = \text{P}(X = x|N = x + y)\text{P}(N = x + y).$$

*It remains to plug in the Binomial and Poisson p.m.f.s:*

$$\begin{aligned}\text{P}(X = x, Y = y) &= \text{P}(X = x|N = x + y)\text{P}(N = x + y) \\ &= \binom{x+y}{x} p^x (1-p)^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} \\ &= \frac{(x+y)!}{x!y!} p^x (1-p)^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \cdot \frac{(1-p)^y \lambda^y}{y!} e^{-\lambda(1-p)},\end{aligned}$$

*which is in fact the product of the p.m.f. of a* $\text{Poi}(p\lambda)$ *and a* $\text{Poi}((1-p)\lambda)$ *random variable. Hence, we conclude that* $X$ *and* $Y$ *are independent and* $X \sim \text{Poi}(p\lambda)$ *and* $Y \sim \text{Poi}((1-p)\lambda)$*.*

## 14.2 Continuous case: Conditional density, conditional distribution and conditional expectation

Let us now consider two jointly continuous random variables $(X, Y)$. We cannot proceed as above to define the conditional distribution $\text{P}(Y \leq y|X = x)$ since we now have that $\text{P}(X = x)$ for all $x \in \mathbb{R}$. Hence we need to condition on an event with non-zero probability. Let $\epsilon > 0$, then we have

$$\begin{aligned}\text{P}(Y \leq y|x \leq X \leq x + \epsilon) &= \frac{\text{P}(Y \leq y, x \leq X \leq x + \epsilon)}{\text{P}(x \leq X \leq x + \epsilon)} \\ &= \frac{\int_{u=x}^{x+\epsilon} \int_{v=-\infty}^{y} f_{X,Y}(u, v) dv du}{\int_x^{x+\epsilon} f_X(u) du} = \frac{\frac{1}{\epsilon}\int_{u=x}^{x+\epsilon} \int_{v=-\infty}^{y} f_{X,Y}(u, v) dv du}{\frac{1}{\epsilon}\int_x^{x+\epsilon} f_X(u) du}.\end{aligned}$$

We define $H(y, u) := \int_{v=-\infty}^{y} f_{X,Y}(u, v) dv$. Then we can write

$$\text{P}(Y \leq y|x \leq X \leq x + \epsilon) = \frac{\frac{1}{\epsilon}\int_{u=x}^{x+\epsilon} H(y, u)}{\frac{1}{\epsilon}\int_x^{x+\epsilon} f_X(u) du}.$$

Now we let $\epsilon \to 0$ and get

$$\lim_{\epsilon \to 0} \mathrm{P}(Y \leq y | x \leq X \leq x + \epsilon) = \frac{H(y,x)}{f_X(x)} = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)} = \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)dv}{f_X(x)} = G(y).$$

So $G$ is a distribution function with density function

$$g(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \text{ for } y \in \mathbb{R}.$$

The derivations above only work in the case when $f_X(x) > 0$. Let us now state our formal definition:

**Definition 14.2.1** (Conditional distribution and conditional density)**.** *For two jointly continuous random variables $X, Y$, we define the* conditional density of $Y$ given $X = x$ *as*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{14.2.1}$$

*for all $y \in \mathbb{R}$ and for all $x \in \mathbb{R}$ for which $f_X(x) > 0$. The corresponding* conditional distribution function of $Y$ given $X = x$ *is then given by*

$$F_{Y|X=x}(y|x) = \frac{\int_{\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)},$$

*for all $y \in \mathbb{R}$ and for all $x \in \mathbb{R}$ for which $f_X(x) > 0$.*

Note that we can now also formulate an independence condition in terms of conditional p.d.f.s: Jointly continuous random variables $X$ and $Y$ are independent if and only if

$$f_{Y|X}(y|x) = f_Y(y),$$

for all $x, y$ such that $f_X(x) > 0$.

**Remark 14.2.2.** *Note that* (14.2.1) *also implies a Bayes' type formula:*

$$f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y),$$

*provided that $f_X(x), f_Y(y) > 0$.*

We note that we can now formulate a continuous version of the law of total probability[1]:

**Proposition 14.2.3.** *Let $(X, Y)$ be two jointly continuous random variables with joint probability density function $f_{X,Y}$ and marginal probability densities denoted by $f_X$ and $f_Y$, respectively. Let $F_{Y|X=x}$ denote the conditional distribution function of $Y$ given $X = x$. Then*

$$\mathrm{P}(Y \leq y) = \int_{\{x: f_X(x) > 0\}} F_{Y|X=x}(y|x) f_X(x) dx, \qquad y \in \mathbb{R}.$$

*Proof.* For $y \in \mathbb{R}$, we have

$$\begin{aligned}
\mathrm{P}(Y \leq y) &= \int_{-\infty}^{y} f_Y(v)dv, \\
f_Y(v) &= \int_{-\infty}^{\infty} f_{X,Y}(x,v)dx, \\
f_{Y|X=x}(y|x) &= \frac{f_{X,Y}(x,v)}{f_X(x)}, \text{for } x \text{ such that } f_X(x) > 0.
\end{aligned}$$

[1] The proof was part of an exam question in 2021.

Hence, we can write

$$\begin{aligned}
\mathrm{P}(Y \leq y) &= \int_{-\infty}^{y} f_Y(v)dv = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f_{X,Y}(x,v)dxdv \\
&= \int_{-\infty}^{y}\int_{\{x:f_X(x)>0\}} \frac{f_{X,Y}(x,v)}{f_X(x)} f_X(x)dxdv \\
&= \int_{-\infty}^{y}\int_{\{x:f_X(x)>0\}} f_{Y|X=x}(v|x)f_X(x)dxdv \\
&= \int_{\{x:f_X(x)>0\}}\int_{-\infty}^{y} f_{Y|X=x}(v|x)dv f_X(x)dx \\
&= \int_{\{x:f_X(x)>0\}} F_{Y|X=x}(y|x)f_X(x)dx.
\end{aligned}$$

We note that the interchange of the integrals can be justified by Tonelli's theorem, which states that for non-negative functions the order of the integration can be interchanged. The proof of Tonelli's theorem is beyond the scope of this course. □

Similarly to the discrete case, we can now define the conditional expectation and formulate the law of total expectation.

**Definition 14.2.4** (Conditional expectation)**.** *For two jointly continuous random variables $X, Y$, we define the* conditional expectation of $Y$ given $X = x$ *as*

$$\mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y\frac{f_{X,Y}(x,y)}{f_X(x)}dy,$$

*provided that $f_X(x) > 0$.*

Recall that, in the discrete case, Theorem 14.1.2 implies that for jointly discrete random variables $X, Y$ with $\mathrm{E}|Y| < \infty$, we have

$$\mathrm{E}(Y) = \sum_{x:\mathrm{P}(X=x)>0} \mathrm{E}(Y|X=x)\mathrm{P}(X=x),$$

whenever the sum converges absolutely. The continuous analogue reads as follows:

**Theorem 14.2.5** (Law of total expectation)**.** *For jointly continuous random variable $X, Y$ with $\mathrm{E}|Y| < \infty$, we have*

$$\mathrm{E}(Y) = \int_{\{x:f_X(x)>0\}} \mathrm{E}(Y|X=x)f_X(x)dx.$$

*Proof.* We use the definition of the expectation, the fact that the marginal density of $Y$ can be obtained by integrating out the joint density and equation (14.2.1):

$$\begin{aligned}
\mathrm{E}(Y) &= \int y f_Y(y)dy = \int\int y f_{X,Y}(x,y)dxdy \\
&= \int\int y f_{Y|X}(y|x)f_X(x)dxdy \\
&= \int \left(\int y f_{Y|X}(y|x)dy\right) f_X(x)dx \\
&= \int \mathrm{E}(Y|X=x)f_X(x)dx,
\end{aligned}$$

where we assume that the integrals range over the appropriate values for $x$ and $y$. □

**End of lecture 19.** . . . . . .

### 14.2.1 Examples

**Example 14.2.6.** *Consider two jointly continuous random variables with joint density (for $\lambda > 0$):*

$$f_{X,Y}(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & \text{for } 0 \le x \le y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

- *Find $f_{Y|X}$:*
  *Recall that $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$. We compute the marginal density of $X$ first:*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_x^{\infty} \lambda^2 e^{-\lambda y} dy = \lambda^2 \frac{(-1)}{\lambda} e^{-\lambda y}\Big|_{y=x}^{\infty} = \lambda e^{-\lambda x},$$

  *for $x \ge 0$ and $f_X(x) = 0$ for $x < 0$. Hence,*

$$f_{Y|X}(y|x) = \begin{cases} \frac{\lambda^2 e^{-\lambda y}}{\lambda e^{-\lambda y}} = \lambda e^{-\lambda(y-x)}, & \text{for } 0 \le x \le y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

- *Find $\mathrm{E}(Y|X=x)$.*
  *For $x \ge 0$, we have*

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_x^{\infty} y\lambda e^{-\lambda(y-x)}dy = e^{\lambda x}\int_x^{\infty} \lambda y e^{-\lambda y} \\ &\overset{u:=\lambda y}{=} e^{\lambda x}\int_{\lambda x}^{\infty} u e^{-u} du \lambda^{-1} = \frac{e^{\lambda x}}{\lambda}\left\{ u(-e^{-u})\big|_{u=\lambda x}^{\infty} - \int_{\lambda x}^{\infty}(-e^{-u})du \right\} \\ &= \frac{e^{\lambda x}}{\lambda}\left\{\lambda x e^{-\lambda x} + e^{-\lambda x}\right\} = \frac{1}{\lambda}(\lambda x + 1). \end{aligned}$$

**Example 14.2.7.** *Let $\rho \in (-1,1)$. The standard bivariate normal distribution has joint density given by*

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[ -\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right]$$

*for $x, y \in \mathbb{R}$. We want to demonstrate some of the concepts introduced earlier.*

1. *What is the marginal density of $X$? We compute*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[ -\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right] dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[ -\frac{1}{2(1-\rho^2)}\left[(y-\rho x)^2 + x^2\left(1-\rho^2\right)\right]\right] dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[ -\frac{(y-\rho x)^2}{2(1-\rho^2)}\right] dy. \end{aligned}$$

   *We observe that the integrand in the above integral is the density of an $\mathrm{N}(\rho x, 1-\rho^2)$ random variable and hence the integral equals 1. Hence*

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

   *I.e. $X \sim \mathrm{N}(0,1)$ (and also $Y \sim \mathrm{N}(0,1)$).*

2. *What is the conditional density of $Y$ given $X = x$? We have*

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2\right)\right] \sqrt{2\pi} e^{x^2/2} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2\right) + x^2/2\right] \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2 - x^2(1-\rho^2)\right)\right] \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(y-\rho x)^2}{2\left(1-\rho^2\right)}\right]. \end{aligned}$$

*This is in fact the density of an $\mathrm{N}(\rho x, 1-\rho^2)$ random variable.*

3. *What is the conditional expectation of $Y$ given $X = x$?*

*Using the definition of the conditional expectation, we have*

$$\mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \rho x,$$

*given our above finding that $Y|X = x \sim \mathrm{N}(\rho x, 1-\rho^2)$.*

4. *Formulate a condition which ensures that $X$ and $Y$ are independent.*

*We know that $X$ and $Y$ are independent, if and only if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all $x, y \in \mathbb{R}$. For any $x, y \in \mathbb{R}$ we have*

$$\begin{aligned} & f_{X,Y}(x,y) = f_X(x) f_Y(y) \\ & \iff \\ & \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2\right)\right] = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\ & \iff \\ & \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(x^2 - 2\rho xy + y^2\right)\right] = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \\ & \iff \rho = 0. \end{aligned}$$

*So $X$ and $Y$ are independent if and only if $\rho = 0$.*

5. *Find the covariance between $X$ and $Y$.*

*We note that since $\mathrm{E}(X) = 0 = \mathrm{E}(Y)$, we have that*

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XY) = \int_{-\infty}^{\infty} \mathrm{E}(XY|X=x) f_X(x) dx,$$

*where we used the law of the total expectation, see Theorem 14.2.5. Note that $\mathrm{E}(XY|X=x) = \mathrm{E}(xY|X=x) = x\mathrm{E}(Y|X=x) = \rho x^2$, hence*

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \int_{-\infty}^{\infty} \rho x^2 f_X(x) dx = \rho \int_{-\infty}^{\infty} x^2 f_X(x) dx = \rho \mathrm{E}(X^2) \\ &= \rho \mathrm{Var}(X) = \rho. \end{aligned}$$

*So, we have*

$$\rho = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y),$$

*which, with our findings above, implies the following important result:* **Assume that $X, Y$ follow a bivariate (standard) normal distribution.** *Then $X$ and $Y$ are independent if and only if $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.*

**Warning:** *As soon as you drop the assumption that you are dealing with jointly normal random variables, then we only know that if we assume that they are independent, then the product formula for the expectations holds. However, if we have only verified that the product formula for the expectations holds, then that does not imply in general independence of the random variables. We will illustrate this in the following remark and example.*

**Remark 14.2.8.**

- *If $(X, Y)$ is bivariate normal and $\mathrm{Cov}(X, Y) = 0 \Rightarrow X, Y$ are independent.*
- *However, if $X$ and $Y$ follow a univariate normal distribution and $\mathrm{Cov}(X, Y) = 0 \not\Rightarrow X, Y$ are independent.*

**Example 14.2.9.** *Let $X \sim \mathrm{N}(0, 1)$. Let $Z$ be a discrete random variable, independent of $X$ with $\mathrm{P}(Z = -1) = \mathrm{P}(Z = 1) = \frac{1}{2}$. Let $Y := Z \cdot X$. We want to show that*

1. *$Y \sim \mathrm{N}(0, 1)$,*
2. *$\mathrm{Cov}(X, Y) = 0$,*
3. *$X$ and $Y$ are not independent.*

1. *$Y \sim \mathrm{N}(0, 1)$:*
   *Let $y \in \mathbb{R}$. Then, using the law of total probability and the independence of $Z$ and $X$, we have*

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(ZX \leq y) \\ &= \mathrm{P}(ZX \leq y | Z = -1)\mathrm{P}(Z = -1) + \mathrm{P}(ZX \leq y | Z = 1)\mathrm{P}(Z = 1) \\ &= \frac{1}{2}\left(\mathrm{P}(-X \leq y) + \mathrm{P}(X \leq y)\right) = \frac{1}{2}(\mathrm{P}(X \geq -y) + \Phi(y)) \\ &= \frac{1}{2}(1 - \mathrm{P}(X \leq -y) + \Phi(y)) = \frac{1}{2}(1 - \Phi(-y) + \Phi(y)) \\ &= \frac{1}{2}(1 - (1 - \Phi(y)) + \Phi(y)) = \Phi(y). \end{aligned}$$

   *Hence $Y \sim \mathrm{N}(0, 1)$.*

2. *$\mathrm{Cov}(X, Y) = 0$:*

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(X^2 Z) \overset{\text{independence of } X,Z}{=} \mathrm{E}(X^2)\mathrm{E}(Z) = 1 \cdot 0 = 0.$$

3. *$X$ and $Y$ are not independent:*
   *We note that $|X| = |Y|$ is always true, hence $X$ and $Y$ are not independent.*

*As a side remark, we note that the sum of $X$ and $Y$ is not normally distributed (and not even continuous!):*

$$\begin{aligned} \mathrm{P}(X + Y = 0) &= \mathrm{P}(X + ZX = 0) = \mathrm{P}(X(1 + Z) = 0) \\ &= \mathrm{P}(X(1 + Z) = 0 | Z = 1)\mathrm{P}(Z = 1) + \mathrm{P}(X(1 + Z) = 0 | Z = -1)\mathrm{P}(Z = -1) \\ &= \mathrm{P}(2X = 0) \cdot \frac{1}{2} + \mathrm{P}(0 = 0) \cdot \frac{1}{2} = \frac{1}{2} \neq 0. \end{aligned}$$

**Conclusions:** *If we only know that $X$ and $Y$ follow univariate normal distributions, we* **cannot** *conclude that*

- $(X, Y)$ *has a bivariate normal distribution,*
- $(X, Y)$ *are jointly continuous,*
- $\mathrm{Cov}(X, Y) = 0 \Rightarrow X, Y$ *are independent.*

**End of lecture 20.** . . . . . .

# Bibliography


Anderson, D. F., Seppäläinen, T. & Valkó, B. (2018), *Introduction to probability*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.

Blitzstein, J. K. & Hwang, J. (2019), *Introduction to probability*, Texts in Statistical Science Series, CRC Press, Boca Raton, FL. Second edition.

Feller, W. (1957), *An introduction to probability theory and its applications. Vol. I*, John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London. 2nd ed.

Grimmett, G. & Welsh, D. (1986), *Probability: an introduction*, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York.

Hájek, A. (2012), Interpretations of probability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2012 edn, Metaphysics Research Lab, Stanford University.

Proschan, M. A. & Shaw, P. A. (2016), *Essentials of probability theory for statisticians*, Chapman & Hall/CRC Texts in Statistical Science Series, CRC Press, Boca Raton, FL.

Ross, S. (2014), *A first course in probability*, ninth edn, Macmillan Co., New York; Collier Macmillan Ltd., London.

Spiegelhalter, D., Pearson, M. & Short, I. (2011), 'Visualizing uncertainty about the future', *Science* **333**(6048), 1393–1400.