Introduction to Probability & Statistics
Lecture notes 21/22

Dr Sam Crawford & Dr Jess Hargreaves
(Based on notes by Dr Gustav W Delius)
sam.crawford@york.ac.uk
jess.hargreaves@york.ac.uk

Contents

Image of dice on front page obtained from http://en.wikipedia.org/wiki/File:Dice.jpg

1 Introduction

Welcome to the lecture notes for *Introduction to Probability and Statistics*! One thing you may have already noticed is that the chapter numbers aren't always consecutive. This is because each chapter in these notes corresponds to the chapter of the course textbook[1] with the same number. Note that **questions on Moodle will refer to the course textbook**, in addition to these notes, so always make sure you have both available when studying for this course. You can access a digital copy of the textbook at this link, or via the Moodle page.

This is version 1 of the lecture notes, which is mostly unchanged from the previous year. As we progress through the course a slightly updated version of these notes will be released. However, the changes will be relatively minor, so please feel free to read ahead through these notes if you wish.

[1] Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P. and Meester, L.E. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How.* London: Springer.

2 Outcomes, events, and probability

(Textbook chapter link)

We now begin to explore the wonderful world of probability, showing how a few simple definitions and rules (or, as mathematicians call them, *axioms*) can lead to a beautiful and often rather subtle subject...

2.1 Sample spaces

Suppose that we are to perform some sort of experiment, whose outcome is not deterministic (*i.e.* there is some degree of randomness involved).

Definition 2.1. An **experiment** is anything with a set of possible **outcomes**. The set Ω of all possible outcomes of an experiment is the **sample space** of the experiment.

Example 2.2. If the experiment involves tossing a coin and recording the outcome, then the sample space is simply

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

where we have denoted the outcome that the coin lands Heads up by H and the outcome that it lands Tails up by T. If the experiment involves tossing a coin twice, then the sample space would be

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\} \ ,$$

(where *e.g.* (H, T) means that the first result was Heads and the second one Tails).

Example 2.3. If the experiment consists of drawing balls from a bag containing balls numbered $1, \ldots, n$ until the bag is empty, then

$$\Omega = \{\text{all permutations of the set } (1, \ldots, n)\} \ .$$

For example when $n = 3$,

$$\Omega = \{(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)\}.$$

Notation: If E is a set then we write $|E|$ to denote the **size** of the set, also known as its cardinality. (At school you may have used the notation $n(E)$.) Thus in Example 2.2, $|\Omega| = 4$, whilst in Example 2.3 $|\Omega| = n!$.

Example 2.4. Consider throwing a dart at a dartboard of radius r and measuring its distance d from the centre. Then the sample space consists of all distances between 0 and r, i.e.,

$$\Omega = [0, r] = \{d \,|\, 0 \leq d \leq r\}.$$

This is an infinite (in fact, uncountably infinite) sample space, $|\Omega| > |\mathbb{N}|$.

2.2 Events

An event is a set containing a number of possible outcomes of the experiment. The event is said to have occurred if the realised outcome is among the outcomes contained in the event.

Example 2.5. Let Ω be the sample space for tossing a coin twice, as in Example 2.2. The event that exactly one result is Heads is the set $\{(H,T),(T,H)\}$. The event that the second is Heads is the set $\{(H,H),(T,H)\}$. The event that both results are Heads is the set $\{(H,H)\}$.

Always remember that events are sets. Therefore the event that both results are Heads is $\{(H,H)\}$ and not just (H,H).

An event E is always a subset of the sample space Ω. We write this as $E \subseteq \Omega$. This notation nicely indicates that a subset can also be the entire set. So $\Omega \subseteq \Omega$; it is the sure event because it contains all possible outcomes. The empty set, which we denote by $\emptyset$, is a subset of any set, so in particular $\emptyset \subset \Omega$; it is the impossible event.

We can create new, more complicated events from simple ones using basic set operations. For any two events A and B we define:

- $A \cup B$ to be the event consisting of all outcomes that belong to either A, or B, or both A and B – this is known as the **union** of A and B;
- $A \cap B$ to be the event consisting of all outcomes that belong to *both* A and B – this is known as the **intersection** of A and B.
- A^c to be the **complement** of A, that is, A^c contains all outcomes which are not members of A.

A good way to visualise set operations is provided by Venn diagrams. You can find Venn diagrams illustrating the operations of intersection, union and complement in Figure 2.1 below.

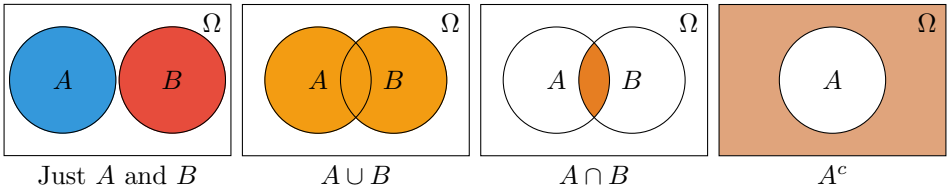


Figure 2.1: Venn diagrams illustrating the union, intersection and complements of sets A and B

Example 2.6. In the setting of Example 2.3, let $n = 3$ and let A denote the event that ball number 2 is drawn out of the bag first. Let B be the event that ball number 3 is drawn second. Then

$$\begin{aligned} A &= \{(2,1,3),(2,3,1)\} \\ B &= \{(1,3,2),(2,3,1)\} \\ A \cup B &= \{(1,3,2),(2,1,3),(2,3,1)\} \\ A \cap B &= \{(2,3,1)\} \\ A^c &= \{(1,2,3),(1,3,2),(3,1,2),(3,2,1)\}\,. \end{aligned}$$

Note that $A \cup B$ is the event that *at least one* of A and B occurs, whilst $A \cap B$ is the event that *both* A and B occur. A^c is the event that A does not occur.

Note that, since Ω is the set of all possible outcomes of the experiment, $\Omega^c = \emptyset$ (the empty set). Also note that for any two events A and B,

$$\begin{aligned} &\emptyset \subseteq A \cap B \subseteq A \subseteq A \cup B \subseteq \Omega \\ \text{and so} \quad &0 \leq |A \cap B| \leq |A| \leq |A \cup B| \leq |\Omega|\ . \end{aligned}$$

We also define unions and intersections of more than two events in the same way: for example if E_1, E_2, E_3 are events then $E_1 \cup E_2 \cup E_3$ is the event that at least one of the events E_1, E_2 or E_3 occur. This works for any countable collection of events.

A convenient way to write expressions involving an arbitrary number of events is to introduce an index set I and to write the collection of events as $\{E_i | i \in I\}$. For example a set of three events E_1, E_2, E_3 could be written as $\{E_i | i \in I\}$ with $I = \{1,2,3\}$ and we could write $E_1 \cup E_2 \cup E_3$ as $\bigcup_{i \in I} E_i$. Thus

- $\bigcup_{i \in I} E_i$ is the event that consists of all outcomes that belong to at least one of the events in the collection;
- $\bigcap_{i \in I} E_i$ is the event that consists of all outcomes that belong to all of the events in the collection.

An infinite countable collection $E_1, E_2, \ldots$ could be written as $\{E_i | i \in \mathbb{N}\}$. In that case there is an alternative notation:

$$\bigcup_{i \in \mathbb{N}} E_i = \bigcup_{i=1}^{\infty} E_i.$$

Proposition 2.7 (De Morgan's laws)**.** *For any countable collection of events* $\{E_i | i \in I\}$*,*

$$\left(\bigcup_{i\in I} E_i\right)^c = \bigcap_{i\in I} {E_i}^c$$
$$\left(\bigcap_{i\in I} E_i\right)^c = \bigcup_{i\in I} {E_i}^c .$$

Proof. You will discuss the proof of this more in Core Algebra. Therefore you can skip the proof below for now, if you like. We also prove only the first of these equalities, with the proof of the second being similar.

Suppose that ω is an outcome belonging to $\left(\bigcup_{i\in I} E_i\right)^c$. This means that ω is not contained in any of the events E_i, and so it must belong to ${E_i}^c$ for all $i \in I$. Thus $\omega \in \bigcap_{i\in I} E_i^c$.

Conversely, suppose that $\omega \in \bigcap_{i\in I} {E_i}^c$. Then ω belongs to all of the ${E_i}^c$, and so does not belong to any of the E_i. This implies that it does not belong to $\left(\bigcup_{i\in I} E_i\right)$, as required. Since we have shown that any outcome belonging to the set on the left hand side belongs to that on the right, and vice versa, the two events must be equal, as required. □

Often we only need a simplified version of De Morgan's law for only two events, A and B. We obtain this by setting $I = \{1, 2\}$ and $E_1 = A, E_2 = B$ in Proposition 2.7:

Corollary 2.8. *Let* A, B *be events. Then*

$$(A \cup B)^c = A^c \cap B^c \text{ and } (A \cap B)^c = A^c \cup B^c.$$

We have done something amazing. We have taken the vague concept of an "event" and have made it into a mathematical object, namely a set. This has allowed us to combine events to form new events via the set operations of union, intersection and complement. Thus we can calculate with events similar with how we can calculate with numbers, where we can form new numbers via the operations of addition and multiplication. Like addition and multiplication, the operations of union and intersection satisfy the properties of commutativity, associativity, distributivity:

Proposition 2.9. *The following rules for unions and intersections hold for all events*

E, F, G:

Commutativity:	$E \cup F = F \cup E$
	$E \cap F = F \cap E$
Associativity:	$(E \cup F) \cup G = E \cup (F \cup G)$
	$(E \cap F) \cap G = E \cap (F \cap G)$
Distributivity:	$(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$
	$(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

Proof. Simply show that any element belonging to the set on the left hand side of an equation belongs to the set on the right, and vice-versa, as in the proof of Proposition 2.7. □

2.2.1 Event Spaces and σ-Algebras

In your Algebra module you will meet the concept of a number field as a set of numbers that is closed under the operations of addition, subtraction, multiplication and division. In exactly the same vein we now define a σ-algebra as a set that is closed under union, intersection and complement.

Definition 2.10. A **σ-algebra** $\mathcal{F}$ on a sample space Ω is a set of subsets of Ω such that

1. $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$,
2. If $E \in \mathcal{F}$ then $E^c \in \mathcal{F}$,
3. If $\{E_i | i \in I\}$ is a countable subset of $\mathcal{F}$ then $\bigcup_{i \in I} E_i \in \mathcal{F}$ and $\bigcap_{i \in I} E_i \in \mathcal{F}$.

We can now say that the set of events for a probability experiment is a σ-algebra. Sometimes the name **event space** is used for this σ-algebra. This summarises our above observations that Ω is an event, namely the certain event, that $\emptyset$ is an event, namely the impossible event, the complement of an event is an event, and the countable unions and intersections of events are events. Thus we can now work with events in the same mathematical way in which we work with numbers. This gives you a good first feel for how modern mathematics operates.

The power set $\mathcal{P}(\Omega)$, the set of all subsets of Ω, is a σ-algebra for Ω. This is the σ-algebra we will use in most of our examples involving a finite set of outcomes. However in cases where the sample space is uncountably infinite (for example an interval of the real line), the power set would be too large and one has to work with a smaller σ-algebra.

Finally, we will need the following definition:

Definition 2.11. A collection $\{E_i | i \in I\}$ of events is called **disjoint** or **mutually exclusive** if no two events share a common element, *i.e.* if $E_i \cap E_j = \emptyset$ for all $i \neq j \in I$.

For example the sets $\{1, 2\}$ and $\{3, 4\}$ are disjoint but the sets $\{1, 2\}, \{3, 4\}, \{4, 5\}$ are not because the second two sets share the element 4.

We have now introduced all the set theory we will need in this module. When we write proofs in this module, we will freely use results from set theory without proving them here.

2.3 Probability

Now that we are happy with the idea of the sample space for an experiment, and with the concept of events, we'd like to be able to say something about how likely various events are: this is what probability is all about! We want to assign a probability to each event. Thinking about a few examples one realises that one cannot assign probabilities arbitrarily.

Let's think about the football match example. There is a football match between Leeds United and Liverpool. From the point of view of Leeds United, there are three possible outcomes: Win, Lose or Draw. Hence, the sample space of this probability experiment is $\Omega = \{\text{Win}, \text{Draw}, \text{Lose}\} = \{\text{W}, \text{D}, \text{L}\}$. You may object that it feels wrong to call a football match a "probability experiment", but that is the language mathematicians use for anything that has a set of possible outcomes. To the event $\{\text{W}\}$ I could assign a probability of 0.6. Clearly that is a very subjective statement. You could assign a probability of 0.1 and I could not prove you wrong. However if I were to also to assign a probability of 0.9 to the event $\{\text{D}\}$, then you could tell me that I am wrong. Why? Because then the probability of the event that Leeds United either win or draw, would be $0.6 + 0.9 = 1.5$ and that is not allowed because probabilities lie between 0 and 1. If I were to assign a probability of 0.6 to the event $\{\text{W}\}$, a probability of 0.2 to the event $\{\text{D}\}$ and a probability of 0.1 to the event $\{\text{L}\}$ you would again object, because all those probabilities add up to 0.9 whereas they should add up to 1.

We clearly have some intuitive feeling for what kinds of rules probabilities have to satisfy in this example. The challenge is to formalise these rules. This was achieved by Kolmogorov in 1933.

Definition 2.12 (Axioms of probability)**.** Let Ω be the sample space of an experiment and $\mathcal{F}$ a σ-algebra of events. A **probability function** P assigns to each event $E \in \mathcal{F}$ a real number $P(E)$ such that

(P1) $P(E) \in [0, 1]$;

(P2) $P(\Omega) = 1$;

(P3) If $\{E_i | i \in I\}$ is a countable disjoint collection of events then

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

The number $P(E)$ is called the **probability** that E occurs. The triple $(\Omega, \mathcal{F}, P)$ is a **probability space**.

In words, (P1) says that a probability is always between 0 and 1, (P2) says that the certain event has probability 1, and (P3) says that the probability of the union of disjoint events is the sum of the probabilities of the individual events. Note that (P3) contains the special case that

$$P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset.$$

The next theorem gives some rules that may be inferred from the above axioms of probability.

Theorem 2.13. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then for any events $A, B \in \mathcal{F}$*

(P4) $P(A^c) = 1 - P(A)$;

(P5) $P(\emptyset) = 0$;

(P6) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(P7) *If $A \subseteq B$ then $P(A) \leq P(B)$.*

Proof. For property (P4) we start with the observation that $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$, so

$$1 \overset{(P2)}{=} P(\Omega) = P(A \cup A^c) \overset{(P3)}{=} P(A) + P(A^c).$$

We can now solve this equation for $P(A^c)$. This gives

$$P(A^c) = 1 - P(A) .$$

Property (P5) follows directly from (P4) since $\emptyset = \Omega^c$ and $P(\Omega) = 1$ by (P2).

For property (P6), writing A as the *disjoint* union

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

we obtain

$$P(A) = P((A \cap B^c) \cup (A \cap B)) \overset{(P3)}{=} P(A \cap B^c) + P(A \cap B) .$$

Similarly, we can write $A \cup B$ as the *disjoint* union

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

which gives

$$P(A \cup B) \overset{(P3)}{=} P(B) + P(A \cap B^c) .$$

Eliminating $P(A \cap B^c)$ from these two equations gives

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

For property (P7) we observe that since $A \subseteq B$, we can write B as the *disjoint* union $B = (B \cap A^c) \cup A$. This gives

$$P(B) = P((B \cap A^c) \cup A) \overset{(P3)}{=} P(B \cap A^c) + P(A).$$

Since $P(B \cap A^c) \geq 0$ by axiom (P1), it follows that $P(B) \geq P(A)$. □

Note the importance in this proof of writing unions as *disjoint* unions, so that we can apply axiom (P3).

Now that we have defined what properties a probability function must have, we should give some examples of probability functions.

Example 2.14. The probability space of the football match example is $(\Omega, \mathcal{F}, P)$, where Ω ={Win, Draw, Lose} = {W,D,L}. As a σ-algebra $\mathcal{F}$ of events we choose the power set of Ω. So $\mathcal{F}$ = {$\emptyset$, {W},{D},{L},{W,D},{W,L}, {D,L},Ω}. The probability function P assigns to each event in $\mathcal{F}$ a real number. A possible assignment is as follows:

Event	Probability
$\emptyset$	0
{W}	0.6
{D}	0.1
{L}	0.3
{W,D}	0.7
{W,L}	0.9
{D,L}	0.4
Ω	1

Note how there is no choice about the first and last line in the table. Also, once the probability of {W} is specified, the probability of $\{D, L\}$ is fixed as well. And once the probability of one more of the elementary events is specified, the probability of all other events is fixed.

At this point you would be excused for saying that you could understand this example without all the heavy machinery of probability spaces, and you would be right. However our heavy machinery is much more general and already the next example may give you a hint for why mathematical rigour might be important.

Example 2.15. Let us revisit the experiment from Example 2.4 where we observed the distance from the centre of a dartboard at which a dart lands. The sample space is $\Omega = [0, 1]$ where we have chosen the radius of the dartboard to be $r = 1$.

Next we need to choose an event space[2]. It is natural to want to include all subintervals of $[0, 1]$ in the event space. The event space then also needs to contain all other subsets of $[0, 1]$ that can be obtained by taking countable unions, intersections and

[2]You may ask why we do not simply use the power set of the sample space, i.e., all subsets of the interval $[0, 1]$, as the set of events. It turns out that if we did that, there would exist no probability function!. So we settle for a slightly smaller event space. Luckily it turns out that it is quite difficult to find examples of subsets of $[0, 1]$ that are not in the Borel algebra (see Wikipedia article for an example) so this is a very suitable event space.

complements because otherwise it would not be a σ-algebra. The smallest σ-algebra containing all intervals is known as the Borel algebra, so we choose $\mathcal{F}$ to be the Borel algebra on $[0, 1]$.

A probability function is now uniquely determined by specifying the probability of closed intervals. We can for example choose

$$P([a, b]) = b - a.$$

This leads to the probability function known as the Borel measure.

The interesting bit of this example is that the probability that the dart lands at a particular distance d is zero for any d: $P(\{d\}) = P([d, d]) = P(d - d) = 0$. This does of course not mean that it is impossible for the dart to land at a certain distance d. In fact, it must land at some distance, even though the probability of landing at any distance is zero!

The important lesson is that an event that has probability zero does not have to be the impossible event. Therefore also $P(A \cap B) = 0$ does not imply that A and B are disjoint or mutually exclusive. Neither does $P(A) = 1$ imply that $A = \Omega$.

Often one has probability experiments where the sample space contains a finite number n of outcomes and all n elementary events have the same probability of $1/n$. Examples include flipping of a coin ($n = 2$), throwing of a die ($n = 6$), picking a card at random from a standard deck of cards ($n = 52$). In these cases, the computation of probabilities reduces to being able to count the number of outcomes in the event of interest. We deal with all such examples at once in the following theorem.

Theorem 2.16. *Suppose the sample space Ω contains exactly n outcomes, $|\Omega| = n$. Let the event space $\mathcal{F}$ contain all subsets of Ω. Set*

$$P(E) = \frac{|E|}{n}$$

for any event $E \in \mathcal{F}$. Then P is a probability function.

Proof. We need to show that P satisfies the probability axioms. The number of elements of any event E must be between 0 and n and it therefore follows that $P(E) = |E|/n$ lies between 0 and 1, as required by (P1). Since $|\Omega| = n$, $P(\Omega) = 1$ and so (P2) is also satisfied. Since there is only a finite number of subsets of Ω (in fact exactly 2^n), we only need to show (P3) for every finite union of disjoint sets. For disjoint sets $E_1, \ldots, E_k$, $|E_1 \cup \cdots \cup E_k| = |E_1| + \cdots + |E_k|$. Hence,

$$\begin{aligned} P(E_1 \cup \cdots \cup E_k) &= \frac{|E_1 \cup \cdots \cup E_k|}{n} = \frac{|E_1| + \cdots + |E_k|}{n} \\ &= \frac{|E_1|}{n} + \cdots + \frac{|E_k|}{n} \\ &= P(E_1) + \cdots + P(E_k)\,, \end{aligned}$$

and so (P3) holds. Thus, P is indeed a probability function. □

Of course we have already seen examples where the different possible outcomes are not equally likely. These are not covered by the above theorem. Here is another such example.

Example 2.17. Consider an experiment consisting of drawing one ball at random from a bag containing four red, six green and three blue balls. The set of possible outcomes of the experiment is $\Omega = \{\text{red, green, blue}\} = \{r, g, b\}$. The probabilities of these outcomes are $p_r = P(\{\text{red}\}) = 4/13$, $p_g = P(\{\text{green}\}) = 6/13$, $p_b = P(\{\text{blue}\}) = 3/13$. The probability for all events can then be calculated as

Event	P(Event)
$\emptyset$	0
{red}	$p_r = \frac{4}{13}$
{green}	$p_g = \frac{6}{13}$
{blue}	$p_b = \frac{3}{13}$
{red,green}	$p_r + p_g = \frac{10}{13}$
{red,blue}	$p_r + p_b = \frac{7}{13}$
{green,blue}	$p_g + p_b = \frac{9}{13}$
{red,green,blue}$= \Omega$	$p_r + p_g + p_b = 1$

We have seen this pattern before: In the case of a countable sample space, once the probabilities of the elementary events (the events containing only a single outcome) have been assigned, all other probabilities can be deduced from those. The only restriction placed on the probabilities of the elementary events is that they need to sum up to 1. We'll formulate this as a theorem:

Theorem 2.18. *Let Ω be a countable sample space. Let the event space $\mathcal{F}$ be the power set of Ω. Choose a set $\{p_\omega | \omega \in \Omega\}$ of non-negative real numbers satisfying $\sum_{\omega\in\Omega} p_\omega = 1$. For any event $E \in \mathcal{F}$ define*

$$P(E) = \sum_{\omega\in E} p_\omega.$$

Then P is a probability function on $\mathcal{F}$ and thus $(\Omega, \mathcal{F}, P)$ is a probability space.

Proof. We need to check the conditions (P1), (P2) and (P3).

(P1) Because all probabilities are sums of elementary probabilities p_ω and all p_ω are non-negative and their total sum is 1, any probability is non-negative and no larger than 1.

(P2) We have $P(\Omega) = \sum_{\omega\in\Omega} p_\omega = 1$.

(P3) This uses the fact that the sum over a disjoint union of index sets can be split up into a sum over individual sums:

$$P\left(\bigcup_{i\in I} E_i\right) = \sum_{\omega\in\bigcup_{i\in I} E_i} p_\omega = \sum_{i\in I}\sum_{\omega\in E_i} p_\omega = \sum_{i\in I} P(E_i)\,.$$

□

In the special case where all the elementary probabilities p_ω are equal, this theorem reduces to Theorem 2.16.

2.4 Products of sample spaces

Suppose we perform two experiments with probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$. Then the possible outcomes of the combined experiment can be described by ordered pairs of outcomes of the individual experiments and thus the combined sample space is

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) | \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}.$$

We have already seen a sample space of this form in Example 2.2 where we discussed the experiment involving tossing a coin twice. That experiment can be seen a combination of two experiments, each consisting of one flip of the coin. The individual sample spaces are $\Omega_1 = \{H, T\} = \Omega_2$ and thus the combined sample space is

$$\Omega = \Omega_1 \times \Omega_2 = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Note that $|\Omega| = |\Omega_1| \cdot |\Omega_2|$.

If Ω_1 and Ω_2 are countable then so is Ω and we can choose as our event space $\mathcal{F}$ the power set of Ω. We can define a probability function on $\mathcal{F}$ by setting

$$P(\{(\omega_1, \omega_2)\}) = P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \quad \text{for all } (\omega_1, \omega_2) \in \Omega.$$

The probabilities for composite events are then determined as in Theorem 2.18. The theorem tells us that the $(\Omega, \mathcal{F}, P)$ that we have constructed in this way for the combined experiment is a valid probability space, provided the sum of the probabilities of all the elementary events is equal to 1. Let us check that this is indeed the case.

$$\begin{aligned}
\sum_{(\omega_1,\omega_2)\in\Omega} P(\{(\omega_1, \omega_2)\}) &= \sum_{\omega_1\in\Omega_1} \sum_{\omega_2\in\Omega_2} P(\{(\omega_1, \omega_2)\}) \\
&= \sum_{\omega_1\in\Omega_1} \sum_{\omega_2\in\Omega_2} P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \\
&= \sum_{\omega_1\in\Omega_1} P_1(\{\omega_1\}) \cdot \sum_{\omega_2\in\Omega_2} P_2(\{\omega_2\}) \\
&= P_1(\Omega_1) \cdot P_2(\Omega_2) = 1 \cdot 1 = 1.
\end{aligned}$$

In the first equality we used that summing over all pairs of outcomes is the same as summing over all possible outcomes of one experiment and then summing that over all outcomes of the other experiment. In the third equality we pulled a factor that does not depend on ω_2 out of the sum over ω_2.

This probability function on the joint sample space is the appropriate choice when the two experiments are completely independent in the sense that the realised outcome of one experiment has no influence on the other. It is important not to use this product

rule in other cases. We will discuss in the next chapter how to deal with the case where the outcomes are not independent.

The above construction of a probability space for the combination of two experiments can be generalised to n experiments. One has the sample space

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n = \{(\omega_1, \omega_2, \ldots, \omega_n) | \omega_i \in \Omega_i, i = 1, \ldots, n\}.$$

When all experiments are independent of each other, we use the probability function that is specified through

$$P(\{(\omega_1, \omega_2, \ldots, \omega_n)\} = P_1(\{\omega_1\}) \cdot P_2(\{\omega_2\}) \cdot \cdots \cdot P_n(\{\omega_n\}).$$

Example 2.19 (Chevalier de Méré)**.** Consider a game consisting of throwing a pair of dice 24 times in which you win if there is at least one double six among the 24 throws and you lose otherwise. What is the probability of winning?

Solution. The sample space for a single throw of the pair of dice contains 36 outcomes. They all have equal probability. Therefore the probability of rolling a double six in one round is $P(\{(6,6)\}) = 1/36$. The probability of NOT rolling a double six is therefore $P(\{(6,6)\}^c) = 1 - 1/36 = 35/36$. Because the individual throws are independent, the probability of losing the game, i.e. of in each of the 24 rounds not rolling a double six, is the product of the probability of not rolling a double six in the first round times the probability of not rolling a double six in the second round times ...

$$P(L) = P_1(\{(6,6)\}^c) \cdot P_2(\{(6,6)\}^c) \cdot \cdots \cdot P_{24}(\{(6,6)\}^c) = \left(\frac{35}{36}\right)^{24}.$$

The probability of winning is then

$$P(W) = 1 - P(L) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.49.$$

3 Conditional Probability and Independence

(Textbook chapter link)

3.1 Conditional probability

Example 3.1. Consider a game: you get to roll a fair die twice and if the total score is higher than 9 you win £6, otherwise you lose £1. The probability of you winning is the probability of getting (4,6) or (6,4) or (5,5) or (5,6) or (6,5) or (6,6) *i.e.* $\frac{6}{36} = \frac{1}{6}$ (there are a total of 36 possible outcomes). Now suppose that the first roll of the die gives you a 6. What is your probability of winning now? The number of possible scores has been drastically reduced from 36 to 6, and the winning outcomes are now (6,4) or (6,5) or (6,6). Thus, you have probability $\frac{1}{2}$ of winning. What happened is that the additional information has shrunk the set of elementary events that has non-zero probability. Before the first die stopped at six, the sample space contains these elements:

$$\begin{array}{cccccc}
(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\
(4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\
(5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\
(6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)
\end{array}$$

whereas after the first die stops at six the sample space consists only of the last row. Let A be the event of you winning the bet and B the event of the first die stopping at six. Then, before the extra information, $P(A) = \frac{1}{6}$, $P(B) = \frac{1}{6}$ and $P(A \cap B) = P(\{6,4\} \cup \{6,5\} \cup \{6,6\}) = \frac{3}{36}$. After the additional information we have a new probability, call it P_B (to be defined shortly), such that $P_B(A) = \frac{1}{2}$ and $P_B(B) = 1$. Note that $\frac{P(A \cap B)}{P(B)} = \frac{3/36}{1/6} = \frac{1}{2}$ and $\frac{P(B \cap B)}{P(B)} = \frac{1/6}{1/6} = 1$; *i.e.* $P_B(A) = \frac{P(A \cap B)}{P(B)}$ and $P_B(B) = \frac{P(B \cap B)}{P(B)}$.

Definition 3.2. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let B be any event such that $P(B) > 0$. For any event A the **conditional probability** of A given B is

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

We will from now on drop the "Let $(\Omega, \mathcal{F}, P)$ be a probability space" at the start of our definitions and theorems. When we use the symbols Ω, $\mathcal{F}$ or P from now on you can always assume that they are the sample space, the event space and the probability function of a probability space.

We now do something amazing that will allow us to reuse all the results from last week in a new context. We prove that the conditional probability also satisfies the axioms of probability and hence automatically all theorems we have proved for probability automatically hold for conditional probability. This gives us a feel for why the axiomatic approach to mathematics is so powerful.

Theorem 3.3. *For any event B with $P(B) > 0$ the function P_B defined by*

$$P_B(A) = P(A|B)$$

for any event A is a probability function on $\mathcal{F}$, so that $(\Omega, \mathcal{F}, P_B)$ is a probability space.

Proof. We need to show that P_B satisfies (P1), (P2) and (P3), using that P does.

(P1) is easy: clearly $P_B(A) \geq 0$ for all A, and since $A \cap B \subseteq B$, property (P7) ensures that $P(A \cap B) \leq P(B)$ and hence that $P_B(A) \leq 1$.

(P2) is also easy:

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

For (P3), if $\{E_i | i \in I\}$ is a countable collection of disjoint events, then the collection $\{E_i \cap B | i \in I\}$ is also disjoint. Thus using (P3) for the probability P,

$$\begin{aligned} P_B\left(\bigcup_{i\in I} E_i\right) &= \frac{P\left(\left(\bigcup_{i\in I} E_i\right) \cap B\right)}{P(B)} = \frac{P\left(\left(\bigcup_{i\in I} E_i \cap B\right)\right)}{P(B)} \\ &= \frac{\sum_{i\in I} P(E_i \cap B)}{P(B)} = \sum_{i\in I} P_B(E_i)\,. \end{aligned}$$

□

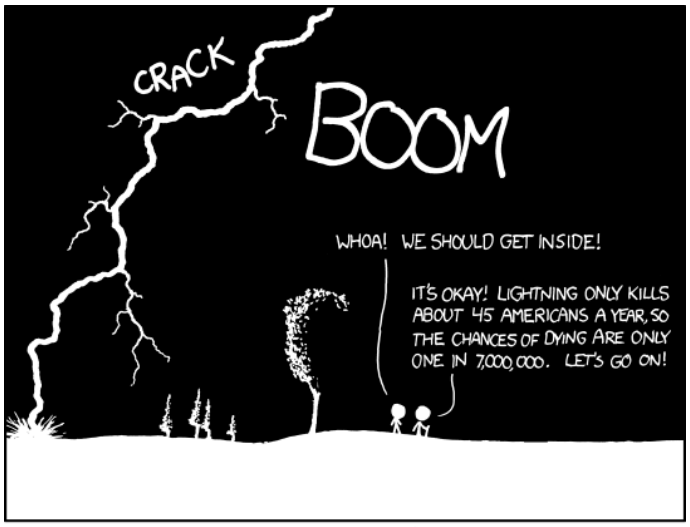
Example 3.4. A certain game involves tossing two fair coins: a player wins if they get at least one tail. Given that Peter played the game and won, what is the probability that he got two tails?

Solution. At first glance you might think that this probability is 1/2, since "the other coin was equally likely to be a head or a tail". However, this is not the correct answer! At the start of the game, the sample space is given by

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

Let $B = \{(H,T), (T,H), (T,T)\}$ be the event that Peter wins the game, and $A = \{(T,T)\}$ be the event that he gets two Tails. Then $P(B) = 3/4$ and $P(A \cap B) = 1/4$. Thus, conditioning on event B, we see that

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



(http://xkcd.com/795)

3.2 The multiplication rule

Example 3.5. Consider drawing two balls out of a bag (without replacement) containing 8 white and 4 red balls. What is the probability that both balls are red?

Solution. Let R_1 be the event that the first ball is red and R_2 the event that the second ball is red. The question asks for $P(R_1 \cap R_2)$. It is easy to determine that $P(R_1) = 1/3$ because all balls are equally likely to be picked and one third of the balls are red. It is less easy to determine $P(R_2)$ because we do not know whether by the time the second ball is picked there are still 4 or only 3 red balls still in the bag, because that depends on the outcome of the first draw. We will return to that at the start of the next section. We can however determine the conditional probability $P(R_2|R_1)$ that the second ball is red given that the first one was already red, because given that information we now know that there are 3 red balls left and there are 11 balls left altogether, so $P(R_2|R_1) = 3/11$. We now recall the definition of the conditional probability,

$$P(R_2|R_1) = \frac{P(R_2 \cap R_1)}{P(R_1)},$$

and solve that for $P(R_2 \cap R_1)$ giving

$$P(R_2 \cap R_1) = P(R_2|R_1)P(R_1) = \frac{3}{11}\frac{1}{3} = \frac{1}{11}.$$

So let us formulate this useful formula as a theorem.

Theorem 3.6 (Multiplication rule)**.** *Let A, B be events with $P(B) > 0$. Then*

$$P(A \cap B) = P(A|B)P(B).$$

Proof.

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

□

3.3 The law of total probability and Bayes' rule

Example 3.5 continued: We now address the question left open in the last section of what is the probability that the second ball is red, $P(R_2)$. We use the trick of splitting up the event R_2 into the union of disjoint events. Because $R_2 = (R_2 \cap R_1) \cup (R_2 \cap R_1^c)$ and $(R_2 \cap R_1) \cap (R_2 \cap R_1^c) = \emptyset$, we can write

$$\begin{aligned} P(R_2) &= P\left((R_2 \cap R_1) \cup (R_2 \cap R_1^c)\right) \\ &\overset{(P3)}{=} P(R_2 \cap R_1) + P(R_2 \cap R_1^c) \\ &\overset{Thm3.6}{=} P(R_2 \mid R_1)\, P(R_1) + P(R_2 \mid R_1^c)\, P(R_1^c). \end{aligned}$$

All the probabilities on the right-hand side are easy to calculate. We already dealt with the first term. For the second term we have $P(R_2|R_1^c) = 4/11$, because if the first ball was not red there are still 4 red balls among the 11 balls in the bag when the second ball is drawn, and $P(R_1^c) = 1 - P(R_1) = 2/3$. Substituting this into the above equation gives

$$P(R_2) = \frac{3}{11}\frac{1}{3} + \frac{4}{11}\frac{2}{3} = \frac{1}{3}.$$

(A student once pointed out after a lecture that it is curious how $P(R_1) = P(R_2)$. It would be an interesting exercise for you to investigate whether that continues to hold when the numbers of balls in the bag are chosen differently.)

In this example we only had to consider two alternatives: either the first ball was red or it was not red. But in general there could of course be any number of alternatives.

Example 3.7. Assume 60% of consumers have a mobile phone from manufacturer A, 30% from manufacturer B and 10% from C. The probability of a phone from manufacture A to spontaneously go up into flames is 1%, for B it is 2% and for C it is 3%. What is the probability that a random customer's phone goes up in flames?

Solution. We introduce the event M_A that the random customer has a phone from manufacture A and similarly we introduce events M_B and M_C. Note that our events

M_A, M_B, M_C are disjoint

$$M_A \cap M_B = M_A \cap M_C = M_B \cap M_C = \emptyset$$

and together cover all of Ω,

$$\Omega = M_A \cup M_B \cup M_C.$$

We also introduce the event F that the customer's phone goes up in flames. We can now translate the information from the problem statement into formulas as

$$P(M_A) = 0.6, \quad P(M_B) = 0.3, \quad P(M_C) = 0.1$$

and

$$P(F|M_A) = 0.01, \quad P(F|M_B) = 0.02, \quad P(F|M_C) = 0.03.$$

We can now split the event F whose probability we want to calculate into three disjoint events,

$$F = (F \cap M_A) \cup (F \cap M_B) \cup (F \cap M_C)$$

and therefore by axiom (P3),

$$\begin{aligned} P(F) &= P((F \cap M_A) \cup (F \cap M_B) \cup (F \cap M_C)) \\ &\overset{(P3)}{=} P(F \cap M_A) + P(F \cap M_B) + P(F \cap M_C) \\ &\overset{Thm\ 3.7}{=} P(F|M_A)P(M_A) + P(F|M_B)P(M_B) + P(F|M_C)P(M_C) \\ &= 0.01 \cdot 0.6 + 0.02 \cdot 0.3 + 0.03 \cdot 0.1 \\ &= 0.015. \end{aligned}$$

So the probability of the phone of a random customer bursting into flames is 1.5%.

You notice that the equations already got a bit long for three alternative events and this would get worse when we consider examples with a larger number of alternatives. We can shorten this with the index set notation we introduced in Section 2.2. In the above example we would introduce the index set $I = \{A, B, C\}$. Then we can rewrite the above equation as

$$M_i \cap M_j = \emptyset \text{ for all } i, j \in I,$$

$$\Omega = \bigcup_{i \in I} M_i,$$

$$F = \bigcup_{i \in I} F \cap M_i,$$

and

$$P(F) = \sum_{i \in I} P(F|M_i)P(M_i).$$

We now use this notation to formalise the idea of splitting the sample space into a union of disjoint sets.

Definition 3.8. We say that a countable collection of events $\{B_i | i \in I\}$ is a **partition** of Ω if it is disjoint and $\Omega = \bigcup_{i \in I} B_i$.

The idea of splitting a difficult probability problem up into simpler ones by partitioning the sample space in such a way that the conditional probability in each component of the partition is easy to calculate is so useful that it has not only one name but two:

Theorem 3.9 (Law of Total Probability or Partition Theorem)**.** *Let $\{B_i | i \in I\}$ be a partition of sample space Ω such that $P(B_i) > 0$, for each $i \in I$. Then*

$$P(A) = \sum_{i \in I} P(A \mid B_i) P(B_i) \quad \text{for all events } A. \tag{3.1}$$

Proof. Since $P(\cup_{i \in I} B_i) = P(\Omega) = 1$, we have that for each event A

$$\begin{aligned} P(A) &= P\left(A \cap \left(\bigcup_{i \in I} B_i\right)\right) = P\left(\bigcup_{i \in I} (A \cap B_i)\right) = \sum_{i \in I} P(A \cap B_i) \\ &= \sum_{i \in I} P(A \mid B_i) P(B_i) \end{aligned}$$

where the second-to-last equality follows from (P3) (remember the B_i are disjoint and so $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ if $i \neq j$) and the last equality follows from Thm 3.6. □

Here is another example of the use of the partition theorem that we skipped in the lecture:

Example 3.10. On Thursday night you can either stay in or go to Fibbers. The probability of you going to Fibbers is $\frac{4}{5}$. If you choose to stay in, the probability of you staying awake during your Calculus lecture on Friday is $\frac{9}{10}$, while the corresponding probability if you choose to go to Fibbers is $\frac{1}{5}$. What is the probability of staying awake during the Calculus lecture on Friday?

Solution. Let A be the event that you stay awake during the lecture and F the event that you go to Fibbers. The events F and F^c form a partition of Ω because $F \cap F^c = \emptyset$ and $F \cup F^c = \Omega$. Applying the Law of Total Probability we obtain

$$\begin{aligned} P(A) &= P(A \mid F) P(F) + P(A \mid F^c) P(F^c) \\ &= \left(\frac{1}{5}\right)\left(\frac{4}{5}\right) + \left(\frac{9}{10}\right)\left(\frac{1}{5}\right) = \frac{17}{50} = 0.34. \end{aligned}$$

Example 3.11 (Monty Hall Problem)**.** In a particular game show, you have to choose one of three doors. Behind one door there's a car, behind each of the other two doors is a goat. (It's assumed that you want to win the car, not a goat...) Once you have chosen a door, and before it is opened, the game show host opens another door behind which he knows there is a goat. You are then given the opportunity to revise your choice. Should you swap to the other unopened door?

Two possible arguments are as follows:

1. after the host has opened one door, the probability that the car is behind each of the remaining two doors is 1/2. So, it doesn't matter whether you swap or not.

2. at the start of the game you had probability 1/3 of choosing the door with the car. If you don't swap then 'nothing has really changed', and so the chance of you winning if you don't swap is still 1/3. Therefore you should swap.

Although many people insist that the first of these arguments is correct[3], it is actually false, and you do have probability 2/3 of winning if you swap to the other unopened door. This becomes really easy if we use the law of total probability.

Solution. Let R be the event "the prize is behind the door you chose initially," and S the event "you win the prize by switching doors." If you chose the right door, switching will make you lose, so $P(S|R) = 0$. If you chose the wrong door, switching will make you win for sure: $P(S|R^c) = 1$. We can now calculate the probability of winning by switching the doors by using the law of total probability:

$$P(S) = P(S|R)P(R) + P(S|R^c)P(R^c) = 0 \cdot 1/3 + 1 \cdot 2/3 = 2/3.$$

So, you should switch doors!

Example 3.10 continued: We had introduced the events F = go to Fibbers and A = stay awake in lecture. We were given the conditional probability $P(A|F) = 1/5$ of staying awake in the lecture if you went to Fibbers. But what is the probability of having gone to Fibbers if you stay awake in the lecture?

Solution. We can calculate the probability $P(F|A)$ by using the Definition 3.2 of conditional probability, the multiplication rule (Thm. 3.6) and the law of total probability

[3]This problem became famous in 1990 when discussed by Marilyn vos Savant in the American magazine *Parade*. She reasoned that argument 2 is correct, but subsequently received many thousands of letters disagreeing with her, many from mathematicians. See http://www.nytimes.com/1991/07/21/us/behind-monty-hall-s-doors-puzzle-debate-and-answer.html

(Thm. 3.9):

$$P(F|A) \overset{3.2}{=} \frac{P(F \cap A)}{P(A)} \overset{3.6}{=} \frac{P(A|F)P(F)}{P(A)}$$
$$\overset{3.9}{=} \frac{P(A|F)P(F)}{P(A|F)P(F) + P(A|F^c)P(F^c)}$$
$$= \frac{1/5 \cdot 4/5}{17/50} = \frac{8}{17}.$$

Example 3.7 continued: We had introduced the events M_i = phone is made by manufacturer i, with $i \in \{A, B, C\}$ and the event F = phone goes up in flames. We had been given conditional probabilities that a phone goes up in flames if it was made by a particular manufacturer. But what about the condition probabilities that a phone was made by a particular manufacturer if it goes up in flames?

Solution. The method is as in the previous example, but now we use the law of total probability with a partition containing three alternatives, M_A, M_B, M_C:

$$P(M_A|F) \overset{3.2}{=} \frac{P(M_A \cap F)}{P(F)} \overset{3.6}{=} \frac{P(F|M_A)P(M_A)}{P(F)}$$
$$\overset{3.9}{=} \frac{P(F|M_A)P(M_A)}{P(F|M_A)P(M_A) + P(F|M_B)P(M_B) + P(F|M_C)P(M_C)}$$
$$= \frac{0.01 \cdot 0.6}{0.015} = \frac{6}{15} = \frac{2}{5} = 0.4.$$

So the probability that a phone that goes up in flames was made by manufacturer A is 0.4. Clearly we could now do the same calculation for the other manufacturers. Using index set notation we can formulate this method in the general case as follows:

Theorem 3.12 (Bayes' theorem)**.** *Let $\{B_i | i \in I\}$ be a partition of sample space Ω, such that $P(B_i) > 0$, for all $i \in I$. If A is an event with $P(A) > 0$, then*

$$P(B_i \,|\, A) = \frac{P(A \,|\, B_i)\, P(B_i)}{P(A)} = \frac{P(A \,|\, B_i)\, P(B_i)}{\sum_{j \in I} P(A \,|\, B_j)\, P(B_j)}, \quad \forall\, i \in I.$$

Proof. The proof just follows the pattern of the two previous examples.

$$P(B_i \,|\, A) \overset{3.2}{=} \frac{P(B_i \cap A)}{P(A)} \overset{3.6}{=} \frac{P(A \,|\, B_i)\, P(B_i)}{P(A)}$$
$$\overset{3.9}{=} \frac{P(A \,|\, B_i)\, P(B_i)}{\sum_{j \in I} P(A \,|\, B_j)\, P(B_j)}.$$

□

We can deduce the following simple rule from this theorem:

Corollary 3.13 (Bayes theorem for two alternatives)**.** *For any events A and B with positive probability,*

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} = \frac{P(A \mid B)\, P(B)}{P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B)^c}\,.$$

Proof. Use the partition $\{B, B^c\}$ in Theorem 3.12. □

This corollary is often useful when we want to know $P(B \mid A)$ but are in a situation where $P(A \mid B)$ is much simpler to calculate.

Bayes' theorem (named after the 18^{th} century English clergyman Thomas Bayes) is arguably the most famous result in conditional probability and forms the basis of Bayesian Statistics. In a nutshell it tells us how to update probabilities after new information has arrived.

The application of Bayes' theorem can lead to surprising results as the following example shows.

Example 3.14. Suppose that a diagnostic test for fresher's flu is 95% accurate both on those who do have fresher's flu and on those who don't. Assuming that 0.5% of the population actually has this flu[4], what is the probability that a particular individual has fresher's flu, given that the test is positive (*i.e.* indicates that the individual has fresher's flu)?

Solution. Let F be the event that the individual has fresher's flu and let T be the event that the test is positive. Note that F and F^c form, by definition, a partition of the sample space. Bayes' theorem then gives

$$\begin{aligned} P(F \mid T) &= \frac{P(T \mid F)\, P(F)}{P(T \mid F)\, P(F) + P(T \mid F^c)\, P(F^c)} \\ &= \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.05)(0.995)} \\ &\approx 0.087. \end{aligned}$$

So in spite of the high accuracy of the test, the probability that the student has fresher's flu is less than 9%.

Before we end this section, there is an alternative form of Bayes' theorem which, whilst mathematically identical to theorem 3.12, is common enough and sufficiently different looking that it is worth writing down. This is the *conditional* Bayes' theorem, where we have an extra "given" event on both sides of the equation. Before we get to the theorem though, it is worth briefly noting how to chain together conditions.

[4]Presumably this test is not conducted during fresher's week.

Theorem 3.15. *Let A, B, C be events such that $P(C) > 0$ and $P(B \cap C) > 0$, then*

$$P_C(A|B) = P(A|B \cap C).$$

Proof. Starting with the definition of conditional probability for P_C, we have

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

We can then expand the definition of P_C to get

$$P_C(A|B) = \frac{\left(\frac{P(A \cap B \cap C)}{P(C)}\right)}{\left(\frac{P(B \cap C)}{P(C)}\right)}.$$

Multiplying the top and bottom of the main fraction by $P(C)$, we get the definition of $P(A|B \cap C)$ as desired. □

Corollary 3.16. *(Conditional Bayes' theorem) Let A, B, and C be events such that $P(C)$, $P(A \cap C)$, and $P(B \cap C)$ are all > 0, then*

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

Proof. Apply Bayes' theorem to $P_C(B|A)$, then use theorem 3.15. □

Example 3.11 revisited: We can use the conditional Bayes theorem for a more explicit treatment of the Monty Hall problem. Intuitively, one might want to phrase the problem as, for example: *"Given I chose door 1, and Monty revealed a goat behind door 2, what is the probability the car is actually behind door 3?"*

To do this, we can introduce three collections of events (which are actually partitions, according to definition 3.8) in each case, $i \in \{1, 2, 3\}$:

- F_i: You initially select door i,
- G_i: Monty reveals a goat behind door i.
- C_i: The car is behind door i,

Then we are looking for $P(C_3|F_1 \cap G_2)$. This is hard to compute, hence the problem! However, we can fairly easily compute $P(G_2|F_1 \cap C_3)$: if the car is behind door 3, and we picked door 1, Monty has to reveal a goat behind door 2. Similarly, we find $P(G_2|F_1 \cap C_2) = 0$ and $P(G_2|F_1 \cap C_1) = 1/2$ (assuming that Monty is equally likely to reveal either door if our first pick was the car). We can then apply the conditional

Bayes' theorem, with everything conditional on our door choice:

$$P(C_3|G_2 \cap F_1) = \frac{P(G_2|C_3 \cap F_1)P(C_3|F_1)}{P(G_2|F_1)}.$$

We then use the Law of Total Proability on the denominator, using the partition $\{C_i|i \in 1,2,3\}$ to get

$$P(C_3|G_2 \cap F_1) = \frac{P(G_2|C_3 \cap F_1)P(C_3|F_1)}{P(G_2|C_1 \cap F_1)P(C_1|F_1) + P(G_2|C_2 \cap F_1)P(C_2|F_1) + P(G_2|C_3 \cap F_1)P(C_3|F_1)}.$$

If we assume that the $P(C_i|F_1)$ are all equal then we can cancel them all out, and we are left with

$$\begin{aligned} P(C_3|G_2 \cap F_1) &= \frac{P(G_2|C_3 \cap F_1)}{P(G_2|C_1 \cap F_1) + P(G_2|C_2 \cap F_1) + P(G_2|C_3 \cap F_1)} \\ &= \frac{1}{\frac{1}{2} + 0 + 1} = \frac{2}{3}, \end{aligned}$$

as before! Note that with this formula, we can also see what happens if Monty *does not* pick at random in the event $C_1 \cap F_1$ where we got it right first time. For example, if we know Monty always reveals the rightmost door with a goat behind it, $P(G_2|C_1 \cap F_1) = 0$, hence $P(C_3|G_2 \cap F_1) = 1$ and we are *guaranteed* to win by switching. However, if Monty always goes for the *leftmost* door, then $P(G_2|C_1 \cap F_1) = 1$, and we thus have only a 50% chance of succeeding no matter what we do. *(If we knew this was his strategy, then which door should we have picked first?)*

3.4 Independence

There are situations where knowledge that an event B occurs does not influence the probability of occurrence of an event A. This gives rise to the notion of *independent events*.

Definition 3.17. Events A and B are **independent** if $P(A \cap B) = P(A)\,P(B)$.

Theorem 3.18. *Let A, B be events with $P(B) > 0$. The following statements are equivalent:*

1. *A and B are independent,*
2. *$P(A|B) = P(A)$.*

Proof. To show that statement 1. and 2. are equivalent we need to show that 1. implies 2. and that 2. implies 1. To show that 1. implies 2. we use the definition of conditional

probability together with 1.

$$P(A|B) \overset{3.2}{=} \frac{P(A \cap B)}{P(B)} \overset{1.}{=} \frac{P(A)P(B)}{P(B)} = P(A).$$

To show that 2. implies 1. we use the multiplication rule together with 2.

$$P(A \cap B) \overset{3.6}{=} P(A|B)P(B) \overset{2.}{=} P(A)P(B).$$

□

Example 3.19. A playing card is picked at random from an ordinary deck of 52 cards. The events $A = \{\text{the card is red}\}$ and $B = \{\text{the card is an eight}\}$ are independent, since

$$P(A \cap B) = P(\text{card is } 8\heartsuit \text{ or } 8\diamondsuit) = \frac{2}{52} = \left(\frac{26}{52}\right)\left(\frac{4}{52}\right) = P(A)\,P(B)\ .$$

Theorem 3.20. *If A and B are independent events then also A and B^c are independent.*

Proof. We have $A = (A \cap B) \cup (A \cap B^c)$ and thus $P(A) = P(A \cap B) + P(A \cap B^c)$ by (P3). Solving this equation for $P(A \cap B^c)$ and using that A and B are independent and hence $P(A \cap B) = P(A)P(B)$ gives

$$P(A \cap B^c) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c),$$

where for the last equality we used property (P4). This establishes that A and B^c are independent. □

So far we have only discussed the independence of two events. The following generalises the concept to an arbitrary number of events.

Definition 3.21. A collection $\mathcal{A} = \{A_i | i \in I\}$ of events is called **independent** if, for all finite subsets J of I,

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)\,.$$

Clearly, Definition 3.17 is a special case of Definition 3.21 when $|I| = 2$. If $|I| = 3$, e.g. $\mathcal{A} = \{A_1, A_2, A_3\}$, then the events A_1, A_2, A_3 are independent if and only if *all* of the following equalities hold:

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

Note that pairwise independence is not enough to guarantee the independence of three events. There is an example of this on the problem sheet.

Example 3.22. A duck has laid 3 eggs, each of which is equally likely to be a male or a female duckling independently of the others. Define the events

$$\begin{aligned} A &= \{\text{all the ducklings are of the same sex}\} \\ B &= \{\text{there is at most one male}\} \\ C &= \{\text{the offspring includes a male and a female duckling}\}. \end{aligned}$$

Show that A is independent of B and that B is independent of C. Is A independent of C? Do the above results hold if males and females are not hatched with equal probability?

Solution. The sample space consists of the ordered triples

$$\Omega = \{m, f\} \times \{m, f\} \times \{m, f\} = \{(i, j, k) | i, j, k \in \{m, f\}\}$$

where m stands for male and f for female in order of hatching (*i.e.* i is the entry for the first duck, j the second and k the third). Then

$$\begin{aligned} P(A) &= P(\{(m,m,m)\} \cup \{(f,f,f)\}) = P(\{(m,m,m)\}) + P(\{(f,f,f\})) \\ &= P(\text{a male is hatched})^3 + P(\text{a female is hatched})^3 \\ &= \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}, \end{aligned}$$

where the third equality follows by independence. Similarly,

$$\begin{aligned} P(B) &= P(\{(f,f,f),(m,f,f),(f,m,f),(f,f,m)\}) \\ &= P(\{(f,f,f)\}) + P(\{(m,f,f)\}) + P(\{(f,m,f)\}) + P(\{(f,f,m)\}) \\ &= 4\left(\frac{1}{2}\right)^3 = \frac{1}{2}, \end{aligned}$$

and since $C = A^c$, $P(C) = P(A^c) = 1 - P(A) = 1 - \frac{1}{4} = \frac{3}{4}$. Now

$$P(A \cap B) = P(\{(f,f,f)\}) = \left(\frac{1}{2}\right)^3 = \frac{1}{8} = P(A)\,P(B).$$

So events A and B are independent. Also,

$$P(B \cap C) = P(\{(m,f,f),(f,m,f),(f,f,m)\}) = 3P\left(\frac{1}{2}\right)^3 = \frac{3}{8} = P(B)\,P(C)$$

so events B and C are independent. However, $A \cap C = \emptyset$, so $P(A \cap C) = 0 \neq P(A)\,P(C)$, so events A and C are not independent.

The above conclusions are very much dependent on the fact that male and female ducklings are equally likely to hatch. If $P(\text{a male duckling hatches}) = p \neq \frac{1}{2}$, then

$P(A \cap B) = (1-p)^3$, whereas

$$P(A)\,P(B) = \left[p^3 + (1-p)^3\right]\left[(1-p)^3 + 3p(1-p)^2\right]$$

so events A and B are no longer independent unless either $p = 0$ (a female duckling hatches with probability 1) or $p = 1$ (a male duckling hatches with probability 1).

Note that we could have worked with a smaller sample space $\Omega = \{0, 1, 2, 3\}$ where the numbers refer to the number of males among the three ducklings. In this notation the events are $A = \{0, 3\}, B = \{0, 1\}, C = \{1, 2\}$. However in this sample space not all outcomes are equally probable. Rather

$$P(\{0\}) = P(\{3\}) = \frac{1}{8}, \quad P(\{1\}) = P(\{2\}) = \frac{3}{8}.$$

Definition 3.23. Events A and B are **conditionally independent** given an event C if

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

Note that this is just exactly the same definition as Definition 3.17 but for the conditional probability function $P(\cdot|C)$ instead of $P(\cdot)$. It follows that also Theorem 3.18 applies so that for example if $P(B|C) > 0$ then the conditional independence of A and B given C implies that

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

It is not really necessary to prove this again, but it is easy to do so by repeated use of the product rule and the conditional independence property:

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

Example 3.24. In Example 3.14 we considered a fresher's flu test. We introduced event F that you have fresher's flu and event T that the test returns positive. Using the given probabilities $P(F) = 0.005$ and $P(T|F) = P(T^c|F^c) = 0.95$ we calculated that the probability of you having fresher's flu if the test returns positive is quite small:

$$P(F|T) = \frac{P(T \mid F)\,P(F)}{P(T \mid F)\,P(F) + P(T \mid F^c)\,P(F^c)} \approx 0.087.$$

Now assume that there is another independent test for fresher's flu. Denote by S the

event that this second test also returns positive. Let us assume that the second test is not quite as reliable as the first but instead the probability that it returns a correct result when you have fresher's flu is $P(S|F) = 0.8$ and the probability that it returns a correct result when you do not have fresher's flu is $P(S^c|F^c) = 0.9$. We want to calculate the probability $P(F|T \cap S)$ that you have fresher's flu given that both tests come back positive.

Solution. We can apply the conditional Bayes' theorem (3.16) with respect to T:

$$P(F|S \cap T) = \frac{P(S|F \cap T)P(F|T)}{P(S|F \cap T)P(F|T) + P(S|F^c \cap T)P(F^c|T)}. \tag{3.2}$$

We have to realise that the fact that the two tests are independent does not mean that the events T and S are independent. Clearly if the first test returns positive it does make it more likely that you have fresher's flu and therefore it is more likely that the second test returns positive, i.e., $P(S|T) > P(S)$, which means T and S are not independent. However, T and S are conditionally independent given F or given F^c, because if it is already known whether you have fresher's flu or not, the result of the first test is not going to affect the result of the second test. This conditional independence implies that (see the note after Definition 3.23)

$$P(S|F \cap T) = P(S|F) = 0.8, \quad P(S|F^c \cap T) = P(S|F^c) = 1 - P(S^c|F^c) = 0.1.$$

Finally $P(F^c|T) = 1 - P(F|T) \approx 0.913$ so that

$$P(F|T \cap S) \approx \frac{0.8 \cdot 0.087}{0.8 \cdot 0.087 + 0.1 \cdot 0.913} \approx 0.43.$$

So after the second test has returned positive as well the probability that you have fresher's flu has risen to about 43%.

The fact that Bayes' theorem can be applied repeatedly when new information comes in to update the probabilities is very useful, especially in the case where the incoming data is conditionally independent of previous data, as in the above case. You will explore this more in the second R lab.

4 Discrete Random Variables

(Textbook chapter link)

4.1 Random variables

We have seen how the probability function assigns probabilities to events. For example the weather forecast gives me a probability that it will rain tomorrow. But I might also be interested in the temperature tomorrow. That is not an event. It is a number, but its value is uncertain. If we think of the weather as a probability experiment, then the temperature tomorrow is only known once the outcome of that experiment is known.

A **random variable** is a quantity of interest that depends on the outcome of a probability experiment.

Example 4.1. Consider the following game: you roll a fair die and (a) win 2 points if the outcome is 5 or 6, (b) lose 1 point if the outcome is 1, 2 or 3, (c) win or lose nothing if the outcome is 4. The sample space of this random experiment is $\Omega = \{1, 2, \ldots, 6\}$. Denoting losing points as gaining a negative amount of points, we can represent the gains from this game as a function $X : \Omega \to \mathbb{R}$ defined as

$$X(\omega) = \begin{cases} -1 & \text{if } \omega = 1, 2, 3 \\ 0 & \text{if } \omega = 4 \\ 2 & \text{if } \omega = 5, 6. \end{cases}$$

The amount you gain is a random variable.

Definition 4.2. A **random variable** is a ($\mathcal{F}$-measurable)[5] function from a sample space Ω into the real numbers $\mathbb{R}$.

Discrete random variables are random variables that take only countably many values.

Definition 4.3. A random variable $X : \Omega \to \mathbb{R}$ is **discrete** if its image $X(\Omega)$ is countable.

The random variable in Example 4.1 is an example of a discrete random variable with image $X(\Omega) = \{-1, 0, 2\}$.

The different values that a random variable can take define different events. In Example 4.1

$$\begin{aligned} X^{-1}(-1) &= \{\omega \in \Omega | X(\omega) = -1\} = \{X = -1\} = \{1, 2, 3\} \\ X^{-1}(0) &= \{\omega \in \Omega | X(\omega) = 0\} = \{X = 0\} = \{4\} \\ X^{-1}(2) &= \{\omega \in \Omega | X(\omega) = 2\} = \{X = 2\} = \{5, 6\} \end{aligned}$$

[5] In this module we will not worry about measurability. But if you are curious: a function $X : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable if $X^{-1}(B) \in \mathcal{F}$ for all Borel subsets of $\mathbb{R}$.

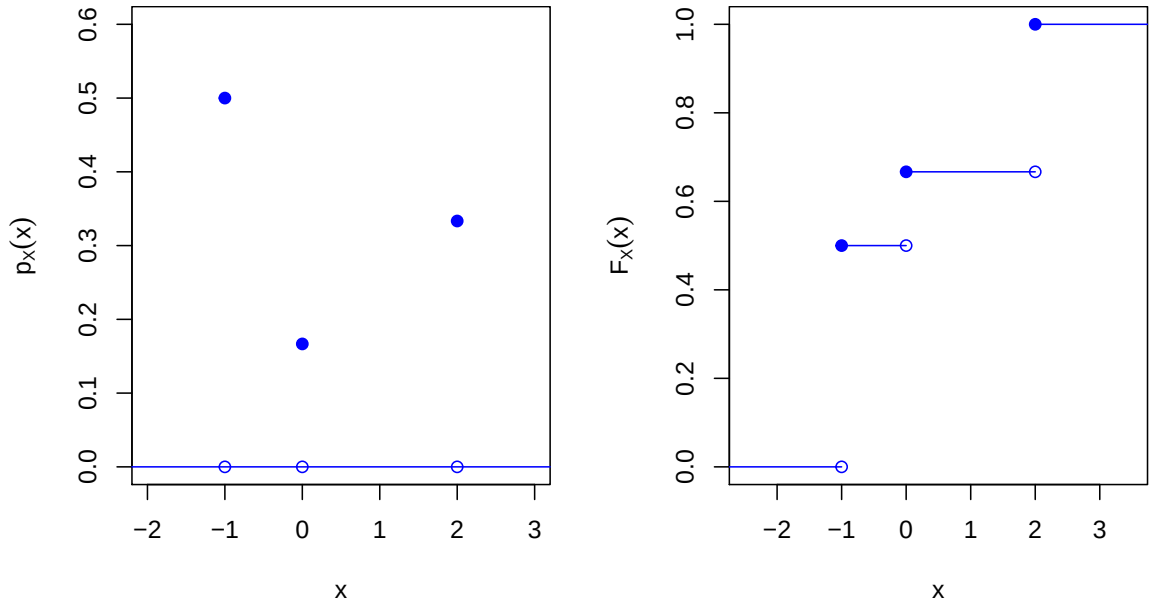


Figure 4.1: Probability mass function and distribution function for Example 4.1

These events form a partition of Ω. We will often be interested in the probability of these events. Usually we are quite lazy and write just $P(X = 2)$ instead of the full form $P(\{\omega \in \Omega | X(\omega) = 2\})$.

4.2 The probability distribution of a discrete random variable

Definition 4.4. The **probability mass function** of a discrete random variable X is the function $p_X : \mathbb{R} \to \mathbb{R}$ defined by

$$p_X(x) = P\left(X = x\right).$$

We will adopt the notation that we use capital letters to denote random variables and lower case letters to denote real numbers.

In Example 4.1 the probability mass function is given by

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = -1 \\ 1/6 & \text{if } x = 0 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{if } x \notin \{-1, 0, 2\}. \end{cases}$$

You can find a graph of this function in Figure 4.1

Mass functions have two defining properties:

Theorem 4.5. *Let X be a random variable with countable image $I = X(\Omega)$. Then its*

probability mass function p_X satisfies

$$p_X(x) \geq 0 \quad \forall x \in \mathbb{R} \text{ and } p_X(x) = 0 \quad \forall x \notin I, \tag{m1}$$

$$\sum_{x \in I} p_X(x) = 1. \tag{m2}$$

Proof. While (m1) is obvious, (m2) can be obtained (using (P3)) as follows:

$$\sum_{x \in I} p_X(x) = \sum_{x \in I} P(X = x) = P\left(\bigcup_{x \in I} \{X = x\}\right) = P(\Omega) = 1\,.$$

□

It is a fundamental result that if a function satisfies (m1) and (m2), it is the mass function of some random variable.

Theorem 4.6. *Consider any countable set $I \subset \mathbb{R}$ and any function $p_X : \mathbb{R} \to \mathbb{R}$ satisfying* (m1) *and* (m2) *above. Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a discrete random variable $X : \Omega \to \mathbb{R}$ such that p_X is the mass function of X.*

Proof. We can simply construct an example of such a probability space and random variable. We choose $\Omega = I$, $\mathcal{F} =$ the set of all subsets of Ω, P the probability function defined according to Theorem 2.18 by

$$P(\{x\}) = p_X(x) \text{ for } x \in \Omega$$

and X the random variable $X : \Omega \to \mathbb{R}, X(x) = x$. We then have

$$P(X = x) = P(\{x\}) = p_X(x)$$

for all $x \in I$ and for $x \notin I$ we have $P(X = x) = P(\emptyset) = 0$. Definition 4.4 then implies that p_X is the mass function of X. □

This will allow us to describe the properties of discrete random variables by just looking at their mass functions, forgetting about sample spaces, events and probabilities. We only need say "let X be a random variable with mass function p_X" and we can be sure that such a random variable exists without having to construct it explicitly.

Events involving values of a random variable can be assigned probabilities via *distribution functions.*

Definition 4.7. Let X be a random variable. The **distribution function** of X is the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by

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

Be careful about notation: X denotes a random variable; x denotes a real number.

In Example 4.1 the probability mass function is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < -1 \\ 1/2 & \text{if } -1 \leq x < 0 \\ 2/3 & \text{if } 0 \leq x < 2 \\ 1 & \text{if } 2 \leq x. \end{cases}$$

You can find a graph of this function in Figure 4.1.

In general, for any discrete random variable X $\{x_0, x_1, x_2, \dots\}$ with $x_0 < x_1 < \dots$,

$$F_X(x) = P(X \leq x) = P\left(\bigcup_{\substack{y \in X(\Omega) \\ y \leq x}} \{X = y\} \right) = \sum_{\substack{y \in X(\Omega) \\ y \leq x}} p_X(y).$$

That is, the distribution function is obtained simply by summing the mass function for all possible values up to x.

A different way of saying this is that at every possible value for X, the distribution function F_X jumps by the probability of that value. So if x is one of the possible values, then the distribution function at x is larger than the distribution function a little bit to the left of x by $p_X(x)$.

$$p_X(x) = F_X(x) - \lim_{\varepsilon \to 0^+} F_X(x - \varepsilon).$$

Note that this allows us to reconstruct the probability mass function p_X from the distribution function F_X.

Here are some general properties satisfied by the distribution function of any random variable:

Theorem 4.8. *The distribution function F_X of any random variable X satisfies:*

*(i) $F_X(x)$ is increasing in x (*i.e. *if $x \leq y$ then $F_X(x) \leq F_X(y)$);*

(ii) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$;

(iii) F_X is right continuous: $\lim_{\varepsilon \to 0^+} F_X(x + \varepsilon) = F_X(x)$ for all $x \in \mathbb{R}$.

Conversely, any function $F_X : \mathbb{R} \to \mathbb{R}$ satisfying these three properties is the distribution function of some random variable.

4.3 Frequently used discrete probability distributions

There are certain types of discrete probability distributions that are used very frequently: we introduce some of these below. (We take $n \in \mathbb{N}$ and $p \in [0, 1]$ throughout.)

Definition 4.9. We say that a discrete random variable X has the **Bernoulli distribution** with parameter p, and write $X \sim \mathrm{Ber}(p)$, if it only takes values 0 and 1, *i.e.* $X(\Omega) = \{0, 1\}$, with

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p\,.$$

The mass function of X is

$$p_X(x) = P(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{if } x \neq 0, 1. \end{cases}$$

The distribution function of X is

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x\,. \end{cases}$$

Their graphs are plotted in Figure 4.2.

Definition 4.10. The **indicator random variable** of an event A is the random variable $\mathbb{1}_A$ defined by

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

Note that $P(\mathbb{1}_A = 1) = P(A)$ and $P(\mathbb{1}_A = 0) = P(A^c) = 1 - P(A)$ and so

$$\mathbb{1}_A \sim \mathrm{Ber}(P(A)).$$

Definition 4.11. We say that the random variable X has the **binomial distribution** with parameters n and p, and write $X \sim \mathrm{Bin}(n, p)$, if $X(\Omega) = \{0, 1, \ldots, n\}$ and it has mass function

$$p_X(k) = P(X = k) = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{if } k = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

Here $\binom{n}{k}$ is the binomial coefficient that is also sometimes denoted as C_k^n.

Note that if $n = 1$ this is simply the $\mathrm{Ber}(p)$ distribution. The fact that the above is

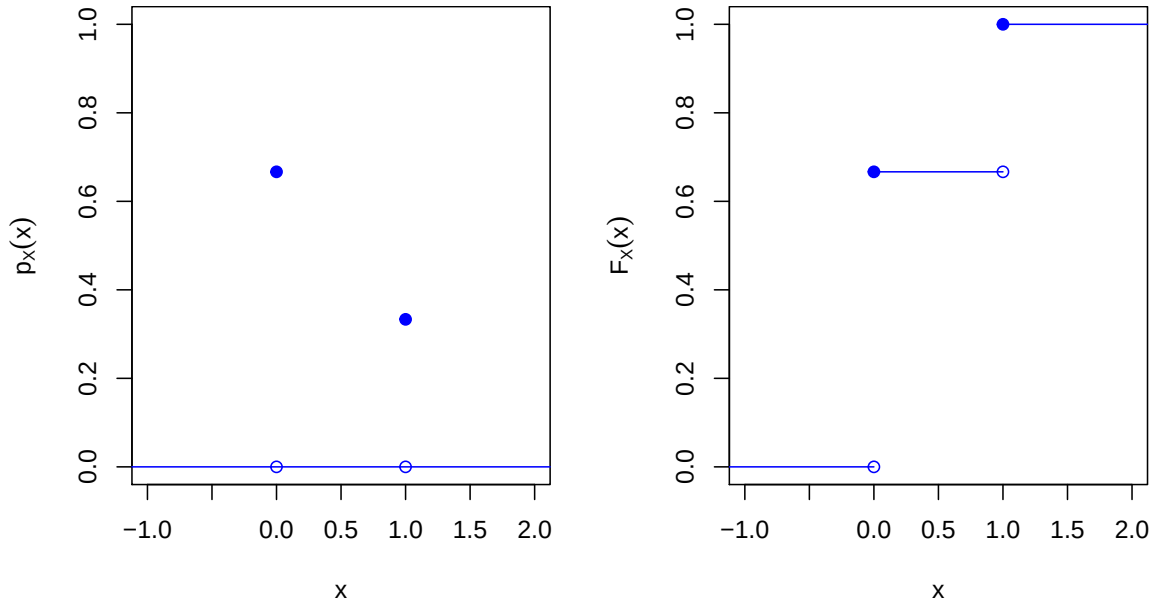


Figure 4.2: Probability mass function and distribution function for the Ber(1/3) distribution.

indeed a mass function follows from the binomial theorem: for each $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k},$$

which gives

$$\sum_{k=0}^{n} P\left(X=k\right) = \sum_{k=0}^{n} \binom{n}{k} p^k \left(1-p\right)^{n-k} = \left[p + (1-p)\right]^n = 1.$$

Example 4.12. Suppose a coin has probability p of showing heads and is tossed n times. The sample space is

$$\Omega = \{H, T\}^n = \{(\omega_1, \ldots, \omega_n) | \omega_i \in \{H, T\}\},$$

i.e. Ω is the set of all n-tuples with entries taken from the set $\{H, T\}$. Associated to the i^{th} toss is a indicator random variable

$$X_i = \mathbb{1}_{\text{i-th toss gives Heads}} = \mathbb{1}_{\{\omega_i = H\}} \sim \text{Ber}(p).$$

Tosses are independent (the outcome of one toss does not affect the others), and the random variable

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

represents the *total* number of heads. The event $\{X = k\}$ is the event of getting exactly

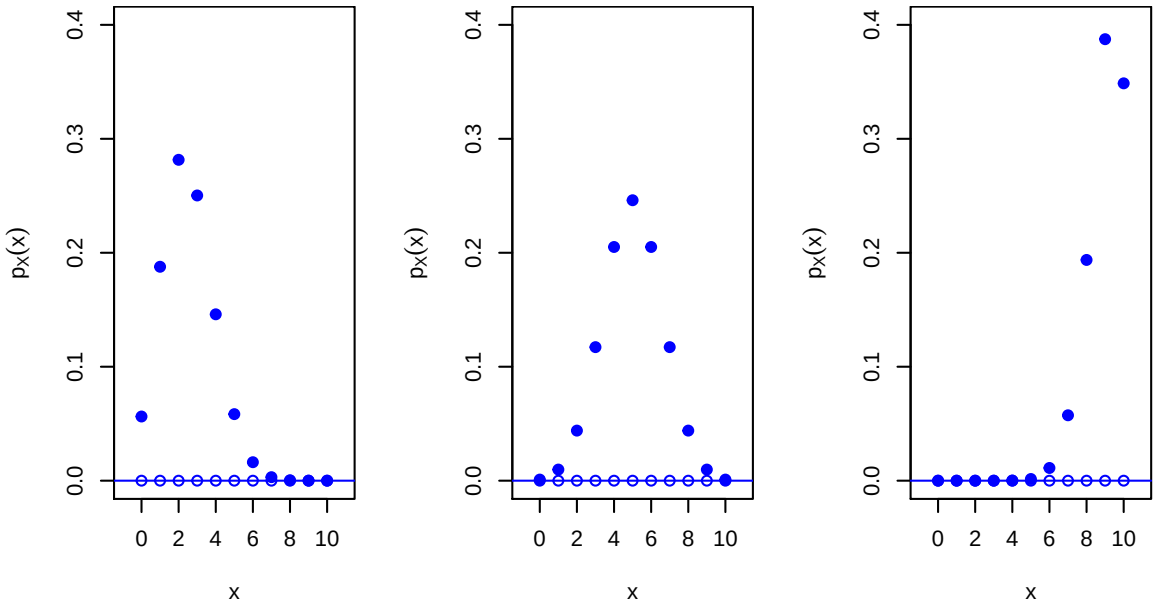


Figure 4.3: Mass functions for (left to right): $\mathrm{Bin}(10, 0.25)$, $\mathrm{Bin}(10, 0.5)$, $\mathrm{Bin}(10, 0.9)$.

k heads out of n tosses. This set contains $\binom{n}{k}$ outcomes (vectors of length n containing k Heads and $n-k$ Tails), each of which has the same probability of occurring (because the order of the outcomes doesn't matter). Thus

$$\begin{aligned} P(X=k) &= \binom{n}{k} P\left(\{(\underbrace{HH\cdots H}_{k \text{ times}}\underbrace{TT\cdots T}_{n-k \text{ times}})\}\right) \\ &= \binom{n}{k} \underbrace{pp\cdots p}_{k \text{ times}} \cdot \underbrace{(1-p)\,(1-p)\cdots(1-p)}_{n-k \text{ times}} \\ &= \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned}$$

i.e. $X \sim \mathrm{Bin}(n, p)$.

In general, the sum of n independent $\mathrm{Ber}(p)$-distributed random variables has the $\mathrm{Bin}(n, p)$ distribution.

Definition 4.13. We say that the random variable X has the **geometric distribution** with parameter $p \in (0, 1]$, and write $X \sim \mathrm{Geo}(p)$, if $X(\Omega) = \mathbb{N}$ has mass function

$$p_X(n) = \begin{cases} p(1-p)^{n-1} & \text{if } n \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases}$$

Let us determine the distribution function

$$F_X(x) = P(X \le x) = \sum_{n \in \mathbb{N} n \le x} P(X = n) = \sum_{n=1}^{\lfloor x \rfloor} P(X = n).$$

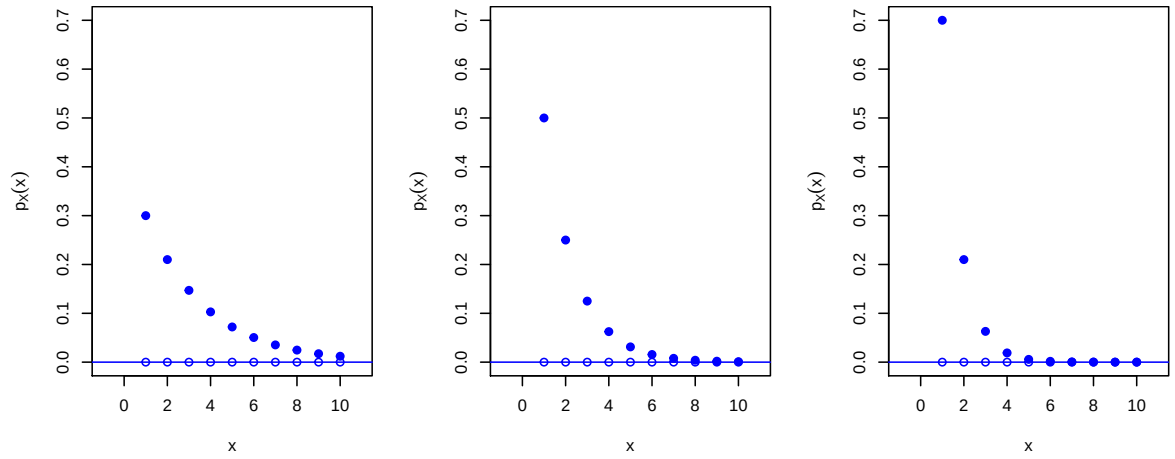


Figure 4.4: Mass functions for (left to right): Geo(0.2), Geo(0.5), Geo(0.7).

In the last equality we introduced the notation $\lfloor x \rfloor$ to denote the largest integer smaller or equal to x. Using the probability mass function given above we then find for $x \geq 1$ that

$$F_X(x) = \sum_{n=1}^{\lfloor x \rfloor} p(1-p)^{n-1} = p \sum_{n=1}^{\lfloor x \rfloor} (1-p)^{n-1} = p \sum_{m=0}^{\lfloor x \rfloor - 1} (1-p)^m.$$

We can now use the formula for the sum of a geometric progression

$$\sum_{m=0}^{N} a^m = \frac{1-a^{N+1}}{1-a}, \quad \text{for } a \in [0,1)$$

with $N = \lfloor x \rfloor - 1$ and $a = 1-p$. This gives, for $x \geq 1$,

$$F_X(x) = p \frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} = 1-(1-p)^{\lfloor x \rfloor}.$$

The distribution function for X is thus given by

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & x < 1 \\ 1-(1-p)^{\lfloor x \rfloor} & x \geq 1 \end{cases}.$$

We see that this has the required property of a distribution function that

$$\lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} 1-(1-p)^{\lfloor x \rfloor} = 1$$

because $1-p$ is smaller than 1.

Example 4.14. You flip a biased coin with $P(\{H\}) = p$ until you get the first Heads (this example will be familiar to you from section 2.5 in the textbook). Let X =number

of the flip on which you get the first Heads. Then

$$P(X = n) = p(\{(T, T, \dots, T, H)\}) = (1-p)^{n-1}p.$$

Thus $X \sim \text{Geo}(p)$.

This example illustrates that a geometric distribution describes the waiting time until a success in a series of Bernoulli trials (trials that lead to either success or failure).

The geometric distribution has the memoryless property:

Theorem 4.15. *Let X be a random variable that has the geometric distribution, $X \sim \text{Geo}(p)$. Then for any $n, k \in \mathbb{N}$*

$$P(X > n + k | X > k) = P(X > n).$$

Proof.

$$\begin{aligned} P(X > n+k | X > k) &= \frac{P(\{X > n+k\} \cap \{X > k\})}{P(X > k)} \text{ by Def.3.2} \\ &= \frac{P(X > n+k)}{P(X > k)} \text{ because } \{X > n+k\} \subset \{X > k\} \\ &= \frac{(1-p)^{n+k}}{(1-p)^k} \text{ using the distribution function given above} \\ &= (1-p)^n = P(X > n). \end{aligned}$$

□

In practice this memoryless property means that for how many trials you have already waited for a success does not affect how many trials you will still have to wait. If you had bad luck for a long time in a game of dice and had to wait for that six from the die for many rounds, that does not have the effect that the six is now bound to come soon; there is no sense of fairness in the game.

There is one more famous discrete probability distribution that I want to list even though we will not discuss it much in this module because it is derived in detail in our Stochastic Processes module.

Definition 4.16. We say that the random variable X has the **Poisson distribution** with parameter $\lambda > 0$, and write $X \sim \text{Pois}(\lambda)$, if $X(\Omega) = \{0, 1, 2, \dots\}$ and it has mass function

$$p_X(n) = \begin{cases} \frac{\lambda^n}{n!} e^{-\lambda} & \text{if } n = 0, 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

The fact that the above is indeed a mass function follows from the Taylor series for the exponential function: for each $x \in \mathbb{R}$,

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Thus,

$$\sum_{n=0}^{\infty} P(X = n) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1.$$

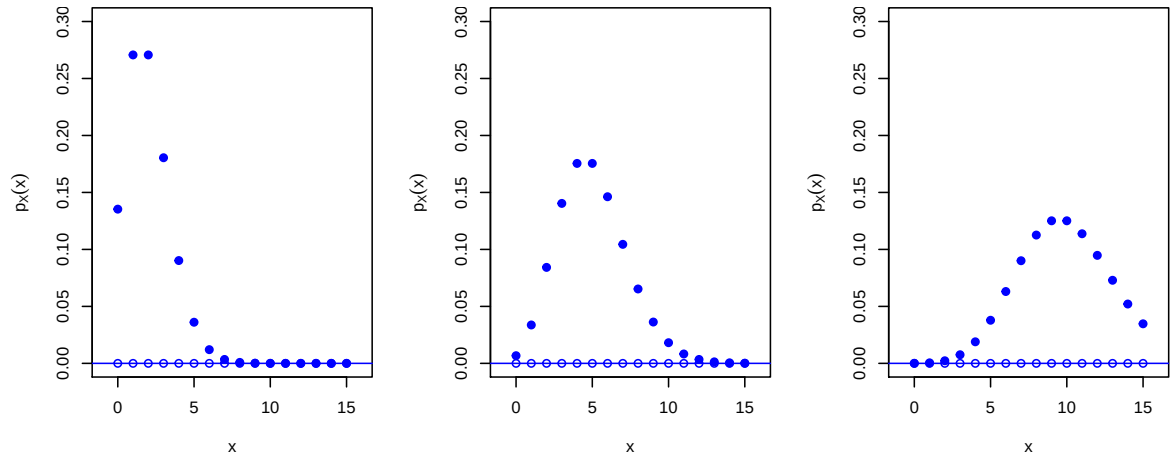


Figure 4.5: Mass functions for (left to right): Pois(2), Pois(5), Pois(10).

5 Continuous Random Variables

(Textbook chapter link)

Discrete random variables take only countably many values. Of course such random variables do not cover all cases of interest (*e.g.* we may want to choose a number at randomly from the interval $[0, 1]$). In this chapter we discuss an important class of random variables whose images are uncountable.

5.1 Probability density functions

We have seen in Theorem 4.8 that distribution functions can have jumps as long as the discontinuity is on the left side of the jump (*i.e.* distribution functions are right-continuous). Indeed, distribution functions of *discrete* random variables are step functions. At the other extreme there are random variables whose distribution functions are differentiable and we call such random variables *continuous*. This name is a bit unfortunate, because a continous random variable is not continuous in the sense in which a function from the reals to the reals can be continuous, a concept that you have discussed in Calculus. A random variable is a function from a sample space Ω into the reals and thus the usual concept of continuity does not apply (because there is no topology on Ω). Continuous random variables are instead characterised by a property of their distribution functions.

Definition 5.1. We call a random variable X **continuous** if its distribution function F_X can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds \qquad \text{for all } x \in \mathbb{R}$$

for some function $f_X : \mathbb{R} \to \mathbb{R}$. In this case, we say that f_X is the **density function** of X.

The fundamental theorem of calculus implies, under some mild regularity conditions, that for each $x \in \mathbb{R}$,

$$\frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(s) ds = f_X(x). \tag{5.1}$$

Thus one can go from the density function to the distribution function by integration and from the distribution function back to the density by differentiation.

For calculating probabilities of events involving random variables, density functions have for continuous random variables the same role that mass functions have for discrete random variables. Similar to the characteristic properties (m1) and (m2) of mass functions from Theorem 4.5, for the density functions we have

Theorem 5.2. *Let X be a continuous random variable then its density function f_X satisfies*

$$f_X(x) \geq 0, \quad \text{for all } x \in \mathbb{R}\,; \tag{d1}$$

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1. \tag{d2}$$

Conversely, any real function f_X satisfying (d1) and (d2) is the density function of some continuous random variable.

Property (d1), the non-negativity of f_X, is necessary to ensure that F_X is an increasing function as required by Theorem 4.8(ii). Also, from Theorem 4.8(ii), we obtain

$$1 = \lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f_X(s)ds = \int_{-\infty}^{\infty} f_X(s)ds,$$

which gives us property (d2).

The similarity between the properties (m1) and (m2) of mass functions and the properties (d1) and (d2) of density functions is striking. The main change is that for continuous random variables there is not longer a discrete set of values that the random variable can take and thus the sum over those values needs to be replaced by an integral over the entire real line in property (d2). The analogy between mass function and density function, however, is not perfect. For one thing, $f_X(\cdot)$ is not a probability, unlike $p_X(\cdot)$: it might well take values greater than 1.

A good way to think about the difference between probability mass and probability density is the way you think about the difference between physical mass and density.

The following result is very useful for calculations.

Theorem 5.3. *If X is a continuous random variable with density function f_X, then for all $a, b \in \mathbb{R}$ with $a \leq b$*

$$P\left(a < X \leq b\right) = \int_a^b f_X(x)dx.$$

.

Proof. We note that

$$\{a < X \leq b\} = \{X \leq b\} \cap \{X > a\} = \{X \leq b\} \cap \{X \leq a\}^c.$$

Furthermore

$$\{X \leq a\} \subseteq \{X \leq b\}.$$

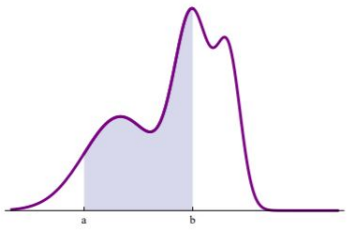
Hence we can now use a fact that you proved in your homework: if A, B are events with

$A \subseteq B$ then $P(B \cap A^c) = P(B) - P(A)$. This gives

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a) \quad \text{by Def. 4.7} \\ &= \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx \quad \text{by Def. 5.1} \\ &= \int_{a}^{b} f_X(x)dx. \end{aligned}$$

For the last equality we used a result about integrals. Actually proving that result will have to wait until you have the mathematical machinery to define integrals rigorously later in your degree. □

Note that this theorem says that we can calculate $P(a < X \leq b)$ simply by calculating the area under the density function between the points a and b.



A startling difference between continuous and discrete random variables is that, while the probability of a discrete random variable taking a value k is $p_X(k)$, the probability of a continuous random variable taking any particular value x is 0.

Theorem 5.4. *If X is a continuous random variable, then for all $x \in \mathbb{R}$,*

$$P(X = x) = 0.$$

As a consequence of the fact that $P(X = a) = 0 = P(X = b)$, it does not matter whether we use weak inequalities ($\leq$) or strict inequalities ($<$) in Theorem 5.3, i.e.,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx.$$

5.2 Frequently used continuous probability distributions

As with discrete random variables, there are a number of continuous random variables which have special places in probability theory.

Definition 5.5. We say that the continuous random variable X has the **uniform distribution** on $[a, b]$, and write $X \sim \mathrm{U}(a, b)$, if the density of X is

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

The above is indeed a density function since $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f_X(x)dx = \frac{1}{b-a}\int_a^b dx = 1.$$

For $x \in [a, b]$, the distribution function of the uniform distribution is given by

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(s)ds \\ &= \int_a^x \frac{1}{b-a}\, ds = \left[\frac{s}{b-a}\right]_a^x \\ &= \frac{x-a}{b-a}. \end{aligned}$$

Thus the full specification of the distribution function is

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b. \end{cases}$$

This distribution is used, for example, when we talk about 'choosing a number at random' from an interval $[a, b]$. An informal but incorrect description would be to say that that random number is equally likely to take any value in the interval. The reason this is not a good way of saying this is because the probability to take any particular value is zero. Thus we should instead say that the probability is spread uniformly over the interval.

Definition 5.6. We say that the continuous random variable X has the **exponential distribution** with parameter $\lambda > 0$, and write $X \sim \mathrm{Exp}(\lambda)$, if the density of X is

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

This is a density function since $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ (notice that this requires $\lambda \geq 0$) and

$$\int_{-\infty}^{\infty} f_X(x)dx = \lambda\int_0^{\infty} e^{-\lambda x}dx = \lambda\frac{1}{\lambda} = 1.$$

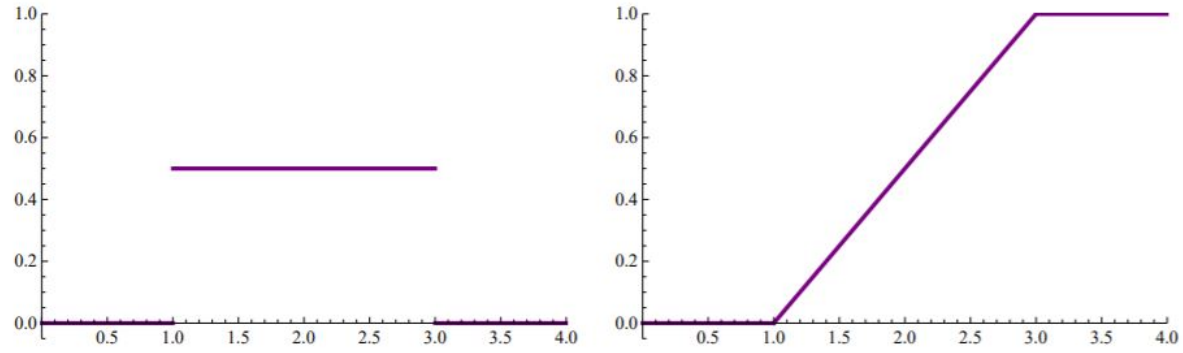


Figure 5.1: Density and distribution function of U[1, 3].

Note that the above required $\lambda \neq 0$. Hence the restriction in the definition to $\lambda > 0$. We can find the distribution function of the exponential distribution by integrating the density function and find that

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

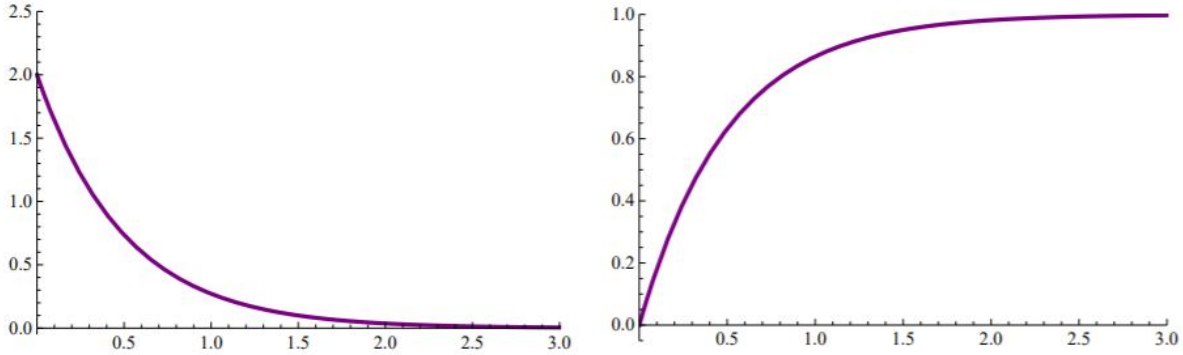


Figure 5.2: Density and distribution function of Exp(2).

The exponential distribution is often used to model waiting times between certain events, such as natural disasters, machine break-downs, or customers joining a queue. If these waiting times are independent and Exp(λ) distributed, then it can be shown that the number of arrivals per unit of time t follows a Pois(λt) distribution. Conversely, if the number of arrivals follows the Pois(λt) distribution, then the waiting times follow the Exp(λ) distribution.

An important property of the exponential distribution is that it satisfies the "memoryless" property. So if the length of time between hurricanes is exponentially distributed, the probability that the next hurricane doesn't occur in the next $t+s$ units of time, given

that we've waited s units already, is simply the same as the probability that the next hurricane doesn't occur in the next t time units. (This seems quite reasonable: nature doesn't decide that there should be a new hurricane soon because there hasn't been one for a while...!) The proof of the memoryless property is very similar to the proof of Theorem 4.15. You will provide it in your homework. One can show that the exponential distribution is the only continuous distribution with this property.

Next we define the Pareto distribution, which is one example of a "power-law distribution" because its density is falling off like a power of x rather than exponentially for large x. Thus the distribution has a "long tail". These distributions show up in many complex systems, in particular social systems. For example the number of friends that a random individual has on a social network tends to be well described by power-law distribution. Wealth and income inequality manifests itself in a power-law distribution for wealth and income, describing the fact that there are some very rich individuals but also a huge number of poor individuals.

Definition 5.7. We say that the continuous random variable X has the **Pareto distribution** with parameter $\alpha > 0$, and write $X \sim \mathrm{Par}(\alpha)$, if the density of X is

$$f_X(x) = \begin{cases} \dfrac{\alpha}{x^{\alpha+1}} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1\,. \end{cases}$$

The distribution function of the Pareto distribution is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 - x^{-\alpha} & \text{if } x \geq 1. \end{cases}.$$

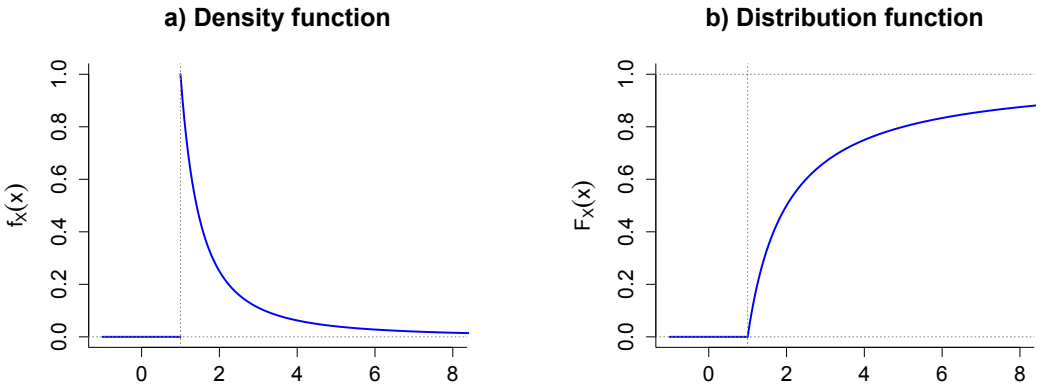


Figure 5.3: Density and distribution function of Par(1).

Definition 5.8. We say that the continuous random variable X has the **normal distribution** with mean μ and variance $\sigma^2 > 0$, and write $X \sim \mathrm{N}(\mu, \sigma^2)$, if the density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } x \in \mathbb{R}.$$

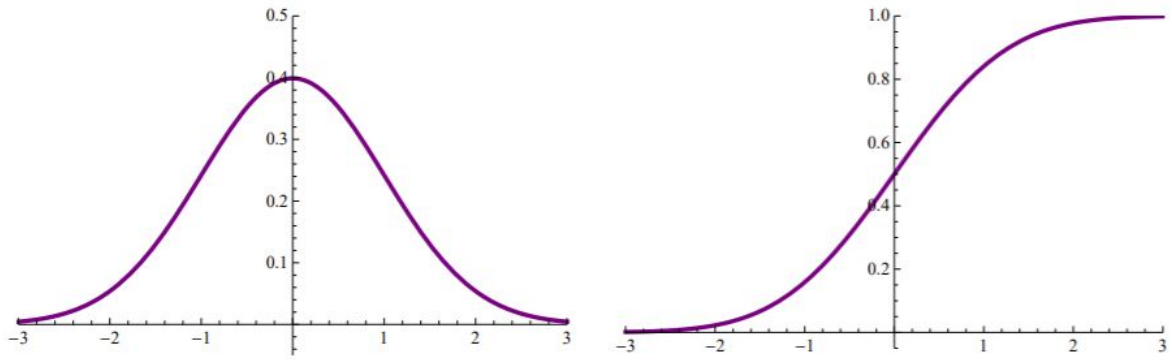


Figure 5.4: Density and distribution function of $\mathrm{N}(0, 1)$.

Because they are used so often in statistics, standard names have been introduced for the density and the distribution function of the standard normal random variable $X \sim \mathrm{N}(0, 1)$. The density is denoted by a lower-case ϕ,

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

and the distribution function is denoted by an upper-case Φ

$$F_X(x) = \Phi(x) = \int_{-\infty}^{x} \phi(s) ds.$$

Unfortunately there is no expression for the result of the integral in terms of standard functions. Traditionally the values of Φ have been tabulated, these days of course calculators and computers can calculate them more easily. You will however find a table of the so-called tail probabilities $1 - \Phi(x)$ in Table B.1 of the textbook.

5.3 Quantiles

Definition 5.9. Let X be a random variable with distribution function F_X and let $p \in [0, 1]$. The pth **quantile** or $(100 \cdot p)$th **percentile** of the distribution of X is the smallest number q_p such that

$$F_X(q_p) = p.$$

The **median** of a distribution is its 50th percentile. The **upper quartile** is the 75th percentile and the **lower quartile** is the 25th percentile respectively.

Example 5.10. Let $X \sim \text{Exp}(\lambda)$. Calculate the median of X.

Solution. Let us denote the median of X by q. To find it we need to solve the equation $F_X(q) = 1/2$. Using the expression for the distribution function of the exponential distribution that we derived earlier, this becomes

$$1 - e^{-\lambda q} = \frac{1}{2}$$

which can be solved for q to give

$$q = \frac{\log(2)}{\lambda}.$$

7 Expectation and variance

(Textbook chapter link)

In this chapter we introduce some quantities that describe features of the shape of a probability distribution, like where it is centred and how broad it is. These will play a big role in the later chapters on statistics.

7.1 Expectation for discrete random variables

The expectation of a random variable can be thought of as the "centre of mass" of the probability distribution.

Definition 7.1. If X is a discrete random variable, the **expectation** of X (or **expected value** of X), denoted by $E[X]$, is defined by

$$E[X] = \sum_{x \in X(\Omega)} x\, p_X(x) = \sum_{x \in X(\Omega)} x\, P(X = x) \tag{7.1}$$

if this series is absolutely convergent. Otherwise the expectation is undefined.

Note that $E[X]$ is a number and **not** a random variable.

Example 7.2. If $X \sim \text{Ber}(p)$, then $X(\Omega) = \{0, 1\}$.

$$E[X] = \sum_{x \in \{0,1\}} x\, p_X(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Example 7.3. If $X \sim \text{Geo}(p)$ then $X(\Omega) = \{1, 2, \dots\}$. Writing $q = 1 - p$:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k q^{k-1} p = p \sum_{k=0}^{\infty} k q^{k-1} = p \sum_{k=0}^{\infty} \frac{d}{dq} q^k \\ &= p \frac{d}{dq} \left(\sum_{k=0}^{\infty} q^k \right) = p \frac{d}{dq} \left(\frac{1}{1-q} \right) \\ &= \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}. \end{aligned}$$

Example 7.4. If $X \sim \text{Pois}(\lambda)$, then

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{-\lambda} \lambda e^{\lambda} = \lambda. \end{aligned}$$

Example 7.5. Consider the following slightly modified dice game: you throw a fair die and lose £1 if $1, 2$ or 3 comes up,neither win nor lose anything when 4 comes up, win £1 if 5 comes up, win £2 if six comes up. These winnings are encoded in the random variable with range $X(\Omega) = \{-1, 0, 1, 2\}$ defined by

$$X(\omega) = \begin{cases} -1 & \text{if } \omega \in \{1, 2, 3\} \\ 0 & \text{if } \omega = 4 \\ 1 & \text{if } \omega = 5 \\ 2 & \text{if } \omega = 6. \end{cases}$$

a) Give and sketch the probability mass function of X. *Solution:*

$$p_X(x) := P(X = x) = \begin{cases} 1/2 & \text{if } x = -1 \\ 1/6 & \text{if } x \in \{0, 1, 2\} \\ 0 & \text{if } x \notin \{-1, 0, 1, 2\}. \end{cases}$$

The plot is shown in Figure 7.1 a).

b) Give and sketch the distribution function of X. *Solution:*

$$F_X(x) := P(X \leq x) = \begin{cases} 0 & \text{if } x < -1 \\ 1/2 & \text{if } -1 \leq x < 0 \\ 2/3 & \text{if } 0 \leq x < 1 \\ 5/6 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } 2 \leq x. \end{cases}$$

The plot is shown in Figure 7.1 b).

c) Calculate the expected gain $E[X]$. *Solution:*

$$E[X] = -1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} = 0.$$

Now assume that the government imposes a 50% tax on all gambling transactions, so that the tax income is given by the random variable

$$T = \frac{1}{2}|X|.$$

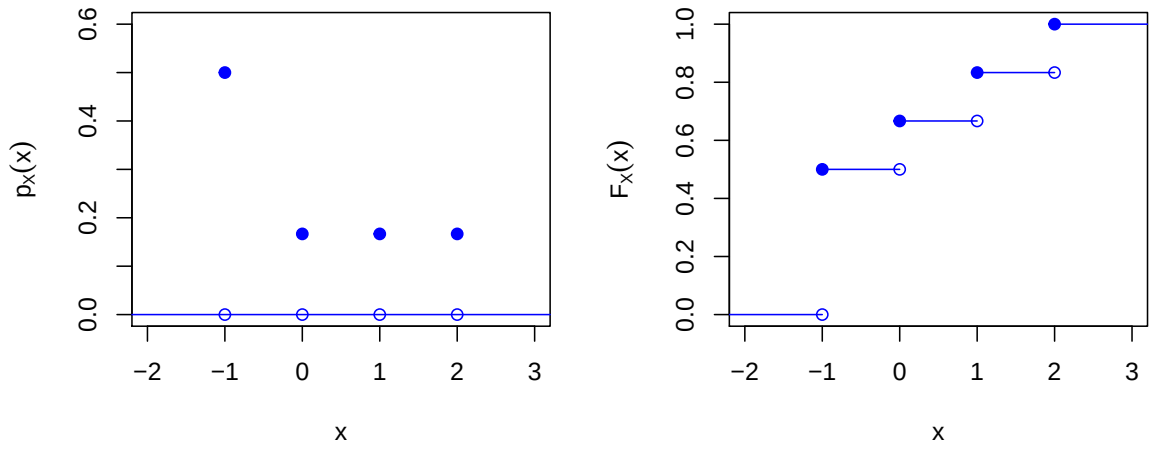


Figure 7.1: Probability mass function $p_X(x)$ and distribution function $F_X(x)$ for your winnings X in Example 7.5.

d) Give and sketch the mass function of T. *Solution:*

$$p_T(t) := P(T = t) = \begin{cases} 2/3 & \text{if } t = 1/2 \\ 1/6 & \text{if } t \in \{0, 1\} \\ 0 & \text{if } t \notin \{0, 1/2, 1\}. \end{cases}$$

The plot is shown in Figure 7.2 a).

e) Give and sketch the distribution function of T. *Solution:*

$$F_T(t) := P(T \leq t) = \begin{cases} 0 & \text{if } t < 0 \\ 1/6 & \text{if } 0 \leq t < 1/2 \\ 5/6 & \text{if } 1/2 \leq t < 1 \\ 1 & \text{if } 1 \leq t. \end{cases}$$

The plot is shown in Figure 7.2 b).

f) Calculate the expected tax income. *Solution:*

$$E[T] = 0 \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{2}{3} + 1 \cdot \frac{1}{6} = \frac{1}{2}.$$

If $X : \Omega \to \mathbb{R}$ is a discrete random variable and $h : \mathbb{R} \to \mathbb{R}$ is a function, then we would like to be able to compute $E[h(X)]$. We may do this as follows.

Theorem 7.6. *If X is a discrete random variable and $h : \mathbb{R} \to \mathbb{R}$ a function so that*

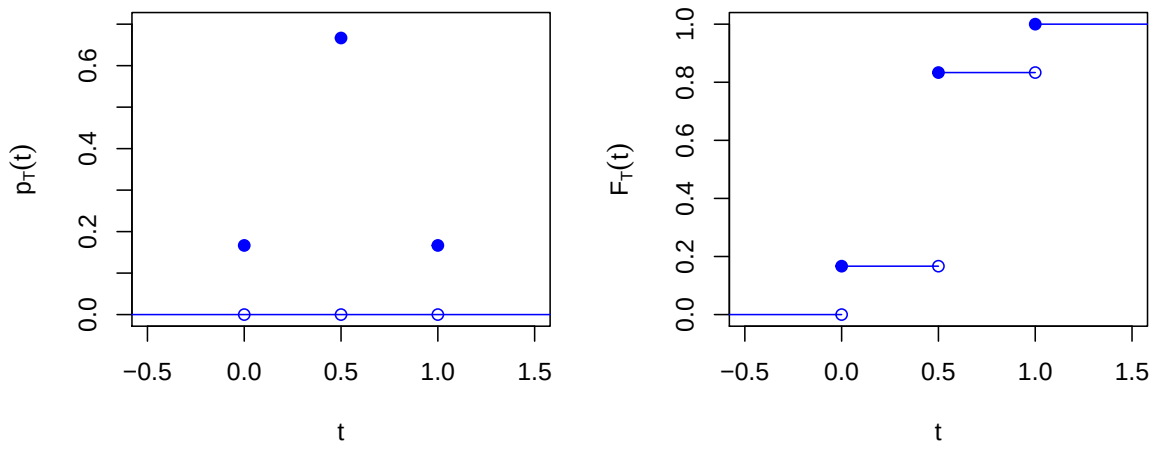


Figure 7.2: Probability mass function $p_T(t)$ and distribution function $F_T(t)$ for the tax T in Example 7.5.

$Y = h(X) = h \circ X$ is a random variable, then

$$E\left[h(X)\right] = \sum_{x \in X(\Omega)} h(x)\, p_X(x) \tag{7.2}$$

if this series is absolutely convergent.

Proof. The only challenge in this proof lies in the notation. If you understood the previous example then you will have an intuitive understanding of why this theorem is true and can skip the proof. We have by Definition 7.1 that

$$E\left[h(X)\right] = \sum_{y \in h(X(\Omega))} yP\left(h(X) = y\right).$$

The probability that $h(x) = y$ is a sum over all the probabilities of all the possible values of X that get mapped to y by h,

$$P\left(h(X) = y\right) = \sum_{\substack{x \in X(\Omega) \\ h(x) = y}} P\left(X = x\right).$$

Thus

$$\begin{aligned} E\left[Y\right] &= \sum_{y \in h(X(\Omega))} y \sum_{\substack{x \in X(\Omega) \\ h(x) = y}} P\left(X = x\right) \\ &= \sum_{y \in h(X(\Omega))} \sum_{\substack{x \in X(\Omega) \\ h(x) = y}} h(x) P\left(X = x\right) \\ &= \sum_{x \in X(\Omega)} h(x)\, p_X(x). \end{aligned}$$

□

Example 7.7. Suppose that $X \sim \text{Pois}(\lambda)$, and we wish to find the expectation of the random variable $Y = e^X$. Taking $h(x) = e^x$ in equation (7.2) we obtain

$$E\left[e^X\right] = \sum_{k=0}^{\infty} e^k p_X(k) = \sum_{k=0}^{\infty} e^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e\lambda)^k}{k!} = e^{-\lambda} e^{e\lambda} = e^{\lambda(e-1)}.$$

7.2 Expectation for continuous random variables

Definition 7.8. If X is a continuous random variable with density function f_X, then the **expectation** of X, denoted once again by $E[X]$, is defined as

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{7.3}$$

whenever the integral converges absolutely[6].

Example 7.9. If $X \sim \text{U}(a, b)$, then

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_a^b \frac{x}{b-a}\, dx \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}\,. \end{aligned}$$

Example 7.10. If $X \sim \text{N}(\mu, \sigma^2)$, then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

We use the change of variable

$$z = \frac{x-\mu}{\sigma}\,.$$

[6] *i.e.* $\int_{-\infty}^{\infty} |x|\, f_X(x) dx < \infty$. In this module the extra requirement about the absolute convergence of the integral will not be discussed further. Random variables that satisfy this requirement are called integrable random variables. It should be understood that all following theorems involving expectations of random variables are applicable only when the random variables are integrable.

Now,

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z+\mu)e^{-\frac{1}{2}z^2}dz \\ &= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2}dz + \mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz \\ &= -\frac{\sigma}{\sqrt{2\pi}}\left[e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} + \mu = \mu, \end{aligned}$$

because the second integral is 1 by (d2), since the integrand is the density function of a $\mathrm{N}(0,1)$ random variable.

If X is a continuous random variable and $h:\mathbb{R}\to\mathbb{R}$ is a (Borel[7]) function (such that $h(X)$ is integrable) then we again have a formula for the expectation of $h(X)$ (compare with the discrete case in Theorem 7.6).

Theorem 7.11. *If X is a continuous random variable with density function f_X and $h:\mathbb{R}\to\mathbb{R}$ is a function, then*

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f_X(x)dx. \tag{7.4}$$

We omit the proof.

Example 7.12. If $X\sim \mathrm{U}(0,1)$ then, letting $h(x)=1/(x+1)$, we can calculate

$$E\left[\frac{1}{X+1}\right] = \int_0^1 \frac{1}{x+1}dx = \left[\log(x+1)\right]_0^1 = \log 2\,.$$

Definition 7.13. The **m-th moment** of a random variable X is the value $E[X^m]$.

Example 7.14. Let $X\sim\mathrm{Exp}(\lambda)$. Then

$$E[X^m] = \frac{m!}{\lambda^m} \quad \text{for all } m\in\mathbb{N}\cup\{0\}.$$

We can show this by induction. If you have not yet met the idea of proof by induction, you can revisit this after you have covered this in other modules.

The statement is true for $m=0$:

$$E[X^0] = E[1] = 1 = \frac{0!}{\lambda^0}.$$

[7]This restriction again has to do with the measurability of random variables, a concept that we agreed not to go into in this module.

We next show that if the statement holds for some $k \in \mathbb{N} \cup \{0\}$ then it holds for $k+1$:

$$\begin{aligned}
E[X^{k+1}] &= \int_{-\infty}^{\infty} x^{k+1} f_X(x) dx = \int_0^{\infty} x^{k+1} \lambda e^{-\lambda x} dx \\
&= \int_0^{\infty} x^{k+1} \frac{d}{dx}\left(-e^{-\lambda x}\right) dx \\
&= -\left[x^{k+1} e^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} (k+1) x^k e^{-\lambda x} dx \\
&= 0 + \frac{k+1}{\lambda} \int_0^{\infty} x^k \lambda e^{-\lambda x} dx \\
&= \frac{k+1}{\lambda} E[X^k] = \frac{k+1}{\lambda} \frac{k!}{\lambda^k} \quad \text{by induction hypothesis} \\
&= \frac{(k+1)!}{\lambda^{k+1}}.
\end{aligned}$$

Thus the statement holds for all $m \in \mathbb{N}$ by induction. In particular $E[X] = 1/\lambda$.

The next example shows that the expectation is not defined for all random variables.

Example 7.15. Let $X \sim \text{Par}(\alpha)$. Then

$$E[X] = \int_{\infty}^{\infty} x\, f_X(x)\, dx = \int_1^{\infty} x\, \frac{\alpha}{x^{\alpha+1}}\, dx = \int_1^{\infty} \alpha\, x^{-\alpha}\, dx.$$

If $\alpha = 1$ then this formula is not applicable because

$$E[X] = \int_1^{\infty} x^{-1} dx = [\log x]_1^{\infty} = \infty.$$

In this case we say that the expectation is undefined. If $\alpha \neq 1$ then the formula gives

$$E[X] = \alpha \int_1^{\infty} x^{-\alpha} dx = \frac{\alpha}{1-\alpha} \left[x^{1-\alpha}\right]_1^{\infty}.$$

We see that the integral does not converge when $\alpha < 1$ and thus the expectation is undefined. However when $\alpha > 1$ then

$$E[X] = \frac{\alpha}{\alpha - 1}.$$

We now give two theorems that will be extremely important when we work with expectations in the future:

Theorem 7.16. *(Linearity of expectation) Let X be a random variable. Then for any $a, b \in \mathbb{R}$*

$$E\left[aX + b\right] = aE\left[X\right] + b$$

Proof. If X is a continuous random variable we can use Theorem 7.11 with $h(x) = ax+b$:

$$\begin{aligned} E[aX + b] &= \int_{-\infty}^{\infty} (ax + b) f_X(x)\, dx \\ &= a \int_{-\infty}^{\infty} x\, f_X(x)\, dx + b \int_{-\infty}^{\infty} f_X(x)\, ds \\ &= aE[X] + b. \end{aligned}$$

Here we used the linearity of the integral. If X is a discrete random variable then we can do a similar proof using Theorem 7.6. We have not defined the expectation for random variables that are neither discrete nor continuous, so we can not really give a proof in that case. □

Theorem 7.17. *Let X be a random variable. Then for any (Borel) functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$,*

$$E\left[h_1(X) \pm h_2(X)\right] = E\left[h_1(X)\right] \pm E\left[h_2(X)\right]$$

whenever these expectations are defined.

Proof. Again we can use either Theorem 7.11 or Theorem 7.6 with $h(x) = h_1(x)+h_2(x)$. I'll leave the details for you to work out. □

7.3 Variance

The expectation $E[X]$ of a random variable is an indication of the "centre" of the distribution of X. Another important quantity associated with X is the *variance* of X, and this is a measure of the degree of dispersion of X about its expectation $E[X]$. Roughly, if the dispersion of X about its expectation is on average very small/large then $|X - E[X]|$ is on average small/large, giving that $E\left[(X - E[X])^2\right]$ is small/large.

Definition 7.18. The **variance** of a random variable X is

$$\operatorname{Var}[X] = E\left[(X - E[X])^2\right]. \tag{7.5}$$

whenever these expectations are defined. The **standard deviation** of X is then defined to be the (positive) square root of $\operatorname{Var}[X]$:

$$\operatorname{sd}[X] = \sqrt{\operatorname{Var}[X]}.$$

The above definition is rarely the most convenient way to calculate the variance of a random variable. The following Theorem gives a very useful identity.

Theorem 7.19.

$$\operatorname{Var}[X] = E\left[X^2\right] - E[X]^2 \tag{7.6}$$

for any random variable X where the variance is defined.

Proof. Recall that $E[X]$ is just a constant number (not a random variable). Thus, $E[E[X]] = E[X]$ and $E[XE[X]] = E[X]E[X] = E[X]^2$ by Thm. 7.16. These together with Thm. 7.17 give

$$\begin{aligned}\operatorname{Var}[X] &= E\left[(X - E[X])^2\right] = E\left[X^2 - 2XE[X] + E[X]^2\right] \\ &= E\left[X^2\right] + E\left[-2XE[X]\right] + E\left[E[X]^2\right] \\ &= E\left[X^2\right] - 2E[X]^2 + E[X]^2 \\ &= E\left[X^2\right] - E[X]^2\end{aligned}$$

□

Now let us calculate variances in some examples.

Example 7.20. (Example 7.5 continued) For the random variable X we found $E[X] = 0$. Thus $X - E[X] = X$ and

$$\operatorname{Var}[X] = E[(X - E[X])^2) = E[X^2] = (-1)^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} = \frac{4}{3}.$$

For the random variable T we found $E[T] = 1/2$. We can similarly calculate $E[T^2]$:

$$E[T^2] = \left(\frac{1}{2}\right)^2 \frac{2}{3} + 1^2 \frac{1}{6} = \frac{1}{3}.$$

Then

$$\operatorname{Var}[T] = E[T^2] - E[T]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

Example 7.21. Let $X \sim \operatorname{Ber}(p)$. Then

$$E\left[X^2\right] = \sum_{k=0}^{1} k^2 p_X(k) = p.$$

So,

$$\operatorname{Var}[X] = E\left[X^2\right] - E[X]^2 = p - p^2 = p(1-p).$$

Example 7.22. If $X \sim \operatorname{Exp}(\lambda)$, then, using the moment calculated in Example 7.14,

$$\operatorname{Var}[X] = E\left[X^2\right] - E[X]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = 1/\lambda^2.$$

Example 7.23. Let $X \sim \text{Geo}(p)$. Then, writing $q = 1 - p$ again:

$$\begin{aligned} E\left[X^2\right] &= \sum_{k=1}^{\infty} k^2 q^{k-1} p = p \sum_{k=0}^{\infty} \frac{d}{dq}\left(kq^k\right) \\ &= p\,\frac{d}{dq}\left(\sum_{k=0}^{\infty} kq^k\right) = p\,\frac{d}{dq}\left(\frac{q}{1-q} E\left[X\right]\right) \\ &= p\,\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right) \\ &= p\left[\frac{1}{p^2} + \frac{2(1-p)}{p^3}\right] = \frac{2}{p^2} - \frac{1}{p}\,. \end{aligned}$$

Therefore

$$\text{Var}\left[X\right] = \frac{1-p}{p^2}\,.$$

Example 7.24. Let $X \sim \text{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} \text{Var}\left[X\right] &= E\left[(X - E[X])^2\right] = E\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx. \end{aligned}$$

Again we make a change of variable to $z = (x - \mu)/\sigma$. This gives

$$\begin{aligned} \text{Var}\left[X\right] &= \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2}\right) \sigma dz \\ &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp\left(-\frac{z^2}{2}\right) dz \\ &= -\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \left(\frac{d}{dz} \exp\left(-\frac{z^2}{2}\right)\right) dz \\ &= -\frac{\sigma^2}{\sqrt{2\pi}} \left(\left[z \exp\left(-\frac{z^2}{2}\right)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz\right) \\ &= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\ &= \sigma^2 \int_{-\infty}^{\infty} \phi(z) dz = \sigma^2, \end{aligned}$$

where $\phi(z)$ is the density function of the standard normal distribution $N(0, 1)$ and for the last equality we used that any density function integrated over the whole real line gives one.

Unlike expectation, variance is not a linear operator in the sense that $\text{Var}\left[aX + b\right] \neq a\text{Var}\left[X\right] + b$. Instead we have

Theorem 7.25.

$$\text{Var}\left[aX + b\right] = a^2 \text{Var}\left[X\right]$$

for any random variable X with $\mathrm{Var}[X] < \infty$ and any real numbers a and b.

Proof. We use Theorems 7.19, 7.16 and 7.17 to calculate that

$$\begin{aligned}\mathrm{Var}[aX+b] &= E\left[(aX+b)^2\right] - E[aX+b]^2 \\ &= E\left[a^2X^2 + 2abX + b^2\right] - (aE[X]+b)^2 \\ &= a^2E\left[X^2\right] + 2abE[X] + b^2 - \left(a^2E[X]^2 + 2abE[X] + b^2\right) \\ &= a^2\left(E\left[X^2\right] - E[X]^2\right) \\ &= a^2\mathrm{Var}[X].\end{aligned}$$

□

Example 7.26. If $X \sim \mathrm{U}(0,1)$, then

$$E\left[X^2\right] = \int_0^1 \frac{x^2}{b-a}\,dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}.$$

Therefore,

$$\mathrm{Var}[X] = E\left[X^2\right] - E[X]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

The transformed random variable $Y = (b-a)X + a$ is still uniformly distributed, but now on the interval from a to b, i.e., $Y \sim \mathrm{U}(a,b)$. We thus can use Theorem 7.25 to calculate the variance of a $\mathrm{U}(a,b)$ random variable:

$$\mathrm{Var}[Y] = (b-a)^2\mathrm{Var}[X] = \frac{(b-a)^2}{12}.$$

Example 7.27. Consider a random variable $X \sim \mathrm{U}(-1,1)$ and another random variable Y whose density function f_y is also vanishing outside the interval $[-1,1]$ but on that interval is given by an upside-down parabola, symmetric around the y axis and reaching zero at $x = -1$ and $x = 1$. Which is bigger, $\mathrm{Var}[X]$ or $\mathrm{Var}[Y]$? Calculate $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$.

Solution. The variance of X will be larger than that of Y because Y has smaller probability density at the values that are further away from the mean. We can calculate the variance of X using the result from the previous example:

$$\mathrm{Var}[X] = \frac{(1-(-1))^2}{12} = \frac{1}{3}.$$

To calculate the variance of Y we first needs its density function. From the description we know that it is

$$f_Y(y) = \begin{cases} \frac{3}{4}(1-y^2) & \text{if } y \in [-1,1] \\ 0 & \text{otherwise.} \end{cases}$$

The factor of $3/4$ in front of $1-y^2$ ensures that the integral under the density function is equal to 1. The expectation of Y is zero due to the symmetry of the density function:

$$E[Y] = \int_{-\infty}^{\infty} y\, f_Y(y) dy = \frac{3}{4}\int_{-1}^{1} y\,(1-y^2)\, dy = 0.$$

We also calculate

$$E[Y^2] = \int_{-\infty}^{\infty} y^2\, f_Y(y) dy = \frac{3}{4}\int_{-1}^{1} y^2\,(1-y^2)\, dy = \frac{3}{4}\left[\frac{y^3}{3} - \frac{y^5}{5}\right]_{-1}^{1} = \frac{3}{4}\,\frac{4}{15} = \frac{1}{5}.$$

Therefore

$$\mathrm{Var}\,[Y] = E[Y^2] - E[Y]^2 = \frac{1}{5}.$$

In order to find the variance of discrete random variables it is sometimes useful to employ the following trick: we apply (7.2) with $h(x) = x(x-1)$, noting that for any random variable with image $\{0, 1, 2, ...\}$,the first two terms of the sum on the right hand side are 0. Here is this trick in action.

Example 7.28. When $X \sim \mathrm{Pois}(\lambda)$ let us show that $\mathrm{Var}\,[X] = \lambda$. For $h(x) = x(x-1)$, (7.2) gives

$$\begin{aligned} E\,[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) p_X(k) = e^{-\lambda} \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda}\lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2, \end{aligned}$$

by letting $k-2 = j$. Using linearity of expectation and the fact that $E\,[X] = \lambda$,

$$\lambda^2 = E\,[X(X-1)] = E\left[X^2 - X\right] = E\left[X^2\right] - E\,[X] = E\left[X^2\right] - \lambda\,,$$

and so $E\,[X^2] = \lambda^2 + \lambda$. Hence $\mathrm{Var}\,[X] = E\,[X^2] - E\,[X]^2 = \lambda$, as claimed.

8 Computations with random variables

(Textbook chapter link)

In this chapter we will investigate how to obtain the probability distribution of a transformed random variable Y that is obtained by applying a function h to another random variable X with known distribution. So we consider $Y = h(X)$ where X is a random variable and $h : \mathbb{R} \to \mathbb{R}$ is a function.

Before we get going, we need to introduce the support of a random variable:

Definition 8.1. The **support** of a random variable X is the smallest set R_X so that $P(X \in R_X) = P(X^{-1}(R_X)) = 1$.

From this definition, we can show that the support of a discrete random variable is the set

$$R_X = \{x \in \mathbb{R} \,|\, p_X(x) > 0\},$$

and the support of a continuous random variable is the set

$$R_X = \{x \in \mathbb{R} \,|\, f_X(x) > 0\}.$$

8.1 Transforming discrete random variables

We already discussed in Example 7.5 the case where X is a discrete random variable. Here is another example:

Example 8.2. Let $X \sim \text{Bin}(3, 1/2)$ and $Y = h(X)$ with $h(x) = \sin(\pi x/2)$. Determine the probability mass function of Y.

Solution. Recall that the binomial distribution $\text{Bin}(n, p)$ describes the number of successes in a series of n trials, with each individual trial independently having a probability p of succeeding. Thus in the case of $n = 3$ there are four possible values for the variable X, namely $X(\Omega) = \{0, 1, 2, 3\} = R_X$. The non-zero values of the probability mass function of X are $p_X(0) = p_X(3) = 1/8, p_X(1) = p_X(2) = 3/8$, according to Definition 4.11.

The support of the random variable $Y = h(X)$ is then

$$R_Y = \{h(0), h(1), h(2), h(3)\} = \{0, 1, 0, -1\} = \{-1, 0, 1\}.$$

The probability mass function is

$$p_Y(y) = P(Y = y) = p(h(X) = y) = \sum_{\substack{x \in X(\Omega) \\ h(x) = y}} P_X(x).$$

This is really all there is to it: to get the probability that Y takes on a particular value y we have to sum up the probabilities of all those values of X that give $h(x) = y$. In our example this gives

$$p_Y(-1) = p_X(3) = \frac{1}{8}, \quad p_Y(0) = p_X(0) + p_X(2) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}, \quad p_Y(1) = p_X(1) = \frac{3}{8}$$

and $p_Y(y) = 0$ if $y \notin \{-1, 0, 1\}$.

8.2 Transforming continuous random variables

The above applies to discrete random variables only, because only they have a probability mass function. If we want to deal with an arbitrary random variable, we have to work with the probability distribution function instead.

Example 8.3. Let X be a continuous random variable with support $R_X = [0, 1]$ and distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \in [0, 1] \\ 1 & \text{if } x > 1. \end{cases}$$

Determine the distribution functions of the random variables $Y = 3X + 2$ and $Z = -X^2$.

Solution. We write $Y = h(X)$ with $h(x) = 3x + 2$. The support of Y is

$$R_Y = h(R_X) = h([0, 1]) = [2, 5].$$

The distribution function is

$$F_Y(y) = P(Y \leq y) = P(3X + 2 \leq y) = P(X \leq (y - 2)/3) = F_X((y - 2)/3).$$

Thus, using the known expression for F_X, we find that

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 2 \\ \dfrac{(y-2)^2}{9} & \text{if } y \in [2, 5] \\ 1 & \text{if } y > 5. \end{cases}$$

Similarly we write $Z = g(X)$ with $g(x) = -x^2$. The support of Z is

$$R_Z = g(R_X) = g([0, 1]) = [-1, 0].$$

The distribution function is

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(-X^2 \leq z) = P(X \geq \sqrt{-z}) = 1 - P(X < \sqrt{-z}) \\ &= 1 - P(X \leq \sqrt{-z}) + P(X = \sqrt{-z}) = 1 - F_X(\sqrt{-z}). \end{aligned}$$

Thus, using the known expression for F_X, we find that

$$F_Z(z) = \begin{cases} 0 & \text{if } z < -1 \\ 1 - (\sqrt{-z})^2 = 1 + z. & \text{if } z \in [-1, 0] \\ 1 & \text{if } z > 0. \end{cases}$$

From the above examples we can already see the general case:

Theorem 8.4. *Let X be a random variable and let $Y = h(X)$ for some function $h : \mathbb{R} \to \mathbb{R}$. If h is strictly increasing on R_X then*

$$F_Y(y) = F_X(h^{-1}(y)) \quad \text{for all } y \in R_Y. \tag{8.1}$$

If h is strictly decreasing on R_X then

$$F_Y(y) = 1 - F_X(h^{-1}(y)) + P(X = h^{-1}(y)) \quad \text{for all } y \in R_Y,$$

which by Thm. 5.4 simplifies to

$$F_Y(y) = 1 - F_X(h^{-1}(y)) \quad \text{for all } y \in R_Y. \tag{8.2}$$

when X is a continuous random variable.

For functions that are not strictly monotonic it is difficult to write down such a general formula.

In the case of continuous random variables we are also interested in the density function of the transformed variable. This can be obtained by differentiating its distribution function. In the case where h is strictly increasing, we differentiate equation (8.1) to get

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(h^{-1}(y)\right) \quad \text{for all } y \in R_Y.$$

We now use the chain rule to evaluate the derivative on the right-hand side, giving

$$f_Y(y) = F_X'\left(h^{-1}(y)\right) \frac{d}{dy} h^{-1}(y),$$

where F_X' is the derivative of the distribution function of X. That derivative gives us

the density function of X. We also use that we can express the derivative of an inverse function in terms of the derivative of the function itself. This gives us

$$f_Y(y) = f_X\left(h^{-1}(y)\right)\frac{1}{h'(h^{-1}(y))}. \tag{8.3}$$

In the case where h is strictly decreasing, we find, using equation (8.2), that

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - F_X(h^{-1}(y))\right) = -f_X\left(h^{-1}(y)\right)\frac{1}{h'(h^{-1}(y))}. \tag{8.4}$$

We can combine the two equations (8.3) and (8.4) by observing that in the decreasing case $h'(x)$ is negative and thus $-h'(x) = |h(x)|$.

We formulate the result as a theorem:

Theorem 8.5. *Let X be a continuous random variable and let a function $h : \mathbb{R} \to \mathbb{R}$ be differentiable and strictly monotonic on R_X. Then the density function of $Y = h(X)$ is given by*

$$f_Y(y) = \begin{cases} \dfrac{f_X\left(h^{-1}(y)\right)}{|h'(h^{-1}(y))|} & \text{if } y \in R_Y \\ 0 & \text{otherwise.} \end{cases}$$

Example. 8.2 continued For the random variables $Y = 3X + 2$ and $Z = -X^2$ determine the density functions f_Y and f_Z using Theorem 8.5.

Solution. The density function of the random variable X is obtained as the derivative of the distribution function given earlier:

$$f_X(x) = \frac{d}{dx}F_X(x) = \begin{cases} 2x & \text{if } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

We had $Y = h(X)$ with $h(x) = 3x + 2$. So $h'(x) = 3$ and $h^{-1}(y) = (y-2)/3$. Then according to Thm. 8.5 for $y \in [2,5]$ we have

$$f_Y(y) = \frac{f_X\left(h^{-1}(y)\right)}{|h'(h^{-1}(y))|} = \frac{2h^{-1}(y)}{|3|} = \frac{2(y-2)/3}{3} = \frac{2}{9}(y-2).$$

We can verify that this is correct by comparing with the result of directly differentiating the distribution function of Y:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\frac{(y-2)^2}{9} = \frac{2}{9}(y-2) \text{ for } y \in [2,5].$$

Of course outside the image of Y its density function is 0.

We also have $Z = h(X)$ with $h(x) = -x^2$. So $h'(x) = -2x$ and $h^{-1}(z) = \sqrt{-z}$ and

thus by Theorem 8.5 for $z \in [-1, 0]$ we have

$$f_Z(z) = \frac{f_X\left(g^{-1}(z)\right)}{|h'(h^{-1}(z))|} = \frac{2h^{-1}(z)}{|-2h^{-1}(z)|} = \frac{2\sqrt{-z}}{|-2\sqrt{-z}|} = 1.$$

Again we can check this by comparing to the derivative of the distribution function:

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}(1+z) = 1 \text{ for } z \in [-1, 0].$$

So we see that $Z \sim \mathrm{U}[-1, 0]$.

Theorem 8.6. *Let X be a continuous random variable and $r, s \in \mathbb{R}$ with $r > 0$. Introduce $Y = rX + s$. Then*

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right), \tag{8.5}$$

$$f_Y(y) = f_X\left(\frac{y-s}{r}\right)\frac{1}{|r|}. \tag{8.6}$$

Proof. Let $h(x) = rx + s$. Because $r > 0$ this is a strictly increasing function, so we can use Equation 8.1 from Thm. 8.4 to obtain the distribution function and Thm 8.5 to obtain the density function of $Y = h(X) = rX + s$. Using that $h^{-1}(y) = (y-s)/r$ and $h'(x) = r$ we obtain

$$F_Y(y) = F_X\left(h^{-1}(y)\right) = F_X\left(\frac{y-s}{r}\right), \tag{8.7}$$

$$f_Y(y) = f_X\left(h^{-1}(y)\right)\frac{1}{h'(h^{-1}(y))} = f_X\left(\frac{y-s}{r}\right)\frac{1}{r}. \tag{8.8}$$

□

A linear transformation of this form turns out to transform any normally distributed random variable into another random variable with transformed mean and variance.

Theorem 8.7. *If $X \sim \mathrm{N}(\mu, \sigma^2)$ then*

$$Y = rX + s \sim \mathrm{N}\left(r\mu + s, (r\sigma)^2\right).$$

Proof. From Thm. 8.6 we obtain

$$f_Y(y) = f_X\left(\frac{y-s}{r}\right)\frac{1}{r}.$$

Subsituting the density function of X from Def. 5.8 gives

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-s}{r} - \mu\right)^2}{2\sigma^2}\right) \frac{1}{r} \\ &= \frac{1}{\sqrt{2\pi}r\sigma} \exp\left(-\frac{(y-s-r\mu)^2}{2(r\sigma)^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left(-\frac{(y-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right) \end{aligned}$$

with $\tilde{\sigma} = r\sigma$ and $\tilde{\mu} = r\mu + s$. We recognize this as the density function of an $\mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$ distribution. Thus $Y \sim \mathrm{N}(r\mu + s, (r\sigma)^2)$ as claimed. □

In particular, this theorem allows one to transform any normally distributed random variable to a standard normal random variable by choosing $r = 1/\sigma$ and $s = -\mu/\sigma$.

Corollary 8.8. *If $X \sim \mathrm{N}(\mu, \sigma^2)$ then*

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

8.3 Extrema (Optional)

Theorem 8.9. *Let $X_1, X_2, \ldots, X_n$ be independent random variables and let $Y = \max\{X_1, X_2, \ldots, X_n\}$ and $V = \min\{X_1, X_2, \ldots, X_n\}$. Then*

$$F_Y(y) = F_{X_1}(y)F_{X_2}(y)\cdots F_{X_n}(y), \tag{8.9}$$

$$F_V(v) = 1 - (1 - F_{X_1}(v))(1 - F_{X_2}(v))\cdots(1 - F_{X_n}(v)). \tag{8.10}$$

Proof.

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(\max\{X_1, X_2, \ldots, X_n\} \le y) \\ &= P(X_1 \le y, X_2 \le y, \ldots, X_n \le y), \end{aligned}$$

where the last equality expresses the fact that the maximum of the variables can only be less than or equal to y if all of the variables individually are less than or equal to y. We now use the independence of the random variables $X_1, \ldots, X_n$. The concept of independence of random variables will only get formally introduced in the next chapter, but it is intuitive enough to quickly introduce it here: it means that all events that can be defined in terms of those random variables are independent. So in particular the events $\{X_1 \le y\}, \{X_2 \le y\}, \ldots \{X_n \le y\}$ are all independent and hence the probability

of their intersection factorizes, giving

$$\begin{aligned} F_Y(y) &= P(X_1 \leq y)P(X_2 \leq y)\cdots P(X_n \leq y) \\ &= F_{X_1}(y)F_{X_2}(y)\cdots F_{X_n}(y). \end{aligned}$$

For V we have

$$\begin{aligned} F_V(v) &= P(V \leq v) = 1 - P(V > v) = 1 - P(\min\{X_1, X_2, \ldots, X_n\} > v) \\ &= 1 - P(X_1 > v, X_2 > v, \ldots, X_n > v). \end{aligned}$$

Again we can use independence to factorize the joint probability and we find

$$\begin{aligned} F_V(v) &= 1 - P(X_1 > v)P(X_2 > v)\cdots P(X_n > v) \\ &= 1 - (1 - P(X_1 \leq v))(1 - P(X_2 \leq v))\cdots(1 - P(X_n \leq v) \\ &= 1 - (1 - F_{X_1}(v))(1 - F_{X_2}(v))\cdots(1 - F_{X_n}(v). \end{aligned}$$

□

Example 8.10. Let $X \sim U(0,1)$ and $Y \sim \text{Exp}(1)$ be independent random variables and let Z be the random variable $Z = \max(X, Y)$. Give the distribution function of Z.

Solution. From section 5.2 we know that $X \sim U(0,1)$ means

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0,1] \\ 1 & \text{if } x > 1 \end{cases}$$

and that $Y \sim \text{Exp}(1)$ means

$$F_Y(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$$

We can now use equation (8.9):

$$F_Z(x) = F_X(x)F_Y(x) = \begin{cases} 0 & \text{if } x < 0 \\ x(1 - e^{-x}) & \text{if } x \in [0,1] \\ 1 - e^{-x} & \text{if } x > 1 \end{cases}$$

9 Joint distributions and independence

(Textbook chapter link)

In this chapter we will consider the case where we are interested in two or more random variables together. Many of the definitions and results in this chapter are direct generalisations of corresponding definitions and results for the single-variable case.

9.1 Joint distributions of discrete random variables

Example 9.1. Consider an experiment consisting of flipping two fair coins. Thus the sample space is $\Omega = \{(HH), (HT), (TH), (TT)\}$ and all these outcomes are equally likely. Introduce two random variables N and S. N is equal to the total number of heads and $S = 1$ if the second coin shows head and $S = 0$ otherwise. This is summarised in the following table:

ω	(HH)	(HT)	(TH)	(TT)
S	1	0	1	0
N	2	1	1	0

We can calculate the probability mass function of S. It is non-zero only on the image $S(\Omega) = \{0, 1\}$ and its non-zero values give the probabilities of the events

$$\{S = 0\} = \{(H, T), (T, T)\}, \quad \{S = 1\} = \{(H, H), (T, H)\} = \{S = 0\}^c$$

and thus

$$p_S(0) = P(S = 0) = P((H, T), (T, T)) = 1/2, \quad p_S(1) = 1 - p_S(0) = 1/2.$$

The mass function vanishes everywhere else: $p_S(x) = 0$ if $x \notin S(\Omega)$.

Similarly we have $N(\Omega) = \{0, 1, 2\}$ so its non-zero values give the probabilities of the events $\{N = 0\} = \{(T, T)\}$, $\{N = 1\} = \{(H, T), (T, H)\}$, $\{N = 2\} = \{(H, H)\}$ and thus

$$p_N(0) = P(\{(T, T)\}) = 1/4, \quad p_N(2) = P(\{(H, H)\}) = 1/4,$$
$$p_N(1) = P(\{(H, T), (T, H)\}) = 1/2 = 1 - p_N(0) - p_N(2).$$

But now we can also consider events defined in terms of both random variables simultaneously, like $\{S = 0 \text{ and } N = 1\}$. For convenience we will often replace the 'and' by a comma, writing $\{S = 0, N = 1\}$. We find

$$\{S = 0, N = 1\} = \{S = 0\} \cap \{N = 1\} = \{(H, T), (T, T)\} \cap \{(H, T), (T, H)\} = \{(H, T)\}$$

and thus

$$P(S = 0, N = 1) = P(\{(H, T)\}) = 1/4.$$

The joint probability mass function is a device to summarise the probability of such events defined in terms of values for both random variables:

Definition 9.2. (compare to Def. 4.4) Given two discrete random variables X and Y, the function $p_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$p_{X,Y}(x, y) = P\left(X^{-1}(x) \cap Y^{-1}(y)\right) = P\left(X = x, Y = y\right)$$

is called the **joint probability mass function** of X and Y.

Example 9.1 cont.: We have already started calculating the joint probability mass function for S and N above:

$$P_{S,N}(0, 1) = P(S = 0, N = 1) = P(\{(H, H)\}) = 1/4.$$

Doing similar calculations we find all other non-zero values of $p_{S,N}$ as

$$\begin{aligned} p_{S,N}(0, 0) &= P(\{(T, T)\}) = \frac{1}{4}, \\ p_{S,N}(1, 1) &= P(\{(T, H)\}) = \frac{1}{4}, \\ p_{S,N}(1, 2) &= P(\{(H, H)\}) = \frac{1}{4}. \end{aligned}$$

We also find

$$p_{S,N}(0, 2) = p_{S,N}(1, 0) = P(\emptyset) = 0.$$

And clearly $p_{S,N}(s, n) = 0$ unless $s \in S(\Omega)$ and $n \in N(\Omega)$.

The values of $p_{S,N}$ can be displayed as a table:

		n		
		0	1	2
s	0	1/4	1/4	0
	1	0	1/4	1/4

It is convenient to also include the probability mass functions of S and N in the margins of the table:

		n			
		0	1	2	$p_S(s)$
s	0	1/4	1/4	0	1/2
	1	0	1/4	1/4	1/2
	$p_N(n)$	1/4	1/2	1/4	

(9.1)

Because of this convention of displaying the mass functions of the individual random variables in the margins of the table they are also often referred to as the marginal probability mass functions. Note how they can be obtained by summing up the values of the joint mass function across rows or columns. Let us summarise that in a theorem:

Theorem 9.3. *Let X and Y be discrete random variables. The probability mass functions of X and Y can be obtained as*

$$p_X(x) = \sum_{y \in Y(\Omega)} p_{X,Y}(x,y), \quad p_Y(y) = \sum_{x \in X(\Omega)} p_{X,Y}(x,y).$$

Proof. This is just a consequence of the fact that the collection of events $\{\{Y = y\} | y \in Y(\Omega)\}$ is a partition of the sample space, i.e.,

$$\bigcup_{y \in Y(\Omega)} \{Y = y_k\} = \Omega \quad \text{and } \{Y = y_1\} \cap \{Y = y_2\} = \emptyset \text{ if } y_1 \neq y_2.$$

Thus we can write the event $\{X = x\}$ as a disjoint union,

$$\{X = x\} = \{X = x\} \cap \Omega = (\{X = x\} \cap \bigcup_{y \in Y(\Omega)} \{Y = y\} = \bigcup_{y \in Y(\Omega)} (\{X = x\} \cap \{Y = y\}).$$

Therefore, by axiom (P3),

$$\begin{aligned} p_X(x) &= P\left(\bigcup_{y \in Y(\Omega)} (\{X = x\} \cap \{Y = y\}) \right) = \sum_{y \in Y(\Omega)} P\left((\{X = x\} \cap \{Y = y\}) \right) \\ &= \sum_{y \in Y(\Omega)} p_{X,Y}(x,y). \end{aligned}$$

The second identity follows similarly with X and Y interchanged. □

Note that while one can always recover the marginal probability mass functions from the joint probability mass function, the converse is certainly not true. The joint probability mass function contains much more information than is contained in the marginal probability mass functions.

Joint mass functions have two defining properties (compare with properties (m1) and (m2) of single-variable mass functions):

(jm1) $p_{XY}(x,y) \geq 0 \ \forall x, y \in \mathbb{R}$. In particular, $p_{XY}(x,y) = 0$ unless $x \in X(\Omega)$ and $y \in Y(\Omega)$.

(jm2) $\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} p_{XY}(x,y) = 1$.

Next we introduce the joint *distribution function* as an alternative way of specifying the probability distribution. This has an advantage over the probability mass function, as it will also work for continuous random variables.

Definition 9.4. (Compare to Def. 4.7) Let X and Y be random variables. The **joint distribution function** of X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F_{X,Y}(x,y) = P\left(X^{-1}(-\infty, x] \cap Y^{-1}(-\infty, y]\right) = P\left(X \leq x, Y \leq y\right).$$

Example 9.1 cont.: The joint distribution function of S and N is

$$F_{S,N}(s,n) = P(S \leq s, N \leq n) = \begin{cases} 0 & \text{if } s < 0 \text{ or } n < 0 \\ 1/4 & \text{if } 0 \leq s, 0 \leq n < 1 \\ 1/2 & \text{if } 0 \leq s < 1, 1 \leq n \\ 3/4 & \text{if } 1 \leq s, 1 \leq n < 2 \\ 1 & \text{if } 1 \leq s, 2 \leq n. \end{cases}$$

The joint distribution function of two discrete random variables is a two-dimensional step function.

We were able to get the marginal mass functions from the joint mass function by summing. How can we get the marginal density functions from the joint density function?

Theorem 9.5. *Let X and Y be random variables and let $F_{X,Y}$ be their joint distribution function. Then their (marginal) distribution functions can be obtained as*

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y), \qquad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y).$$

This is true because

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq \infty) = \lim_{y\to\infty} P(X \leq x, Y \leq y) = \lim_{y\to\infty} F_{X,Y}(x,y)$$

and similarly for F_Y.

Example 9.1 cont.:

$$F_S(s) = \lim_{n\to\infty} F_{S,N}(s,n) = F_{S,N}(s,2)$$

because $F_{S_N}(s,n) = F_{S,N}(s,2)$ for all $n \geq 2$. Hence we find

$$F_S(s) = \begin{cases} 0 & \text{if } s < 0 \\ 1/2 & \text{if } 0 \leq s < 1 \\ 1 & \text{if } s \geq 1 \end{cases}.$$

Similarly

$$F_N(n) = \lim_{s\to\infty} F_{S,N}(s,n) = F_{S,N}(1,n)$$
$$= \begin{cases} 0 & \text{if } n < 0 \\ 1/4 & \text{if } 0 \le n < 1 \\ 3/4 & \text{if } 1 \le n < 2 \\ 1 & \text{if } n \ge 2 \end{cases}.$$

9.2 Joint distributions of continuous random variables

Definition 9.6. (Compare to Def. 5.1) We call two random variables X and Y **jointly continuous** if their joint distribution function $F_{X,Y}$ can be written as

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(\hat{x},\hat{y})\, d\hat{y}d\hat{x} \quad \forall x,y \in \mathbb{R}$$

for some non-negative function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$. In this case, we say that $f_{X,Y}$ is the **joint density function** of X and Y.

Example 9.7. Consider the uniform distribution on a rectangle where the probability density is evenly spread over the rectangle. Let's take a rectangle parallel to the x and y axes so that the x-coordinate of a point in the rectangle lies in an interval $[a,b]$ and the y-coordinate in an interval $[c,d]$. Because the area of the rectangle is $(b-a)(d-c)$ and the probability density is spread evenly, the density function is

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{(b-a)(d-c)} & \text{if } x \in [a,b] \text{ and } y \in [c,d] \\ 0 & \text{otherwise.} \end{cases}$$

In the lectures we skipped the remainder of this section. It uses two-variable calculus and therefore may look more scary than it is. You are encouraged to study it but this is not required for the understanding of the rest of the module.

The joint distribution function for jointly continuous random variables is continuous and even differentiable almost everywhere. The fundamental theorem of calculus implies, under some mild regularity conditions, that for each $(x,y) \in \mathbb{R}^2$,

$$\frac{d}{dx}\frac{d}{dy}F_{X,Y}(x,y) = f_{X,Y}(x,y). \tag{9.2}$$

Joint density functions have the following two properties (compare with properties

(d1) and (d2) of single-variable density functions):

$$f_{X,Y}(x,y) \geq 0, \quad \text{for all } x,y \in \mathbb{R}\,; \tag{jd1}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dydx = 1. \tag{jd2}$$

Theorem 9.8. *(Compare to theorem 5.3) If X and Y are jointly continuous random variables with joint density function $f_{X,Y}$, then for all $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 \leq b_1$, $a_2 \leq b_2$,*

$$P(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f_{X,Y}(x,y)dydx.$$

Weak inequality can be replaced by strict inequality anywhere on the LHS of the above equation.

Theorem 9.9. *(Compare to theorem 9.3) Let $f_{X,Y}$ be the joint density of X and Y. Then their (marginal) density functions are*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Example 9.10. Suppose X and Y have joint density function

$$f_{X,Y}(x,y) = \begin{cases} xe^{-x-y} & \text{for } x \geq 0, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the joint distribution function is

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(\hat{x},\hat{y})d\hat{y}d\hat{x} \\ &= \int_0^x\int_0^y \hat{x}e^{-\hat{x}-\hat{y}}d\hat{y}d\hat{x} = \int_0^x \hat{x}e^{-\hat{x}}\int_0^y e^{-\hat{y}}d\hat{y}d\hat{x} \\ &= \left(1-e^{-y}\right)\int_0^x \hat{x}e^{-\hat{x}}d\hat{x} = \left(1-e^{-y}\right)\left(1-(1+x)e^{-x}\right) \\ &= 1-(1+x)e^{-x}-e^{-y}+(1+x)e^{-x-y}. \end{aligned}$$

We can check our calculation of the distribution function by using equation (5.1):

$$\begin{aligned} \frac{d}{dx}\frac{d}{dy}F_{X,Y}(x,y) &= \frac{d}{dx}\frac{d}{dy}\left(1-(1+x)e^{-x}-e^{-y}+(1+x)e^{-x-y}\right) \\ &= \frac{d}{dx}\left(e^{-y}-(1+x)e^{-x-y}\right) = -e^{-x-y}+(1+x)e^{-x-y} \\ &= xe^{-x-y} = f_{X,Y}(x,y). \end{aligned}$$

Using theorem 9.5 we obtain the marginal distribution functions:

$$\begin{aligned} F_X(x) &= \lim_{y\to\infty} F_{X,Y}(x,y) \\ &= \lim_{y\to\infty} \left(1-(1+x)e^{-x}-e^{-y}+(1+x)e^{-x-y}\right) \\ &= 1-(1+x)e^{-x}, \\ F_Y(y) &= \lim_{x\to\infty} F_{X,Y}(x,y) \\ &= \lim_{x\to\infty} \left(1-(1+x)e^{-x}-e^{-y}+(1+x)e^{-x-y}\right) \\ &= 1-e^{-y}. \end{aligned}$$

Using theorem 9.9 we obtain the marginal density functions. For example:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^{\infty} xe^{-x-y}dy = xe^{-x}\int_0^{\infty} e^{-y}dy = xe^{-x},$$

To check this, we can also obtain the density function as a derivative of the distribution function:

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}\left(1-(1+x)e^{-x}\right) = -e^{-x}+(1+x)e^{-x} = xe^{-x}.$$

We get the same result, as must be the case.

9.3 More than two random variables

The message is that everything we have done and will do for two random variables in this chapter generalises in a straightforward way to any number of random variables. Only the notation becomes more cumbersome.

9.4 Independent random variables

We call two random variables **independent**, if knowing the value of one of them tells us nothing about the value of the other. Luckily there is a simple way to make this idea mathematically concrete:

Definition 9.11. Two random variables X and Y are independent, $X \perp\!\!\!\perp Y$, if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \text{ for all } x,y \in \mathbb{R}.$$

If you don't happen to have the distribution functions handy, you can also check independence by looking at the probability mass function or the probability density function:

Theorem 9.12. *If X and Y are discrete random variables, they are independent if and only if*

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ for all } x,y \in \mathbb{R}.$$

If X and Y are jointly continuous random variables they are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } x,y \in \mathbb{R}.$$

Example 9.1 cont.: We observe from table (9.1) that for example

$$p_S(1)p_N(0) = \frac{1}{2}\cdot\frac{1}{4} \neq 0 = p_{S,N}(1,0).$$

This one counterexample is sufficient to show that X and Y are not independent.
Example 9.7 cont.: The density function for the x-coordinate is

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

Similarly for the y coordinate

$$f_Y(y) = \begin{cases} \dfrac{1}{d-c} & \text{if } y \in [c,d] \\ 0 & \text{otherwise.} \end{cases}$$

$$F_X(x)F_Y(y) = \begin{cases} \dfrac{1}{(b-a)(d-c)} & \text{if } x \in [a,b] \text{ and } y \in [c,d] \\ 0 & \text{otherwise.} \end{cases} = F_{X,Y}(x,y)$$

for all $x,y \in \mathbb{R}$. Hence X and Y are independent.
Example 9.10 cont.: We observe that

$$F_X(x)F_Y(y) = \left(1-(1+x)e^{-x}\right)\left(1-e^{-y}\right) = F_{X,Y}(x,y)$$

for all $x,y \in \mathbb{R}$. Hence X and Y are independent.

9.5 Propagation of independence

Theorem 9.13 (Propagation of independence)**.** *Let $X_1, X_2, \ldots X_n$ be independent random variables and $h_1, h_2, \ldots h_n : \mathbb{R} \to \mathbb{R}$ be functions. Then the random variables $h_1(X_1), h_2(X_2), \ldots h_n(X_n)$ are independent.*

10 Covariance and correlation

(Textbook chapter link)

10.1 Expectation and joint distributions

Theorem 10.1. *(Compare to theorems 7.6 and 7.11) Let X and Y be random variables and let $h : \mathbb{R}^2 \to \mathbb{R}$ be a function so that $h(X, Y)$ is a new random variable. If X and Y are discrete then*

$$E\left[h(X,Y)\right] = \sum_{x\in X(\Omega)} \sum_{y\in Y(\Omega)} h(x,y) p_{XY}(x,y). \tag{10.1}$$

If X and Y are jointly continuous with joint density function f_{XY}, then

$$E\left[h(X,Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{XY}(x,y) dy dx. \tag{10.2}$$

Example 9.1 cont.: We can use the above theorem with $h(x, y) = xy$ to calculate $E[SN]$ as follows:

$$E[SN] = \sum_{s=0}^{1} \sum_{n=0}^{2} s\, n\, p_{SN}(s,n) = 1 \cdot p_{SN}(1,1) + 2 \cdot p_{SN}(1,2) = \frac{1}{4} + 2\frac{1}{4} = \frac{3}{4}.$$

Similarly we can calculate $E[S + N]$ by using $h(x, y) = x + y$:

$$\begin{aligned} E[S+N] &= \sum_{s=0}^{1} \sum_{n=0}^{2} (s+n)\, p_{SN}(s,n) \\ &= 1 \cdot p_{SN}(0,1) + 1 \cdot p_{SN}(1,0) + 2 \cdot p_{SN}(0,2) + 2 \cdot p_{SN}(1,1) + 3 \cdot p_{SN}(1,2) \\ &= \frac{1}{4} + 0 + 2 \cdot 0 + 2\frac{1}{4} + 3\frac{1}{4} = \frac{3}{2}. \end{aligned}$$

Note that also

$$E[S] = \sum_{s=0}^{1} s\, p_S(s) = \frac{1}{2} \quad \text{and} \quad E[N] = \sum_{n=0}^{2} n\, p_N(n) = \frac{1}{2} + 2\frac{1}{4} = 1,$$

so that $E[S] + E[N] = 3/2 = E[S + N]$. That is actually no coincidence as the next theorem below shows.

Theorem 10.2 (Linearity of expectations)**.** *Let X and Y be random variables and let $r, s, t \in \mathbb{R}$. Then*

$$E[rX + sY + t] = rE[X] + sE[Y] + t.$$

Proof. Let us first proof the case where X and Y are discrete.

$$\begin{aligned} E[rX+sY+t] &= \sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)}(rx+sy+t)p_{XY}(x,y) \text{ by Thm. 10.1} \\ &= r\sum_{x\in X(\Omega)} x\left(\sum_{y\in Y(\Omega)} p_{XY}(x,y)\right) + s\sum_{y\in Y(\Omega)} y\left(\sum_{x\in X(\Omega)} p_{XY}(x,y)\right) \\ &\quad + t\sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} p_{XY}(x,y). \end{aligned}$$

Above we made use of the fact that we can exchange the order of the summations, always under the assumption that the expectations actually exist. We now make use of Theorem 9.3 and of property (9.1) to perform some of the sums above and then use the definition of expectation, Def. 7.1. This gives

$$\begin{aligned} E[rX+sY+t] &= r\sum_{x\in X(\Omega)} x\,p_X(x) + s\sum_{y\in Y(\Omega)} y\,p_Y(y) + t \\ &= rE[X] + sE[Y] + t. \end{aligned}$$

Next we give the proof in the case where X and Y are jointly continuous. We skipped this in the lecture and so can you if you like.

$$\begin{aligned} E[rX+sY+t] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(rx+sy+t)f_{XY}(x,y)dxdy \text{ by Thm. 10.1} \\ &= r\int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{XY}(x,y)dy\right)dx + s\int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f_{XY}(x,y)dx\right)dy \\ &\quad + t\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy. \end{aligned}$$

Above we made use of the fact that we can exchange the order of the integrations, always under the assumption that the expectations actually exist. We now make use of Theorem 9.9 and of property (jd2) to perform some of the integrals above and then use the definition of expectation, Def. 7.8. This gives

$$\begin{aligned} E[rX+sY+t] &= r\int_{-\infty}^{\infty} x f_X(x)dx + s\int_{-\infty}^{\infty} y f_Y(y)dy + t \\ &= rE[X] + sE[Y] + t. \end{aligned}$$

Note how the proof in the continuous case is almost identical to the proof in the discrete case, just with sums replaced by integrals and probability mass functions replaced by probability density functions. We will not prove other cases where X and Y are neither both discrete nor jointly continuous because that would require a generalised notation

that we do not introduce in this module. □

Example 10.3. Let $Y_1, \ldots, Y_n$ be a sequence of independent Bernoulli trials, each with probability of success p, i.e., $Y_i \sim \text{Ber}(p)$. Then $X = \sum_{k=1}^{n} Y_k$ is equal to the total number of successes in n trials. As we discussed in Example 4.12, $X \sim \text{Bin}(n, p)$. Theorem 10.2 allows us to calculate $E[X]$ as

$$E[X] = E\left[\sum_{k=1}^{n} Y_k\right] = \sum_{k=1}^{n} E[Y_k] = \sum_{k=1}^{n} p = np\,.$$

We used that the expectation of a $\text{Ber}(p)$ random variable is p.

10.2 Covariance

Linearity does not hold for variance:

$$\begin{aligned}
\text{Var}[X+Y] &= E\left[(X+Y-E[X+Y])^2\right] \text{ by Def. 7.18} \\
&= E\left[(X-E[X]+Y-E[Y])^2\right] \text{ by Theorem 10.2} \\
&= E\left[(X-E[X])^2 + (Y-E[Y])^2 + 2(X-E[X])(Y-E[Y])\right] \\
&= E\left[(X-E[X])^2\right] + E\left[(Y-E[Y])^2\right] + 2E\left[(X-E[X])(Y-E[Y])\right] \\
&= \text{Var}[X] + \text{Var}[Y] + 2E\left[(X-E[X])(Y-E[Y])\right].
\end{aligned}$$

We used Theorem 10.2 (linearity of the expectation) several times above. We now give a name to the third term above.

Definition 10.4. Let X and Y be random variables. The **covariance** between X and Y is defined as

$$\text{Cov}[X,Y] = E\left[(X-E[X])(Y-E[Y])\right].$$

If $\text{Cov}[X,Y] = 0$ we say that X and Y are **uncorrelated**, otherwise they are **correlated**.

We see that if the covariance is positive, larger X leads us to expect larger Y and vice versa.

We can slightly generalise our above observation for the variance of a sum:

Theorem 10.5. *Let X and Y be random variables and let $r, s, t \in \mathbb{R}$. Then*

$$\text{Var}[rX+sY+t] = r^2\text{Var}[X] + s^2\text{Var}[Y] + 2rs\text{Cov}[X,Y].$$

The proof is a simple exercise.

As was the case for the variance, also for the covariance there is an alternative expression for it that sometimes is easier to compute.

Theorem 10.6. *(Compare to Theorem 7.19) Let X and Y be random variables. Then*

$$\mathrm{Cov}\,[X,Y] = E[XY] - E[X]E[Y].$$

I leave the proof as an exercise.

Example 9.1 cont.: When the total number of heads N is larger, then we have a higher expectation that the second coin lands head. Thus we expect S and N to be positively correlated. We confirm that with a calculation:

$$\mathrm{Cov}\,[S,N] = E[SN] - E[S]E[N] = \frac{3}{4} - \frac{1}{2}\cdot 1 = \frac{1}{4}.$$

Example 10.7. Smarties come in 8 colours: Red, Green, Blue, Yellow, Orange, Brown, Violet and Pink. Let us denote the probability for a random smartie to be red by p_R and similarly for $p_G, p_B, p_Y, p_O, p_{Br}, p_V, p_P$. Consider a small box with n randomly drawn smarties. Let Y be the number of yellow smarties in the box. Then $Y \sim \mathrm{Bin}(n, p_Y)$. Similarly let B be the number of blue smarties in the box. Then $B \sim \mathrm{Bin}(n, p_B)$. Calculate $\mathrm{Cov}\,[Y,B]$.

Solution. According to Theorem 10.6 we can calculate the covariance as

$$\mathrm{Cov}\,[Y,B] = E[YB] - E[Y]E[B].$$

We have already calculated the expectation of binomially distributed random variables in Example 10.3, giving us

$$E[Y] = n\,p_Y, \quad E[B] = n\,p_B.$$

We still need to calculate $E[YB]$, for which we can use Theorem 10.1,

$$E[YB] = \sum_{y\in Y(\Omega)} \sum_{b\in B(\Omega)} y\,b\,p_{YB}(y,b).$$

For this we need the joint mass function

$$p_{YB}(y,b) = P(Y=y, B=b) = p_Y^y\, p_B^b (1-p_Y-p_B)^{n-y-b} \binom{n}{y+b}\binom{y+b}{b}.$$

The binomial factors count the number of ways to choose the yellow and blue smarties from all n smarties. Doing the doubly infinite sum to calculate $E[YB]$ in this way is a good exercise in arithmetic, but we will not do that here and instead follow an alternative method that is more enlightening and uses tricks that will be useful in other contexts. We will use the same method we used in Example 10.3 of writing a binomially distributed random variable as sum over indicator random variables. So in this case we introduce

the random variables

$$Y_i = \mathbb{1}_{i\text{-th smartie is yellow}} = \begin{cases} 1 & \text{if } i\text{-th smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

and

$$B_i = \mathbb{1}_{i\text{-th smartie is blue}} = \begin{cases} 1 & \text{if } i\text{-th smartie is blue} \\ 0 & \text{otherwise} \end{cases}$$

and then

$$Y = \sum_{i=1}^{n} Y_i, \quad B = \sum_{i=1}^{n} B_i.$$

Using this we have

$$\operatorname{Cov}[Y, B] = \operatorname{Cov}\left[\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} B_j\right].$$

It would be very convenient if we could write this covariance of sums as sums of covariances. This is indeed possible due to the following theorem:

Theorem 10.8. *If X, Y and Z are random variables and $r, s, t \in \mathbb{R}$ then*

$$\operatorname{Cov}[rX + sY + t, Z] = r \operatorname{Cov}[X, Z] + s \operatorname{Cov}[Y, Z]$$

Proof. The proof is by calculation, using the definition of covariance and the linearity of expectation, Theorem 10.2.

$$\begin{aligned} \operatorname{Cov}[rX + sY + t, Z] &= E\left[(rX + sY + t - E[rX + sY + t])\,(Z - E[Z])\right] \\ &= E\left[(r(X - E[X]) + s(Y - E[Y]))\,(Z - E[Z])\right] \\ &= rE\left[(X - E[X])\,(Z - E[Z])\right] + sE\left[(Y - E[Y])\,(Z - E[Z])\right] \\ &= r \operatorname{Cov}[X, Z] + s \operatorname{Cov}[Y, Z] \end{aligned}$$

□

Even though we wrote this theorem for the sum of two random variables, this can of course be used repeatedly to write the covariance of a sum of an arbitrary number of

terms as the sum of the covariances of individual terms. We find

$$\begin{aligned}
\operatorname{Cov}[Y, B] &= \operatorname{Cov}\left[\sum_{i=1}^{n} Y_i, B\right] \\
&= \sum_{i=1}^{n} \operatorname{Cov}[Y_i, B] \text{ by using Thm. 10.8 repeatedly} \\
&= \sum_{i=1}^{n} \operatorname{Cov}[B, Y_i] \text{ because the covariance is symmetric} \\
&= \sum_{i=1}^{n} \operatorname{Cov}\left[\sum_{j=1}^{n} B_j, Y_i\right] \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \operatorname{Cov}[B_j, Y_i] \text{ by using Thm. 10.8 repeatedly} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}[Y_i, B_j] \text{ because the covariance is symmetric.}
\end{aligned}$$

The covariance of two indicator random variables is really easy to calculate. We distinguish the case where both refer to the same smartie and the case where they refer to different smarties. In the first case we can use that a smartie can not be both yellow and blue at the same time, so that $Y_i B_i = 0$, giving

$$\operatorname{Cov}[Y_i, B_i] = E[Y_i B_i] - E[Y_i]E[B_i] = 0 - p_Y\, p_B.$$

For the second case we can use that one smartie being yellow is independent of another smartie being blue, so $Y_i \perp\!\!\!\perp B_j$, and thus we can use the following theorem:

Theorem 10.9. *If two random variables are independent, then their covariance is zero.*

Proof. First we give the proof for the case where the two random variables X and Y and $X \perp\!\!\!\perp Y$. We calculate

$$\begin{aligned}
E[XY] &= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} x\, y\, p_{XY}(x_k, y_l) \text{ by Theorem 10.1} \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} x\, y\, p_X(x_k) p_Y(y_l) \text{ by independence and Thm. 9.12} \\
&= \sum_{x \in X(\Omega)} x\, p_X(x_k) \sum_{y \in Y(\Omega)} y\, p_Y(y_l) \\
&= E[X]E[Y] \text{ by Definition 7.1.}
\end{aligned}$$

Therefore

$$\begin{aligned}\operatorname{Cov}[X,Y] &= E[XY] - E[X]E[Y] \text{ by Theorem 10.6}\\ &= E[X]E[Y] - E[X]E[Y] = 0.\end{aligned}$$

The proof for jointly continuous random variables X and Y is very similar and therefore we skipped it in the lecture:

$$\begin{aligned}E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)dxdy \text{ by Theorem 10.1}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x)f_Y(y)dxdy \text{ by independence and 9.12}\\ &= \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy\\ &= E[X]E[Y] \text{ by Definition 7.8.}\end{aligned}$$

Hence again $\operatorname{Cov}[X,Y] = E[XY] - E[X]E[Y] = 0$. □

Note that we have established that

$$X \perp\!\!\!\perp Y \implies E[XY] = E[X]E[Y]$$

but the implication does not go in the other direction. Factorisation of the expectation of the product into the product of expectations is not enough to imply independence.

We can now complete our calculation of the covariance of the number of yellow and the number of blue smarties:

$$\begin{aligned}\operatorname{Cov}[Y,B] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}[Y_i,B_j]\\ &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\operatorname{Cov}[Y_i,B_j] + \sum_{i=1}^{n}\operatorname{Cov}[Y_i,B_i]\\ &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} 0 + \sum_{i=1}^{n} -p_y\, p_b\\ &= -n\, p_Y\, p_B.\end{aligned}$$

Note how we split up the sum over all pairs of indices i, j into those where $i \neq j$ and those where $i = j$.

10.3 The correlation coefficient

The covariance is not a perfect measure of the strength of correlation between two random variables because it depends on the choice of units for the random variables. One can however combine the covariance between X and Y with the variances of X and Y in such a way to cancel that dependence.

Definition 10.10. Let X and Y be random variables. The **correlation coefficient** $\rho(X, Y)$ is defined as

$$\rho(X,Y) = \begin{cases} \dfrac{\operatorname{Cov}[X,Y]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}} & \text{if } \operatorname{Var}[X]\operatorname{Var}[Y] > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The next theorem summarises why the correlation coefficient is convenient as a measure of correlation between two random variables: it does not change as you rescale and it always lies between -1 and 1.

Theorem 10.11. *Let X and Y be random variables and let $r, s, t, u \in \mathbb{R}$. Then*

1.

$$\rho(rX+s, tY+u) = \begin{cases} \rho(X,Y) & \text{if } rt > 0 \\ 0 & \text{if } rt = 0 \\ -\rho(X,Y) & \text{if } rt < 0 \end{cases}$$

2.

$$-1 \leq \rho(X,Y) \leq 1.$$

Example 10.12. Let us calculate the correlation coefficient for the number of yellow and the number of blue smarties in a box of n smarties. For that we need, besides the covariance that we have already calculated, the variances. To calculate these we again use the trick of writing the variables as sums over indicator random variables and thus find

$$\begin{aligned} \operatorname{Var}[Y] &= \operatorname{Var}\left[\sum_{i=1}^{n} Y_i\right] \\ &= \sum_{i=1}^{n} \operatorname{Var}[Y_i] \text{ by independence of the } Y_i \\ &= \sum_{i=1}^{n} p_Y(1-p_Y) \text{ see Example 7.21} \\ &= n\, p_Y(1-p_Y). \end{aligned}$$

Similarly, $\mathrm{Var}\,[B] = n\,p_B(1-p_B)$. Putting these results in the definition of the correlation coefficient gives

$$\rho(Y,B) = \frac{\mathrm{Cov}\,[Y,B]}{\sqrt{\mathrm{Var}\,[Y]\,\mathrm{Var}\,[B]}} = \frac{-n\,p_Y\,p_B}{\sqrt{n\,p_Y(1-p_Y)\,n\,p_B(1-p_B)}} = -\sqrt{\frac{p_Y\,p_B}{(1-p_Y)(1-p_B)}}.$$

13 The law of large numbers

(Textbook chapter link)

13.1 Averages vary less

Consider a probability experiment and a random variable X. As an example we take the throwing of a die with X being the number on the top of the die. Then consider n independent repetitions of the experiment and thus n random variables $X_1, X_2, \ldots X_n$. In our example the random variable X_i then represents the number that comes up in the i-th throw of the die. We assume that there is no change in experimental conditions between the repetitions of the experiment (for example we do not swap the die for an unfair die) and therefore all the X_i have the same distribution. Furthermore, the outcome of one throw does not in any way influence the outcome of other throws and therefore the X_i are all independent.

This situation is so common that there is a name for it:

Definition 13.1. Let X be a random variable. A collection $X_1, \ldots, X_n$ of independent random variables that all have the same distribution as X is called a **i.i.d. sample** from the distribution of X of size n. (i.i.d. stands for independently and identically distributed.) The average $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$ is called the **sample mean**.

Our textbook refers to an i.i.d. sample as a 'random sample', and you may catch me do so as well from time to time. But generally I prefer the term 'i.i.d. sample' because it is more informative.

In Figure 13.1 the values of the first 30 throws of a die are shown in black. The values are connected by black lines just to graphically emphasize how much the values jump around. In red we show the the cumulative averages of the values. These behave much more predictably. They appear to converge towards the expectation $E[X] = 3.5$, which we have indicated with a horizontal dotted line. We will now set out to prove that this must be the case.

The average of the first n measurements is the value of the sample mean $\bar{X}_n$. The expectation of this sample mean is

$$E[\bar{X}_n] = \left(E[X_1] + E[X_2] + \cdots + E[X_n]\right)/n = nE[X]/n = E[X],$$

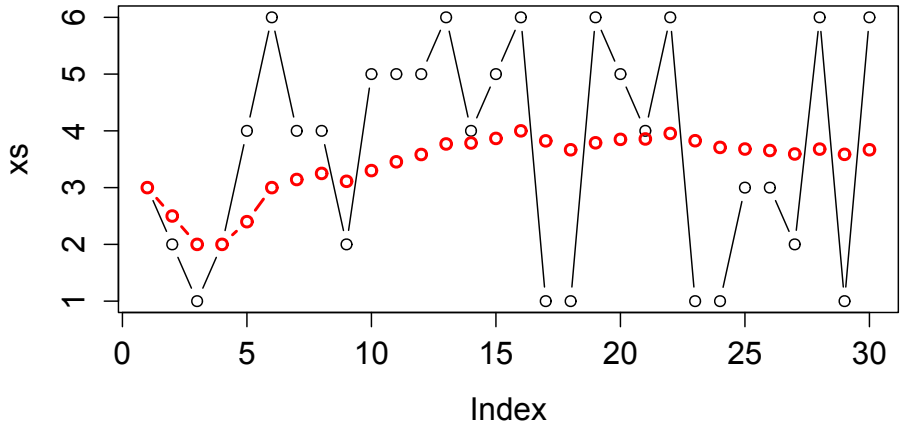


Figure 13.1: In black: the values of $X_1, \ldots, X_{30}$ from repeated throwing of a die. In red: the values of the averages $\bar{X}_1, \ldots \bar{X}_{30}$. The horizontal dotted line indicates the expected value of X.

where we used the linearity of expectation, Theorem 10.2. The variance is

$$\begin{aligned}\mathrm{Var}\left[\bar{X}_n\right] &= \mathrm{Var}\left[X_1/n\right] + \mathrm{Var}\left[X_2/n\right] + \cdots + \mathrm{Var}\left[X_n/n\right] \\ &= \frac{1}{n^2}\left(\mathrm{Var}\left[X_1\right] + \mathrm{Var}\left[X_2\right] + \cdots + \mathrm{Var}\left[X_n\right]\right) \\ &= \frac{1}{n^2} n \mathrm{Var}\left[X\right] = \frac{\mathrm{Var}\left[X\right]}{n}.\end{aligned}$$

For the first equality above we used the independence of the X_i, for the second we used the transformation property of the variance, Theorem 7.25.

13.2 Chebychev's inequality

The previous theorem shows that the variance of the sample mean goes down as $1/n$. Given our intuitive understanding of the variance as a measure for the likelihood that the random variable deviates from its mean, we can now understand why it is that the cummulative averages in Figure 13.1 tend towards the mean. However we still need to provide a formal basis for that intuition. This is provided by Chebychev's inequality.

Theorem 13.2. *Let X be a random variable and let $a \in \mathbb{R}$ with $a > 0$. Then*

$$P(|X - E[X]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}\left[X\right]. \tag{13.1}$$

Proof. We give a proof for the case where X is a continuous random variable. We write

$E[X] = \mu$. Then

$$\begin{aligned}\operatorname{Var}[X] = E[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx \quad \text{by Thm. 7.11} \\ &\geq \int_{|x-\mu]\geq a} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{|x-\mu]\geq a} a^2 f_X(x)dx = a^2 P(|X-\mu| \geq a).\end{aligned}$$

The case of a discrete random variable can be done in the same manner, but using the mass function instead of the density, and this is one of the homework questions. □

An alternative proof via Markov's inequality is given in the textbook by Ross in section 8.2.

There is an alternative formulation of Chebychev's inequality that is often useful. It gives a lower bound for the probability that X takes a value less than a certain number k of standard deviations away from its expectation. We just have to set $a = k \operatorname{sd}[X]$ in Theorem 13.2 to obtain:

Corollary 13.3. *Let X be a random variable with finite expectation $E[X] = \mu$ and finite variance σ^2 and let $k \in \mathbb{R}$ with $k > 0$. Then*

$$P\left(|X - E[X]| \geq k \operatorname{sd}[X]\right) \leq \frac{1}{k^2} \tag{13.2}$$

and thus

$$P(|X - E[X]| < k \operatorname{sd}[X]) \geq 1 - \frac{1}{k^2}. \tag{13.3}$$

Example 13.4. Assume the probability for a smartie to be yellow is $p_Y = 1/8$. As in Example 10.7 let Y be the number of yellow smarties in a box of n smarties. Let $n = 40$. You would then expect $E[Y] = np_Y = 40/8 = 5$ yellow smarties. Use Chebychev to get an upper bound on the probability to get 11 or more yellow smarties.

Solution. Because $E[Y] = 5$, we can write the event $\{Y \geq 11\}$ equivalently as $\{|Y - E[Y]| \geq 6\}$, which makes it easy to apply Chebychev's inequality. Using that $\operatorname{Var}[Y] = 35/8$ we get

$$P(Y \geq 11) = P(|Y - E[Y]| \geq 6) \leq \frac{1}{6^2}\operatorname{Var}[(]Y) = \frac{1}{36}\frac{35}{8} \approx 0.12.$$

So the probability of getting 11 yellow smarties or more is no more than about 12%. However in this case we actually know the distribution of Y, so we can calculate that probability. Because $Y \sim \operatorname{Bin}(40, 1/8)$ we get

$$P(Y \geq 11) = 1 - F_Y(10) \approx 0.008.$$

This shows that the upper bound from Chebychev's inequality is not very good.

13.3 The law of large numbers

In the lectures I talked a bit about the meaning and the importance of the law of large numbers. In these notes I will only state it and its proof.

Theorem 13.5. *For any $n \in \mathbb{N}$, let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a distribution with finite expectation μ and finite variance σ^2. Then*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

for any $\epsilon > 0$ (the weak law of large numbers, convergence in probability) and also

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1,$$

(the strong law of large numbers, convergence almost surely).

Proof of the weak law. We have that

$$E[\bar{X}_n] = E[X_i] = \mu$$

and

$$\mathrm{Var}\left[\bar{X}_n\right] = \mathrm{Var}\left[X_i\right]/n = \sigma^2/n.$$

From Chebychev's inequality (Theorem 13.2) we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathrm{Var}\left[\bar{X}_n\right] = \frac{\sigma^2}{n\epsilon^2}.$$

Now take the limit $n \to \infty$ on both sides:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

But we also know that probabilities are always non-negative, and the limit of a sequence of non-negative numbers is also non-negative, so the limit must be zero.

Note that in this proof we did not use that the X_i are identically distributed, just that they all have the same expectation and variances. Even that can be relaxed further, see problem sheet. The proof of the strong law, that we did not present here, does require the X_i to be identically distributed. □

13.4 Consequences of the law of large numbers

Now we will discuss how we can also estimate the probability of any event A by performing independent repetitions of the probability experiment. The intuitive idea is that the probability of the event could be approximated by the relative frequency with which the event occurs in the sample. To formalise this intuition, we are going to use indicator random variables for the event A, i.e., the random variable $\mathbb{1}_A$ defined by

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

To understand the utility of the indicator random variable in this context we calculate its expectation:

$$E[\mathbb{1}_A] = 1 \cdot P(\mathbb{1}_A = 1) + 0 \cdot P(\mathbb{1}_A = 0) = P(\mathbb{1}_A = 1) = P(A)$$

We see that the probability of an event can be expressed in terms of the expectation of its indicator random variable. We already know from the law of large numbers how to estimate expectations from an i.i.d. sample, so this will now also allow us to estimate probabilities of events from an i.i.d. sample.

To estimate the probability $P(A)$ we take an i.i.d. sample $X_1, X_2, \ldots, X_n$ from $X = \mathbb{1}_A$. The sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ is equal to the proportion of the first n repetitions of the probability experiment in which A occurs. Also

$$E[\bar{X}_n] = E[X_i] = P(\mathbb{1}_A = 1) = P(A).$$

From the law of large numbers

$$\lim_{n\to\infty} \bar{X}_n = E[\bar{X}_n] = P(A)$$

almost surely.

Given that we can estimate probabilities of any event, we can also estimate the probability distribution function F_X of X because $F_X(x)$ is just the probability of the event $\{X \leq x\}$. Thus $F_x(x)$ will be approximately equal to $1/n$ times the number of X_i that are less or equal to x.

We can furthermore estimate the probability density with a histogram. You have seen this in the practicals. We approximate the probability density at a point x using the number of sample values that lie in a small interval $[x - h, x + h]$ around that point,

for h small.

$$\begin{aligned} f_X(x) &\approx \frac{1}{2h} P(X \in [x-h, x+h]) \\ &\approx \frac{1}{2h} \cdot \frac{1}{n} \cdot \text{number of } X_i \text{ that lie in } [x-h, x+h]. \end{aligned}$$

A histogram shows bars for many such small intervals, giving an approximation to f_X.

14 Central limit theorem

(Textbook chapter link)

The following property of the normal distribution is the reason why it plays its role in the central limit theorem:

Theorem 14.1. *Let* $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ *be two independent random variables with* $\mu_x, \sigma_X, \mu_Y, \sigma_Y \in \mathbb{R}$*. Then*

$$X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Note that we already know why the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances. Why the distribution is normal will be easier to show next year after you have been introduced to generating functions, so we will skip the proof for now.

Now consider an i.i.d. sample $X_1, \ldots, X_n$ from a random variable X with finite expectation and variance. We already know that the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ has expectation and variance given by

$$E[\bar{X}_n] = E[X], \quad \mathrm{Var}\left[\bar{X}_n\right] = \frac{\mathrm{Var}\,[X]}{n}.$$

You have seen in computer demonstration at the start of the lecture and in the R lab, that for large sample size the sample mean is approximately normally distributed, even if the distribution of X is not normal at all. Let us introduce the symbol $\dot{\sim}$ to express that a random variable approximately follows a given distribution. Then we can formulate the following rule of thumb

Rule of thumb 14.2. For sufficiently large sample size n the sample mean is approximately normally distributed,

$$\bar{X}_n = (X_1 + \cdots + X_n)/n \dot{\sim} \mathrm{N}(E[X], \mathrm{Var}\,[X]/n).$$

Equivalently, multiplying the sample mean by n gives

$$X_1 + \cdots + X_n \dot{\sim} \mathrm{N}(nE[X], n\mathrm{Var}\,[X]).$$

The amazing fact is that, no matter what the distribution of the random variable X is, for sufficiently large sample size n the sample mean is approximately normally distributed.

The above observation is not in the form of a theorem, because it is rather vague about how large n has to be and exactly how well the normal distribution describes the

distribution of $\bar{X}_n$. To get a precise statement that deserves the name Theorem, we have to take the limit of $n \to \infty$. However X_n does not have a nice limit as $n \to \infty$ because its variance goes towards zero. Therefore we consider instead the standardized random variable

$$Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{\mathrm{Var}\left[\bar{X}_n\right]}} = \frac{\bar{X}_n - E[X_i]}{\sqrt{\mathrm{Var}\left[X_i\right]/n}},$$

which for all n has zero mean and variance one. Now we can formulate the central limit theorem:

Theorem 14.3. *(Central limit theorem) For any $n \in \mathbb{N}$ let $X_1, X_2, \ldots$ be an i.i.d. sample from a distribution with finite expectation μ and finite variance σ^2 and let*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then at any point $x \in \mathbb{R}$

$$\lim_{n\to\infty} F_{Z_n}(x) = \Phi(x),$$

(convergence in distribution) where Φ is the distribution function of the standard normal distribution $\mathrm{N}(0,1)$.

As an aside notice that we have now met three different ways in which a sequence of random variables can converge: 'convergence in distribution',' convergence in probability' and 'convergence almost surely'. These are successively stronger in that 'convergence almost surely' implies 'convergence in probability' which implies 'convergence in distribution'. Perhaps this gives you a glimpse of how much mathematical richness is hiding in probability theory that we unfortunately have to brush over in this introductory module but that I hope you will be able to appreciate later in your degree.

17 Statistical models

(Textbook chapter link)

We now start to discuss how to learn something about the real world by using data. This process involves the building of a statistical model of the real world and then estimating the interesting parameters in that model. In this chapter we introduce the concept of a statistical model and illustrate it with a number of examples of models. In the next chapter we then start discussing the estimation of the model parameters.

The central method underlying our scientific understanding of the world is to build a mathematical model of the phenomena one wants to describe.

Example 17.1. [Galileo] Consider a ball rolling on an inclined plane. Let x be the variable giving the distance travelled by the ball and t the variable giving the time elapsed since the ball started rolling. The relationship between these two variables is given by the mathematical equation

$$x = \frac{1}{2}at^2.$$

The parameter a appearing in this relationship is called the acceleration. To determine the value of the parameter we need to perform the experiment. We let the ball roll and stop it after t seconds and measure the distance x travelled in meters. Then $a = 2x/t^2$. This is trivial but there is a hitch: each time the experiment is repeated one gets a slightly different value for the parameter a, due to random effects leading to measurement errors both in the time measurement and in the distance measurement. By repeating the experiment n times one obtains a dataset $x_1, x_2 \ldots, x_n$. How do we determine the correct value of the parameter a from these experiments. To answer this question we need to extend the simple deterministic relationship between x and t to include the uncertainty arising from the random effects. Such a model is called a *statistical model*. The observations $x_1, x_2, \ldots x_n$ are modelled as the values of a random variables $X_1, X_2, \ldots X_n$. A possible model for Galileo's experiment would be

$$X_i = \frac{1}{2}a\left(t + U_i\right)^2 + V_i.$$

The random variables U_i model the random error in the time measurement and the random variables V_i describe the random error in the distance measurement. We assume that these errors are normally distributed:

$$U_i \sim \mathrm{N}(0, \sigma_U^2), \quad V_i \sim \mathrm{N}(0, \sigma_V^2).$$

Because the time measurement is independent of the distance measurement, and because the repetitions of the experiment are independent, we assume that the U's and the V's

are independent. With such a concrete model in place, it is now possible to estimate the parameter a from the observed data.

You have seen other statistical models already. In Example 10.7 you saw how we modelled the observations of the numbers of smarties of different colours in a box of smarties.

Let us abstract from these examples what we generally mean by a statistical model:

A (parametric) **statistical model** (also referred to as a probabilistic model or stochastic model) for a numerical dataset:

1. views the data as the values of a set of random variables;

2. gives a partial specification of the joint probability distribution for these random variables (and possibly additional unobserved random variables), including in particular information on independence/dependence among them. This distribution is often referred to as the **model distribution**.

3. contains a set of parameters of the distribution that are unknown but of interest, to be estimated from the dataset. These are the **model parameters**.

Statistical inference is the process of inferring the value of the model parameters from the observation. The remainder of these notes and of your future statistics modules is about this process.

In summary: the way a mathematical statistician learns information from data is to make a statistical model for the data and to use the data to deduce the value of the unknown parameters in that model.

We will now proceed to give further of examples of statistical models to get some feel for the wide range of possibilities. You are not expected to memorise these examples, they are for illustrative purposes only.

Example 17.2. A statistical model for the outcomes $x_1, \ldots, x_n \in \{0, 1\}$ of n coin flips, where 0 encodes Tail and 1 encodes Head:

1. Random variables $X_1, \ldots, X_n$.

2. The X_i are an i.i.d. sample from the Bernoulli distribution, i.e., they are independently and identically distributed (i.i.d) with $X_i \sim \text{Ber}(p),\ i = 1, \ldots, n$.

3. The model parameter is p, the probability of the coin landing heads up.

The above two examples modelled the data as realisations of i.i.d. samples, i.e., all the random variables were taken to be independent and identically distributed. In the next example that kind of independence clearly does not hold.

Example 17.3. A statistical model for the daily share price of a company. Let x_i be the closing share price on day i. The dataset $x_1, \ldots, x_n$ is called a time-series. We model this with

1. Random variables $X_1, \ldots, X_n$ for the share price on different days and $U_1, \ldots, U_n$ for random fluctuations on each day.

2. The share price on one day is of course not independent of that on the previous day. So the X_i are certainly not independent. There are many important statistical models for such time series. The simplest is the AR(1) model

$$X_{i+1} = \alpha X_i + U_i, \quad \text{with } U_i \sim \mathrm{N}(0, \sigma^2).$$

3. The model parameters are α, which gives the mean rate of return of the stock, and σ^2, which describes the volatility of the share price.

19 Unbiased estimators

(Textbook chapter link)

We now turn to the problem of estimating the model parameters from the data.

Let us now abstract from this example to general definitions of the terms we introduced in this example.

Definition 19.1. Let $(x_1, x_2, \ldots, x_n)$ be a dataset modelled by random variables $X_1, X_2, \ldots, X_n$ and let $t = h(x_1, \ldots, x_n)$ be an **estimate** for the value of a model parameter θ, expressed as a function h evaluated on the dataset values $x_1, \ldots, x_n$. Then the random variable $T = h(X_1, \ldots, X_n)$ is called an **estimator**. Such an estimator is **unbiased** if $E[T] = \theta$, irrespective of the value of θ, otherwise it is **biased**. The difference $E[T] - \theta$ is called the **bias** of T.

Theorem 19.2. *Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution with expectation $\mu < \infty$ and variance $\sigma^2 < \infty$. Then the sample mean*

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

is an unbiased estimator for μ and the sample variance

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

is an unbiased estimator for σ^2.

Note that the normalisation factor in the expression for the sample variance is $1/(n-1)$ rather than the $1/n$ one might expect for a variance.

Proof. That $E[\bar{X}_n] = \mu$ we already know from our calculation at the start of chapter 13. So we only need to show that $E[S_n^2] = \sigma^2$. First we use linearity of expectation, giving

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

Then we observe that, using Theorem 7.19,

$$E[(X_i - \bar{X}_n)^2] = \mathrm{Var}\left[X_i - \bar{X}_n\right] + \left(E[X_i - \bar{X}_n]\right)^2 = \mathrm{Var}\left[X_i - \bar{X}_n\right].$$

For the last equality we used that $E[X_i - \bar{X}_n] = E[X_i] - E[\bar{X}_n] = \mu - \mu = 0$. To calculate $\mathrm{Var}\left[X_i - \bar{X}_n\right]$ we use the trick of writing

$$X_i - \bar{X}_n = X_i - \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i} X_j.$$

We can then use the fact that X_j is independent of X_i for $j \neq i$ to write

$$\begin{aligned}
\mathrm{Var}\left[X_i - \bar{X}_n\right] &= \mathrm{Var}\left[\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i} X_j\right] \\
&= \mathrm{Var}\left[\frac{n-1}{n}X_i\right] + \sum_{j\neq i}\mathrm{Var}\left[-\frac{1}{n}X_j\right] \\
&= \frac{(n-1)^2}{n^2}\mathrm{Var}\left[X_i\right] + \frac{1}{n^2}\sum_{j\neq i}\mathrm{Var}\left[X_j\right] \\
&= \frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}(n-1)\sigma^2 \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

For the third equality we used the transformation property of the variance, Theorem 7.25. Putting this all together we find

$$E[S_n^2] = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{Var}\left[X_i - \bar{X}_n\right] = \frac{1}{n-1}n\frac{n-1}{n}\sigma^2 = \sigma^2,$$

as required. □

Example 17.1 cont.: An estimator for the acceleration in the inclined plane experiment. Galileo has given us a functional relationship between the distance travelled x and the time of travel t: $x = \frac{1}{2}\,a\,t^2$. We can solve this for the acceleration: $a = 2x/t^2$. However each time the experiment is repeated one will get a different result x_i due to random errors. We modelled these observations by random variables

$$X_i = \frac{1}{2}a(t + U_i)^2 + V_i$$

where $U_i \sim \mathrm{N}(0, \sigma_U^2)$, $V_i \sim \mathrm{N}(0, \sigma_V^2)$ and these errors are all independent. We could try to estimate a by taking an average of the observations $x_1, \ldots, x_n$:

$$a \approx \frac{2}{t^2}\frac{1}{n}\sum_{i=1}^{n} x_i.$$

The corresponding estimator is

$$A = \frac{2}{t^2}\bar{X}_n.$$

To check whether this estimator is unbiased we calculate

$$E[A] = E\left[\frac{2}{t^2}\bar{X}_n\right] = \frac{2}{t^2}E[\bar{X}_n] = \frac{2}{t^2}E[X_i].$$

So we need the expectation of X_i:

$$\begin{aligned} E[X_i] &= E\left[\frac{1}{2}a(t+U_i)^2 + V_i\right] = \frac{1}{2}aE[(t+U_i)^2] + E[V_i] \\ &= \frac{1}{2}a\left(\mathrm{Var}\left[t+U_i\right] + (E[t+U_i])^2\right) = \frac{1}{2}a\left(\mathrm{Var}\left[U_i\right] + t^2\right) \\ &= \frac{1}{2}a\left(\sigma_U^2 + t^2\right) \end{aligned}$$

This gives

$$E[A] = \frac{2}{t^2}E[X_i] = a\frac{\sigma_U^2 + t^2}{t^2} \neq a.$$

So this estimator is not unbiased. Taking the average is going to consistently overestimate the value of a. Luckily we can fix this by rescaling the estimator. The estimator

$$\tilde{A} = \frac{t^2}{\sigma_U^2 + t^2}A = \frac{2}{\sigma_U^2 + t^2}\bar{X}_n$$

is unbiased.

Example 19.3. (R.A. Fisher 1925) This is an example on which we worked together in the lecture.

Leaves of maize plants can be divided into four different types: 1: starchy-green, 2: starchy-white, 3: sugary-green and 4: sugary-white. In an experiment in which $n = 3839$ plants were grown, the number of plants of each type were recorded $n_1 = 1997, n_2 = 906, n_3 = 904, n_4 = 32$. These four numbers constitute our dataset.

We model this dataset with random variables N_1, N_2, N_3, N_4. According to genetic theory, the types occur with probabilities $p_1 = (\theta+2)/4, p_2 = p_3 = (1-\theta)/4$, and $p_4 = \theta/4$, respectively, where $0 < \theta < 1$. This implies that the number of plants N_i of type i is binomially distributed with parameter p_i, $N_i \sim \mathrm{Bin}(n, p_i)$. However in this example the random variables are not independent. Instead, their joint distribution is the multinomial distribution,

$$(N_1, N_2, N_3, N_4) \sim \mathrm{Mult}(n, p_1, p_2, p_3, p_4).$$

The joint probability mass function is

$$\begin{aligned} p_{N_1,N_2,N_3,N_4}(n_1,n_2,n_3,n_4) &= P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) \\ &= \frac{n!}{n_1!n_2!n_3!n_4!}p_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4}. \end{aligned}$$

The model parameter is θ.

Giving just the above information, I asked you to come up with estimators for θ. The

first suggestion I received from the audience was to use an estimator

$$T_4 = \frac{4}{n} N_4.$$

Together we checked that this is indeed an unbiased estimator for θ:

$$E[T_4] = E\left[\frac{4}{n} N_4\right] = \frac{4}{n} E[N_4] = \frac{4}{n}\left(n\frac{\theta}{4}\right) = \theta.$$

On our dataset this estimator leades to an estimate for θ of

$$\theta \approx t_4 = \frac{4}{n} n_4 = \frac{4}{3839} 32 \approx 0.033.$$

The next suggestion was to use an estimator

$$T_1 = \frac{4}{n} N_1 - 2.$$

Again we checked together that it was unbiased:

$$E[T_1] = E\left[\frac{4}{n} N_1 - 2\right] = \frac{4}{n} E[N_1] - 2 = \frac{4}{n}\left(n\frac{1}{4}(\theta + 2)\right) - 2 = \theta.$$

On our dataset this estimator leads to an estimate for θ of

$$\theta \approx t_1 = \frac{4}{n} n_1 - 2 = \frac{4}{3839} 1997 - 2 \approx 0.081.$$

The values predicted by T_1 and T_4 are different. Which one should we believe more? To decide this, we should consider which estimator we should expect to have a smaller error. This question motivated the following definition:

Definition 19.4. Let T be an estimator for a parameter θ. The **mean squared error** of T is the number

$$\mathrm{MSE}(T) = E\left[(T - \theta)^2\right].$$

Note that if the estimator T is unbiased, then the means square error of T is equal to the variance of T:

$$\mathrm{MSE}(T) = E\left[(T - E[T])^2\right] = \mathrm{Var}\left[T\right].$$

Example 19.3 cont. We calculate the variances of the two estimators T_4 and T_1:

$$\mathrm{Var}\left[T_4\right] = \mathrm{Var}\left[\frac{4}{n} N_4\right] = \frac{16}{n^2} \mathrm{Var}\left[N_4\right] = \frac{16}{n^2} n p_4 (1 - p_4) = \frac{16}{n}\frac{\theta}{4}\left(1 - \frac{\theta}{4}\right) = \frac{1}{n}\theta(4 - \theta).$$

$$\mathrm{Var}\left[T_1\right] = \mathrm{Var}\left[\frac{4}{n}N_1 - 2\right] = \frac{16}{n^2}np_1(1-p_1) = \frac{16}{n}\frac{\theta+2}{4}\left(1 - \frac{\theta+2}{4}\right) = \frac{1}{n}(\theta+2)(2-\theta).$$

So the variance in each case is a quadratic function of θ. We do not know a priori what the value of θ is, other than that it lies between 0 and 1. By plotting the variances against θ (do it, sketching parabolas is easy once you know their zeros) we see that in this interval, $\mathrm{Var}\left[T_4\right] \leq \mathrm{Var}\left[T_1\right]$. Hence T_4 is the better estimator because it has the smaller mean squared error.

Next let's look at a more complicated estimator $T_{14} = (T_1 + T_4)/2$. This is clearly an unbiased estimator because

$$E[T_{14}] = E[(T_1 + T_4)/2] = (E[T_1] + E[T_4])/2 = (\theta + \theta)/2 = \theta.$$

To get its mean square error we have to calculate the variance,

$$\begin{aligned}\mathrm{Var}\left[T_{14}\right] &= \mathrm{Var}\left[\frac{1}{2}(T_1 + T_4)\right] = \frac{1}{4}\mathrm{Var}\left[T_1 + T_4\right] \\ &= \frac{1}{4}\left(\mathrm{Var}\left[T_1\right] + \mathrm{Var}\left[T_4\right] + 2\mathrm{Cov}\left[T_1, T_4\right]\right).\end{aligned}$$

To calculate the covariance $\mathrm{Cov}\left[T_1, T_4\right]$ we need to use the joint distribution of N_1 and N_4. Doing the calculation directly from the joint probability mass function will be tedious. So we again use our trick of using indicator random variables. We introduce indicator random variables Y_{ai} so that

$$Y_{ai} = \begin{cases} 1 & \text{if the } i\text{-th leaf is of type } a \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$N_a = \sum_{i=1}^{n} Y_{ai}.$$

For each leaf type a the indicator random variables Y_{ai} form an i.i.d. sample from a Bernoulli distribution $\mathrm{Ber}(p_a)$. Leaves from different plants are independent and hence Y_{ai} is independent from Y_{bj} for all a and b if $i \neq j$. Each specific plant i can only have a single leaf type. Hence if $Y_{ai} = 1$ for some type a then $Y_{bi} = 0$ whenever $b \neq a$. So $Y_{ai}Y_{bi} = 0$ if $a \neq b$.

With this more detailed model with its specification of the dependence and indepen-

dence among its variables we can now calculate the covariance. We begin by calculating

$$E[N_1N_4] = E\left[\sum_{i=1}^{n} Y_{1i} \sum_{j=1}^{n} Y_{4j}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_{1i}Y_{4j}]$$
$$= \sum_{i=1}^{n} E[Y_{1i}Y_{4i}] + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} E[Y_{1i}Y_{4j}],$$

Here we first used the linearity of the expectation and then we just split up the double sum into those terms where both variables refer to the same plant and those where they refer to different plants. In the first sum we can now use that the i-th leave can not at the same time be of type 1 and of type 4 and hence $Y_{1i}Y_{4i} = 0$. In the second sum we can use the independence of the outcomes for the different plants, so that the expectation factorises, giving us

$$E[N_1N_4] = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} E[Y_{1i}]E[Y_{4j}] = \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} p_1p_4 = n(n-1)p_1p_4.$$

We already know that $N_1 \sim \text{Bin}(n, p_1)$ and thus $E[N_1] = np_1$, but this is also easy to calculate:

$$E[N_1] = E\left[\sum_{i=1}^{n} Y_{1i}\right] = \sum_{i=1}^{n} E[Y_{1i}] = np_1,$$

and similarly $E[N_4] = np_4$. This allows us to calculate the covariance using Theorem 10.6:

$$\text{Cov}\,[N_1, N_4] = E[N_1N_4] - E[N_1]E[N_4] = n(n-1)p_1p_4 - np_1np_4 = -np_1p_4.$$

Using this we find that

$$\text{Cov}\,[T_1, T_4] = \text{Cov}\left[\frac{4}{n}N_1 - 2, \frac{4}{n}N_4\right] = \frac{16}{n^2}\text{Cov}\,[N_1, N_4]$$
$$= -\frac{16}{n}\frac{\theta+2}{4}\frac{\theta}{4} = -\frac{1}{n}(\theta+2)\theta.$$

The variance of the estimator T_{14} therefore is

$$\text{Var}\,[T_{14}] = \frac{1}{4n}\left((\theta+2)(2-\theta) + \theta(4-\theta) - 2(\theta+2)\theta\right) = \frac{1}{n}(1-\theta)(1+\theta).$$

By plotting this we see that this estimator has a lower mean square error for most θ but not for very small values. It is always better than the estimator T_1 however.

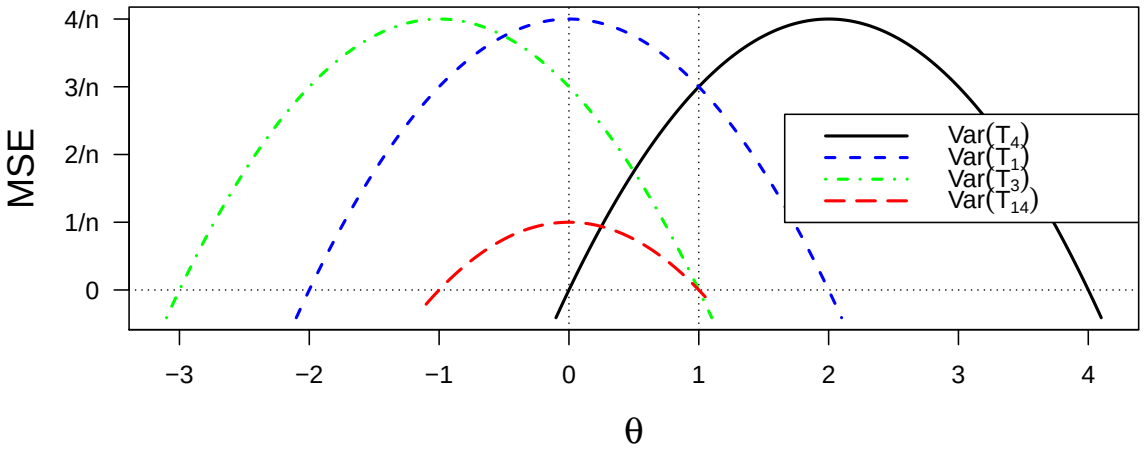


Figure 19.1: Plot of the mean square error (MSE) for the various estimators for the parameter θ in Example 19.3. We are only interested in the region for $\theta \in [0, 1]$, indicated by dotted lines.

Evaluated on our data, the estimator T_{14} gives us an estimate for θ of

$$\theta \approx t_{14} = (t_1 + t_4)/2 \approx 0.057.$$

Another possible unbiased estimator is $T_3 = 1 - \frac{4}{n}N_3$. You checked that it is indeed unbiased, i.e., that $E[T_3] = \theta$ and you calculated its variance:

$$\begin{aligned}\operatorname{Var}[T_3] = \operatorname{Var}\left[1 - \frac{4}{n}N_3\right] = \frac{16}{n^2}\operatorname{Var}[N_3] = \frac{16}{n^2}p_3(1-p_3)\\ = \frac{16}{n}\frac{1-\theta}{4}\left(1 - \frac{1-\theta}{4}\right) = \frac{1}{n}(1-\theta)(3+\theta).\end{aligned}$$

Plotting this parabola we see that if we do not know where in the interval $[0, 1]$ the true value of θ lies, then T_3 and T_4 look equally good. T_3 has a smaller variance for large θ but a large variance for small θ while for T_4 the situation is reversed. However we know from the estimates that θ is quite small and thus in the region where T_4 is the better estimator.

Evaluated on our data, the estimator T_3 gives us an estimate for θ of

$$\theta \approx t_3 = 1 - \frac{4}{n}n_3 = 1 - \frac{4}{3839}904 \approx 0.058.$$

You will investigate more examples of estimators in some of the problems on the problem sheet and will be able to use similar techniques to the one above to calculate their mean squared error. For the particular estimator T_{14} there happens to be an easier

way we could have used to calculate the MSE. We could have observed that

$$T_{14} = \frac{1}{2}(T_1 + T_4) = \frac{1}{2}\left(\frac{4}{n}N_1 - 2 + \frac{4}{n}N_4\right) = \frac{2}{n}(N_1 + N_4) - 1.$$

The random variable $N_1 + N_4$ is the number of plants that have either type 1 or type 4 and because the probability of an individual plant to have either type 1 or type 4 is $p_1 + p_4$ we know that

$$(N_1 + N_4) \sim \text{Bin}(n, p_1 + p_4)$$

and hence

$$\text{Var}[N_1 + N_4] = n(p_1 + p_4)(1 - (p_1 + p_4)).$$

We can thus calculate

$$\begin{aligned} MSE(T_{14}) = \text{Var}[T_{14}] = \text{Var}\left[\frac{2}{n}(N_1 + N_4)\right] = \frac{4}{n^2}\text{Var}[N_1 + N_4] \\ = \frac{4}{n}(p_1 + p_4)(1 - (p_1 + p_4)) = \frac{1}{n}(\theta + 1)(1 - \theta). \end{aligned}$$

21 Maximum likelihood

(Textbook chapter link)

The maximum likelihood principle gives us a rationale for how to estimate a value for a model parameter. The maximum likelihood principle can be stated prosaically as: "Given a dataset, choose the value of the model parameters in such a way that the data is most likely".

Example 21.1. You have a coin that has a probability p of landing heads, but you do not know this probability. You want to estimate it from observations. So you flip the coin 5 times and observe the sequence H H T H H.

Let us first assume that you already know that there are only two possible values for the probability p, namely $p = 3/4$ and $p = 2/3$. If you had to guess which of the two it is, what would you guess? You might start by calculating the probability of observing the data for each of the two possibilities:

$$P\left(\text{data} \,\middle|\, p = \frac{3}{4}\right) = \left(\frac{3}{4}\right)^4 \frac{1}{4} = \frac{81}{1024} \approx 0.079,$$

$$P\left(\text{data} \,\middle|\, p = \frac{2}{3}\right) = \left(\frac{2}{3}\right)^4 \frac{1}{3} = \frac{16}{243} \approx 0.066.$$

On the basis that the data is more likely to arise when $p = 3/4$ you might guess that, on the balance of probabilities, $p = 3/4$.

Next let us assume you have no information about p. Then you can still calculate

$$P(\text{data } |p) = p^4(1-p) = p^4 - p^5$$

and then there is one value of p that makes the probability of the observed data largest. The way to find this probability we use that the slope of a function is zero at its maximum. In this case we have

$$\frac{d}{dp}(p^4 - p^5) = p^3(4 - 5p) = 0$$

if $p = 4/5$. This value is the maximum likelihood estimate for p.

Definition 21.2. Let a dataset $x_1, x_2, \ldots x_n$ be modelled by random variables $X_1, X_2, \ldots X_n$ with a joint distribution with parameter θ. Then the **likelihood** $L(\theta)$ is the probability(density) of observing the data. If X is discrete,

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n) = p_{X_1,\ldots X_n}(x_1, \ldots, x_n).$$

If X is continuous,

$$L(\theta) = f_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

The **maximum likelihood estimate** for θ is a value $\hat{\theta}$ that maximises $L(\theta)$. If we write $\hat{\theta} = h(x_1,\dots,x_n)$ then

$$\hat{\Theta} = h(X_1,\dots,X_n)$$

is the **maximum likelihood estimator** for θ.

Example 21.3. A dataset $x_1,\dots,x_n$ is modelled as an i.i.d. sample $X_1,\dots X_n$ from an Exp(λ) distribution, i.e.,

$$f_{X_i}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

We are interested in infering the value of the parameter λ. The likelihood is given by

$$\begin{aligned} L(\lambda) &= f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n) \\ &= \lambda e^{-\lambda x_1}\cdots\lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}. \end{aligned}$$

Here we used that the variables $X_1,\dots,X_n$ are an i.i.d. sample and therefore are independent. Hence the joint density factorises into the product of the densities of the individual variables. The likelihood function $L(\lambda)$ is a continuous function and we can find its extrema by looking for the zeros of the derivative

$$\begin{aligned} \frac{d}{d\lambda}L(\lambda) &= n\lambda^{n-1}e^{-\lambda(x_1+\cdots+x_n)} - \lambda^n(x_1+\cdots+x_n)e^{-\lambda(x_1+\cdots+x_n)} \\ &= n\lambda^{n-1}e^{-\lambda(x_1+\cdots+x_n)}(1-\lambda(x_1+\cdots+x_n)/n) \\ &= n\lambda^{n-1}e^{-\lambda(x_1+\cdots+x_n)}(1-\lambda\bar{x}_n). \end{aligned}$$

We know that $L(\lambda)$ has an extremum at $\lambda = \hat{\lambda}$ only if

$$\frac{dL}{d\lambda}(\hat{\lambda}) = 0.$$

We are not interested in the extremum at $\hat{\lambda} = 0$. This implies that

$$\hat{\lambda} = \frac{1}{\bar{x}_n}.$$

To see that this extremum is a maximum rather than a minimum we can either make a qualitative sketch of the likelihood function or observe that

$$L(0) = 0, \quad L(\hat{\lambda}) = \bar{x}_n^{-n}e^{-n} > 0, \quad L(\infty) = 0.$$

So we have found the maximum likelihood estimate $\hat{\lambda}$ for λ and thus the maximum likelihood estimator for λ is

$$\hat{\Lambda} = \frac{1}{\bar{X}_n}.$$

If, as in this example, the likelihood consists of many factors, it is usually easier to work with the log likelihood

$$l(\theta) = \log L(\theta)$$

because, where the likelihood is a product of many factors, requiring us to use the product rule when we differentiate, the log likelihood is just a sum of many terms. The log likelihood takes its maximum at the same location as the likelihood itself. This is so because

$$\frac{d}{d\theta} l(\theta) = \frac{d}{d\theta} \log L(\theta) = \frac{1}{L(\theta)} \frac{dL}{d\theta}(\theta)$$

and thus

$$\frac{dl}{d\theta}(\hat{\theta}) = 0 \quad \Leftrightarrow \frac{dL}{d\theta}(\hat{\theta}) = 0.$$

Furthermore, because the logarithm is a strictly increasing function, a maximum of $l(\theta)$ is also a maximum of $L(\theta)$. The log likelihood in this example is

$$\begin{aligned} l(\lambda) = \log L(\lambda) &= \log\left(\lambda e^{-\lambda x_1}\right) + \cdots + \log\left(\lambda e^{-\lambda x_n}\right) \\ &= n \log \lambda - \lambda(x_1 + \cdots + x_n). \end{aligned}$$

Differentiating $l(\lambda)$ with respect to λ gives

$$\frac{dl}{d\lambda}(\lambda) = \frac{n}{\lambda} - (x_1 + \cdots + x_n) = n\left(\frac{1}{\lambda} - \bar{x}_n\right).$$

We know that $l(\lambda)$ has an extremum at $\lambda = \hat{\lambda}$ only if

$$\frac{dl}{d\lambda}(\hat{\lambda}) = 0.$$

Thus

$$\hat{\lambda} = \frac{1}{\bar{x}_n}.$$

It is now easy to check that this extremum is indeed a maximum by calculating the second derivative and observing that it is negative:

$$\frac{d^2 l}{d\lambda^2}(\hat{\lambda}) = -\frac{n}{\hat{\lambda}^2} < 0.$$

Thus the maximum likelihood estimator for λ is

$$\hat{\Lambda} = \frac{1}{\bar{X}_n}.$$

Example 21.4 (Example 19.3 continued)**.** In this example the dataset was modelled by random variables N_1, N_2, N_3, N_4 that are jointly multinomially distributed, i.e.,

$$P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) = p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} \frac{n!}{n_1! n_2! n_3! n_4!},$$

where $p_1 = (\theta + 2)/4, p_2 = p_3 = (1 - \theta)/4$, and $p_4 = \theta/4$. We are interested in the parameter θ. We already studied a range of unbiased estimators for θ. Now we want to derive the maximum likelihood estimator for θ.

The likelihood is just the probability to obtain the data, i.e.,

$$\begin{aligned} L(\theta) &= P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) = p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} \frac{n!}{n_1! n_2! n_3! n_4!} \\ &= (\theta + 2)^{n_1} (1 - \theta)^{n_2 + n_3} \theta^{n_4} \frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!}. \end{aligned}$$

Again it will be easier to work with the log likelihood

$$\begin{aligned} l(\theta) &= \log \left((\theta + 2)^{n_1} (1 - \theta)^{n_2 + n_3} \theta^{n_4} \frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!} \right) \\ &= n_1 \log(\theta + 2) + (n_2 + n_3) \log(1 - \theta) + n_4 \log(\theta) + \log \left(\frac{1}{4^n} \frac{n!}{n_1! n_2! n_3! n_4!} \right). \end{aligned}$$

To find the maximum we take the derivative with respect to θ,

$$\frac{d}{d\theta} l(\theta) = \frac{n_1}{\theta + 2} - \frac{n_2 + n_3}{1 - \theta} + \frac{n_4}{\theta}.$$

The maximum is at the value $\theta = \hat{\theta}$ where this derivative vanishes, hence

$$\frac{d}{d\theta} l(\hat{\theta}) = \frac{n_1}{\hat{\theta} + 2} - \frac{n_2 + n_3}{1 - \hat{\theta}} + \frac{n_4}{\hat{\theta}} = 0.$$

Multiplying by $(\theta + 2)(1 - \theta)\theta$ this becomes

$$\begin{aligned} 0 &= n_1(1 - \theta)\theta - (n_2 + n_3)(\theta + 2)\theta + n_4(\theta + 2)(1 - \theta) \\ &= -n\theta^2 + (n_1 - 2(n_2 + n_3) - n_4)\theta + 2n_4. \end{aligned}$$

To save writing we introduce the shorthand $m = n_1 - 2(n_2 + n_3) - n_4$. The positive

solution of this quadratic equation is then

$$\hat{\theta} = \frac{m + \sqrt{m^2 + 8nn_4}}{2n}.$$

Putting in the numerical values from the example, $n_1 = 1997, n_2 = 906, n_3 = 904, n_4 = 32$ and $n = 3839$ gives the estimate

$$\hat{\theta} \approx 0.0357.$$

The estimator for the parameter θ is

$$\hat{\Theta} = \frac{M + \sqrt{M^2 + 8nN_4}}{2n},$$

where $M = N_1 - 2(N_2 + N_3) - N_4$.

We should now compare the quality of this estimator to the estimators we considered earlier. Ideally we would like to calculate the expectation and the mean square error for this estimator. However this will be a very long calculation. Instead you can apply the simulation technique from R lab 6 to approximate the sampling distribution.

Example 21.5. We have observations of the number of earthquakes in the UK in three different years: $n_1 = 16, n_2 = 12, n_3 = 20$. We model these as realisations of three independent Poisson-distributed random variables $N_i \sim \text{Pois}(\lambda)$ for $i = 1, 2, 3..$ This means that

$$P(N_i = n_i) = \begin{cases} \frac{\lambda^n}{n!} e^{-\lambda} & \text{if } n \in \{0, 1, \dots\} \\ 0 & \text{otherwise} \end{cases}.$$

We want to find the maximum likelihood estimator for the parameter λ. The likelihood is

$$\begin{aligned} L(\lambda) &= p_{N_1,N_2,N_3}(n_1, n_2, n_3) = P_{N_1}(n_1) P_{N_2}(n_2) P_{N_3}(n_3) \\ &= \frac{\lambda^{n_1}}{n_1!} e^{-\lambda} \frac{\lambda^{n_2}}{n_2!} e^{-\lambda} \frac{\lambda^{n_3}}{n_3!} e^{-\lambda} = \frac{\lambda^{n_1+n_2+n_3}}{n_1! n_2! n_3!} e^{-3\lambda}. \end{aligned}$$

Again it will be nicer to work with the log likelihood

$$l(\lambda) = \log L(\lambda) = (n_1 + n_2 + n_3) \log(\lambda) - 3\lambda - \log(n_1! n_2! n_3!).$$

At the maximum we will have zero derivative, so we solve

$$0 = \frac{dl}{d\lambda} = \frac{n_1 + n_2 + n_3}{\lambda} - 3,$$

which tells us that the maximum is at

$$\hat{\lambda} = \frac{n_1 + n_2 + n_3}{3} = 16.$$

The corresponding maximum likelihood estimator is

$$\hat{\Lambda} = \frac{N_1 + N_2 + N_3}{3} = \bar{N}_3.$$

This is an unbiased estimator because

$$E[\hat{\Lambda}] = E\left[\frac{N_1 + N_2 + N_3}{3}\right] = \frac{1}{3}\left(E[N_1] + E[N_2] + E[N_3]\right) = \lambda.$$

Its mean squared error is

$$\begin{aligned} MSE(\hat{\Lambda}) &= \mathrm{Var}\left[\hat{\Lambda}\right] = \mathrm{Var}\left[\frac{N_1 + N_2 + N_3}{3}\right] \\ &= \frac{1}{9}\left(\mathrm{Var}\left[N_1\right] + \mathrm{Var}\left[N_2\right] + \mathrm{Var}\left[N_3\right]\right) = \frac{\lambda}{3}. \end{aligned}$$

Example 21.6. Let a dataset $x_1, \ldots, x_n > 0$ be modelled by an i.i.d. sample from the uniform distribution $X \sim \mathrm{U}(0, \theta)$. We want to find the maximum likelihood estimate for θ. The likelihood function is

$$L(\theta) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) = \begin{cases} 0 & \text{if } \theta < \max\{x_1, \ldots, x_n\} \\ (1/\theta)^n & \text{otherwise} \end{cases}$$

because if any of the observed values is larger than θ then the corresponding density function is zero and if it is below θ the density is equal to $1/\theta$. The maximum likelihood estimate $\hat{\theta}$ for θ is the value for θ at which $L(\theta)$ takes its maximal value, and thus

$$\hat{\theta} = \max\{x_1, x_n\}.$$

The corresponding maximum likelihood estimator is

$$\hat{\Theta} = \max\{X_1, \ldots, X_n\}.$$

Next we turn to a very important example in which we have two model parameters:

Example 21.7. Let a dataset $x_1, \cdots, x_n$ be modelled as an i.i.d. sample $X_1, \ldots, X_n$ from a normal distribution $\mathrm{N}(\mu, \sigma^2)$. We want to determine maximum likelihood estimators for μ and σ^2. So in this case the model parameter is $\theta = (\mu, \sigma^2)$. This just means that the likelihood will be a function of both μ and σ^2 and we have to maximise it with

respect to both. The likelihood is

$$L(\mu, \sigma^2) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$

where the probability density for the $\mathrm{N}(\mu, \sigma^2)$ distributed random variables is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This is one of the formulas that it is worth learning by heart. We will work with the log likelihood, so we need to calculate

$$\log f_X(x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}.$$

This gives us for the log likelihood

$$\begin{aligned} l(\mu, \sigma) &= \log L(\mu, \sigma) = \log f_X(x_1) + \cdots + \log f_X(x_n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right). \end{aligned}$$

We want to find the extrema of this function of two variables. It turns out (and you will learn about this properly next year) that this can be done in a way very similar to finding the maximum of a function of a single variable by setting the derivatives to zero. Only there are now two derivatives, one with respect to μ and one with respect to σ^2, and they both vanish at the extremum. Let us calculate these derivatives.

$$\begin{aligned} \frac{\partial l}{\partial \mu}(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}\left(-2(x_1-\mu) - \cdots - 2(x_n-\mu)\right) \\ &= \frac{1}{\sigma^2}(x_1 + \cdots + x_n - n\mu) = \frac{n}{\sigma^2}(\bar{x}_n - \mu), \\ \frac{\partial l}{\partial \sigma^2}(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right) \\ &= -\frac{n}{2\sigma^4}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

For reasons that don't have to concern us here, in the case where the function depends on more than one variable, the derivatives are written with rounded ∂ rather than d, but these so-called partial derivatives are just the usual derivatives you know. While differentiating with respect to one variable you simply treat the other variable as a constant.

There is an extremum at $\mu = \hat{\mu}, \sigma^2 = \hat{\sigma^2}$ only if

$$\frac{\partial l}{\partial \mu}(\hat{\mu}, \hat{\sigma^2}) = 0 = \frac{\partial l}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma^2}).$$

From the first condition we find $\hat{\mu}$:

$$0 = \frac{\partial l}{\partial \mu}(\hat{\mu}, \hat{\sigma^2}) = \frac{n}{\hat{\sigma}^2}(\bar{x}_n - \hat{\mu}) \quad \Longrightarrow \quad \hat{\mu} = \bar{x}_n.$$

We can then use this in the second condition,

$$0 = \frac{\partial l}{\partial \sigma}(\hat{\mu}, \hat{\sigma}) = -\frac{n}{2\hat{\sigma}^4}\left(\hat{\sigma}^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)$$

and see that this implies

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

This extremum is in fact a maximum, as can be discovered by thinking about the shape of the function, but we did not discuss this in detail.

We have thus found the maximum likelihood estimates for the mean and the standard deviation. Correspondingly, the maximum likelihood estimator for the mean is

$$\hat{M}_n = \bar{X}_n,$$

and the maximum likelihood estimator for the variance is

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

Note that this estimator has a different normalisation from the unbiased estimator S_n^2 that we met in Theorem 19.2. This means that the maximum likelihood estimator is not unbiased. Instead

$$E\left[\hat{\Sigma_n^2}\right] = \frac{n+1}{n}E[S_n^2] = \frac{n+1}{n}\sigma^2.$$

However the bias gets smaller as the sample size n increases and goes away in the limit $n \to \infty$,

$$\lim_{n\to\infty} E\left[\hat{\Sigma_n^2}\right] = \sigma^2.$$

We say that the estimator is asymptotically unbiased.

One of the reasons why maximum likelihood estimators are so important is that in many situations they can be shown to have desirable properties. We will not go into details in this module.

22 Simple linear regression

(Textbook chapter link)

In the lectures I gave some motivation and drew a picture, here I just reproduce the definition (see however Figure 22.1 for example 22.3).

Definition 22.1. A **simple linear regression model** model for a bivariate dataset $(x_1, y_1), \ldots, (x_n, y_n)$ consists of an i.i.d. sample $(X_1, Y_1, R_1), \ldots, (X_n, Y_n, R_n)$. The conditional probability distribution of Y_i given that $\{X_i = x_i\}$ is specified by

$$Y_i|\{X_i = x_i\} = \alpha + \beta x_i + R_i$$

for $i = 1, \ldots, n$. The model parameters are the intercept α and the slope β of the **regression line** $y = \alpha + \beta x$. The R_i are the **residuals**.

Graphically the values r_i of the residuals give the vertical displacement of the data points from the regression line,

$$r_i = y_i - \alpha - \beta x_i.$$

By choosing to model the data in this way, we assume that X influences Y. But a scatter plot with a trend like that in Figure 22.1 could arise also because Y influences X or because some other variable Z influences both X and Y.

We now want to estimate the parameters α and β using the maximum likelihood principle. For that we need to make an assumption about the distribution of the R_i. We assume that

$$R_i \sim \mathrm{N}(0, \sigma^2).$$

This then implies that

$$Y|\{X = x\} \sim \mathrm{N}(\alpha + \beta x, \sigma^2).$$

Because the linear regression model specifies only the distribution of the Y_i given the X_i, we only need to maximize the likelihood of the y_i given that $X_i = x_i$.

$$L(\alpha, \beta) = f_{Y_1|\{X_1=x_1\}}(y_1) \cdot \cdots \cdot f_{Y_n|\{X_n=x_n\}}(y_n)$$

with

$$f_{Y_i|\{X_i=x_i\}}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right).$$

It will be convenient to work with the log likelihood, so we observe that

$$\log f_{Y_i|\{X_i=x_i\}}(y_i) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

and therefore

$$l(\alpha,\beta) = \log L(\alpha,\beta) = \sum_{i=1}^{n} \log(f_{Y_i|\{X_i=x_i\}}(y_i))$$
$$= n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2.$$

We see that maximizing the log likelihood is the same as minimizing the sum of the squares of the residuals,

$$S(\alpha,\beta) = \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 = \sum_{i=1}^{n} r_i^2.$$

For this reason this estimation procedure is also called the least squares estimation. The idea of minimizing the square of the residuals actually pre-dates the development of the maximum likelihood principle, but it lacked a rational justification.

Below we will save space by writing just $\sum$ for $\sum_{i=1}^{n}$. The function $S(\alpha,\beta)$ is a quadratic function in α and β. The graph of the function looks like a parabolic bowl, with its minimum at $(\alpha,\beta) = (\hat{\alpha},\hat{\beta})$. The location of the minimum is found from the condition

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha},\hat{\beta}) = 0 = \frac{\partial S}{\partial \beta}(\hat{\alpha},\hat{\beta}).$$

We calculate

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha},\hat{\beta}) = 2n\hat{\alpha} - 2\sum y_i + 2\hat{\beta}\sum x_i = 2n(\hat{\alpha} - \bar{y}_n + \hat{\beta}\bar{x}_n).$$

This is zero iff

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n. \tag{22.1}$$

Also

$$\frac{\partial S}{\partial \beta}(\hat{\alpha},\hat{\beta}) = 2\sum\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)(-x_i)$$
$$= 2\left(-\sum x_i y_i + \sum x_i(\bar{y}_n - \hat{\beta}\bar{x}_n) + \hat{\beta}\sum (x_i)^2\right)$$
$$= 2\left(-\sum x_i(y_i-\bar{y}_n) + \hat{\beta}\sum x_i(x_i-\bar{x}_n)\right).$$

Clearly this is zero if and only if

$$\hat{\beta} = \frac{\sum x_i(y_i-\bar{y}_n)}{\sum x_i(x_i-\bar{x}_n)}.$$

Alternative ways of writing the same expression are

$$\hat{\beta} = \frac{\sum(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum(x_i - \bar{x}_n)^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}. \tag{22.2}$$

We have derived the maximum likelihood estimates. As usual, we obtain the corresponding estimators by replacing the sample values by their corresponding random variables. For the parameters α and β in the linear regression model it is conventional to use the same symbols $\hat{\alpha}$ and $\hat{\beta}$ to denote both the estimates and the estimators. We can summarize what we have learned in the following theorem:

Theorem 22.2. *The* **least squares estimators** *for the parameters α and β of the linear regression model are*

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{X}_n,$$
$$\hat{\beta} = \frac{\sum(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum(X_i - \bar{X}_n)^2}.$$

Example 22.3. Throughout this chapter we will illustrate the methods on a very simple dataset consisting of only three observations:

	x	y
1	1	2
2	3	1.8
3	5	1

which we could also write in terms of pairs (x_i, y_i) as $(1, 2), (3, 1.8), (5, 1)$. Figure 22.1 shows a scatter plot of these data points as well as the regression line.

We now calculate the estimates for α and β by substituting the values from the dataset into the expressions for the estimators. First we calculate the sums

$$\sum x_i = 1 + 3 + 5 = 9, \quad \sum y_i = 2 + 1.8 + 1 = 4.8,$$

$$\sum x_i^2 = 1 + 9 + 25 = 35, \quad \sum x_i y_i = 2 + 5.4 + 5 = 12.4.$$

We also have $n = 3$. Substituting these values into the expression (22.2) for $\hat{\beta}$ gives

$$\hat{\beta} = \frac{3 \cdot 12.4 - 9 \cdot 4.8}{3 \cdot 35 - 9^2} = -\frac{1}{4}.$$

Then the expression (22.1) for $\hat{\alpha}$ gives

$$\hat{\alpha} = \frac{4.8}{3} - \frac{-1}{4} \cdot \frac{9}{3} = 2.35.$$

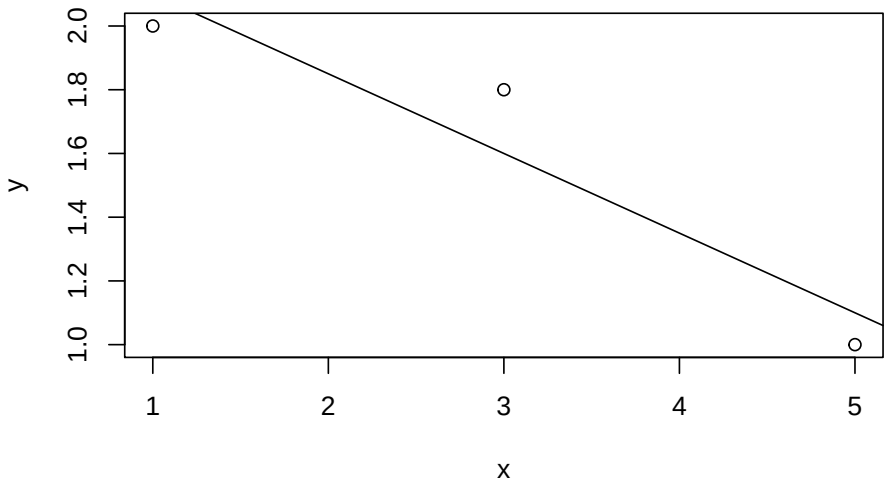


Figure 22.1: Scatterplot with regression line for dataset in Example 22.3

We can use the regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$ to make predictions for y value at given x. We have

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 2.35 - 0.25 \cdot x.$$

For example at $x = 2$ this predicts

$$\hat{y} = 2.35 - 0.5 = 1.85.$$

Usually the variation in the measurements of one variable Y has many causes, of which the explanatory variable X is just one. The coefficient of determination is a tool to measure how much of the variation in Y is caused by X.

Definition 22.4. The **coefficient of determination** R^2 of a linear regression model is defined as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where RSS is the residual sum of squares,

$$\text{RSS} = \sum_{i=1}^{n} R_i^2$$

and TSS is the total sum of squares,

$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

You can think of TSS as the sum of squares of the residuals in a model that fits a horizontal, x-independent line through the data. If such an x-independent model fits the data as well as the linear regression model, then R^2 will be equal to 0. This would tell us that the variable x does not contribute at all to the explanation of the variation in y. At the other extreme, if R^2 was equal to 1 then the variable x would explain the variation in y completely, because $\mathrm{RSS} = 0$ would mean that the data was described perfectly by the regression line. In a real situation R^2 will lie somewhere between 0 and 1.

Example 22.3. (continued) We find

$$\begin{aligned} r_1 &= y_1 - \hat{\alpha} - \hat{\beta}x_1 = 2 - 2.35 + 0.25 = -0.1, \\ r_2 &= y_2 - \hat{\alpha} - \hat{\beta}x_2 = 1.8 - 2.35 + 3 \cdot 0.25 = 0.2, \\ r_3 &= y_3 - \hat{\alpha} - \hat{\beta}x_3 = 1 - 2.35 + 5 \cdot 0.25 = -0.1. \end{aligned}$$

We also have

$$\bar{y}_n = \frac{\sum y_i}{n} = \frac{4.8}{3} = 1.6$$

and thus

$$\begin{aligned} \mathrm{RSS} &= (-0.1)^2 + 0.2^2 + (-0.1)^2 = 0.06, \\ \mathrm{TSS} &= (2 - 1.6)^2 + (1.8 - 1.6)^2 + (1 - 1.6)^2 = 0.56, \\ R^2 &= 1 - \frac{0.06}{0.56} = \frac{25}{28} \approx 0.8929. \end{aligned}$$

It is important to recognise that the simple linear regression model makes various assumptions about the data, which should be checked before blindly using the model:

1. On average, y is a linear function of x.
2. The residuals are identically distributed. In particular, the variance of the residuals is independent of x. This feature is referred to as "**homoscedasticity**". The lack of this feature is known as "**heteroscedasticity**".
3. The observations are independent.
4. The residuals are approximately normally distributed (so that the least squares estimation is justified).

In the lecture this was discussed further with the help of R code examples, which can be found on Moodle. You will investigate more examples in the R lab.

23 Confidence intervals for the mean

(Textbook chapter link)

Consider the distribution of an estimator $\hat{\Theta}$ for some model parameter θ. If the estimator is at all good, its density will be concentrated near the true value θ. In this chapter we assume that the estimator is a continuous random variable. Figure 23.1 sketches an example of a density function. We know however, that the probability that

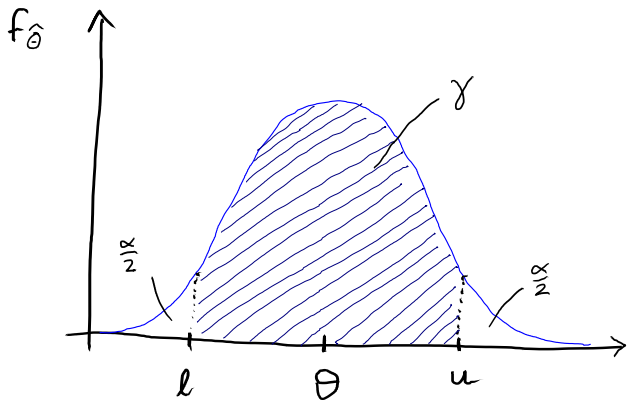


Figure 23.1: Sketch of the density function of an estimator $\hat{\Theta}$

the estimator will give exactly the correct result is zero, because, according to Theorem 5.4, for any continuous random variable $\hat{\Theta}$, $P(\hat{\Theta} = \theta) = 0$. The best we can hope for is that with high probability the estimator gives a value that is close to the true value, say within an interval from $\theta - a$ to $\theta + b$ for some $a, b \in \mathbb{R}$. So we consider the probability

$$P(\theta - a \leq \hat{\Theta} \leq \theta + b) = \gamma = 1 - \alpha.$$

The probability γ is the area of the shaded region under the density function in Figure 23.1. In the special case where the density function $f_{\hat{\Theta}}(x)$ is symmetric around $x = \theta$ and $a = b$, the remaining probability $\alpha = 1 - \gamma$ is split equally between the right and left tail, again indicated in Figure 23.1. The above equation is not yet very useful to us, because we do not know the true value θ and therefore also do not know the location of the interval $(\theta - a, \theta + b)$. However we can use that

$$P(\theta - a < \hat{\Theta} < \theta + b) = P(\hat{\Theta} - b < \theta < \hat{\Theta} + a) = \gamma.$$

Now we have a random interval $(L, U) = (\hat{\Theta} - b, \hat{\Theta} + a)$ that contains the true value with probability γ. When we evaluate the random variables L and U on our data we obtain the so-called confidence interval.

The random variables L and U giving the lower and upper end of the random interval

are not always in the form given above. The following definition allows more general cases:

Definition 23.1. Suppose a dataset $x_1, \ldots, x_n$ is modelled by random variables $X_1, \ldots, X_n$. Let θ be the model parameter and let $\gamma \in [0, 1]$. If there exist random variables $L = g(X_1, \ldots, X_n)$ and $U = h(X_1, \ldots, X_n)$ such that

$$P(L < \theta < U) = \gamma$$

for any value of θ. Then the interval

$$(l, u)$$

is a $100\gamma\%$ **confidence interval** for θ, where $l = g(x_1, \ldots, x_n)$ and $u = h(x_1, \ldots, x_n)$. γ is the **confidence level**. If we only have $P(L < \theta < U) \geq \gamma$ then we speak of a **conservative confidence interval**.

Note that while the random interval (L, U) contains the true value for θ with probability γ, it would be incorrect to say that therefore the interval (l, u) contains the true value θ with probability γ. Once we have evaluated the random variables using the data, a traditional statistician would no longer speak of probability. We now have a definite interval and the true value either lies in it or it does not. It is the same as when your football team has played, the game is over, but you are away and have not yet heard the result. Even though you don't know the result, your team has either won or they have lost. There is nothing probable about it any more. You can still speak about how confidently you believe that they have won, but you should not speak of the probability that they have won. Hence we call γ the confidence level.[8]

In this module we will concentrate on the case where the data is modelled as an i.i.d. sample and the model parameter for which we want to know a confidence interval is the expectation of the model distribution.

Let us first consider the case where the i.i.d. sample is from a normal distribution.

Notation: We introduce notation for the percentiles of the standard normal distribution.

$$z_p = \Phi^{-1}(1 - p) = (1 - p) \text{ quantile.}$$

Here $\Phi(x)$ is the standard normal distribution function. Equivalently, if $Z \sim \mathrm{N}(0, 1)$, then

$$P(Z > z_p) = 1 - P(Z \leq z_p) = 1 - \Phi(z_p) = 1 - (1 - p) = p.$$

Because of the symmetry of the standard normal distribution, $\Phi(z) = 1 - \Phi(-z)$, we have

$$z_{1-p} = -z_p.$$

[8] A Bayesian statistician on the contrary is happy to use probability to describe confidence.

Also

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= P(Z > -z_{\alpha/2}) - P(Z > z_{\alpha/2}) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha. \end{aligned}$$

Theorem 23.2. *Suppose a dataset $x_1, \ldots, x_n$ is modelled as an i.i.d. sample $X_1, \ldots, X_n$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution with unknown mean but known variance. Then the interval*

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \ , \ \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

is a $100(1-\alpha)\%$ confidence interval for μ.

Proof. According to Definition 23.1 we need to show that

$$P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha.$$

We know that the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ is normally distributed, $\bar{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$. Therefore

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

Therefore

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right) \\ &= P\left(-\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \\ &= P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right). \end{aligned}$$

□

Example 23.3. You are given a dataset $x_1, \ldots, x_{36}$ with $\bar{x}_{36} = 74.8$ and model it with an i.i.d. sample $X_1, \ldots, X_{36}$ from a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\sigma = 12$. Give a 95% confidence interval for μ.

Solution. Writing 95% as $100(1-\alpha)\%$ gives $\alpha = 0.05$ and hence $\alpha/2 = 0.025$. The value $z(0.025) \approx 1.96$ can be found in statistical tables or with the R command `qnorm(1-0.025)`. Using the expression for the $100(1-\alpha)\%$ confidence interval from

Proposition 23.2 then gives

$$\begin{aligned}&\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\\ =&\left(74.8 - z(0.025)\frac{12}{\sqrt{36}}, 74.8 + z(0.025)\frac{12}{\sqrt{36}}\right)\\ \approx&(74.8 - 3.92, 74.8 + 3.92)\\ =&(70.88, 78.72).\end{aligned}$$

Let us assume that the true value of μ is $\mu = 72$. The above confidence interval *certainly* contains the true value of μ. It is not true that there is only a 95% probability that it includes the true value.

Now consider the case where a new independent observation gives new sample values $x_1, \ldots, x_{36}$ with $\bar{x}_n = 76.68$. Then the 95% confidence interval is

$$\begin{aligned}&\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\\ \approx&(76.68 - 3.92, 76.68 + 3.92)\\ \approx&(72.76, 80.60).\end{aligned}$$

This confidence interval certainly does *not* contain the true value of $\mu = 72$. The only correct probability statement is that the random interval

$$\left(\bar{X}_{36} - 3.92, \bar{X}_{36} + 3.92\right)$$

contains the true value of $\mu = 72$ with probability 0.95.

Note that in the proof of Proposition 23.2 we only used that the sample mean $\bar{X}_n$ is normal. This is guaranteed when the sample distribution is normal. But it is also approximately true for any sufficiently large i.i.d. sample as long as the sample distribution has finite mean and finite variance. In all those cases we can use the above formula for the confidence interval for the mean, and the error we make will be negligible provided the sample is large enough.

Rule of thumb 23.4. Let n be large ($n \gtrapprox 20$) and let a dataset be modelled as an i.i.d. sample $X_1, \ldots, X_n$ from a distribution with expectation μ and variance σ^2. Then the interval

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;\;,\;\; \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for μ.

Note that the above expressions for the confidence interval require us to know the variance of the distribution. Often the variance is not known either and needs to be

estimated from the data. We know from Theorem 19.2 that the sample variance S_n^2 is an unbiased estimator for the variance. So we would like to replace σ by S_n in Proposition 23.2, i.e., instead of working with the variable

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

we work with the variable

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

Replacing σ by its estimate introduces extra uncertainty and thus our confidence will be lowered. The effect of that is that T follows not the standard normal distribution but instead the t-distribution, which is slightly broadened compared to the standard normal distribution.

Definition 23.5. A continuous random variable T_m has a t-distribution with parameter $m > 1$ (degrees of freedom) if its density function is given as

$$f_{T_m}(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in \mathbb{R},$$

where k_m is the normalisation constant that ensures that the area under the density curve is 1. We write $T_m \sim t(m)$.

In this module you will not be expected to work with this analytic expression for the density function of the t-distribution but will only need to be able to look up its quantiles in the statistical tables or ask a computer to calculate it.

We can now state without proof the distribution of T:

Theorem 23.6. *Let $X_1, \ldots X_n$ be an i.i.d. sample from $\mathrm{N}(\mu, \sigma^2)$. Then*

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1),$$

where $\bar{X}_n$ is the sample mean and S_n^2 is the sample variance (defined in Theorem 19.2).

Notation: We denote the $(1-p)$th quantile of the $t(m)$ distribution by $t_{m,p}$, so that if $T_m \sim t(m)$,

$$P(T_m > t_{m,p}) = p.$$

We can therefore now state the analogue of Proposition 23.2 for the case where the variance of the sample distribution is not known but needs to be estimated by replacing $Z \sim \mathrm{N}(0,1)$ by $T \sim t(n-1)$ and thus replacing $z_{\alpha/2}$ by $t_{n-1,\alpha/2}$.

Proposition 23.7. *Suppose a dataset $x_1, \ldots, x_n$ is modelled as an i.i.d. sample $X_1, \ldots, X_n$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution. Then the interval*

$$\left[\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \;\;,\;\; \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ, where $\bar{x}_n$ is the value of the sample mean on the dataset and s_n is the value of the square root of the sample variance,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

Example 23.8. The reading age of children about to start secondary school is a measure of how good they are at reading and understanding printed text. A child's reading age, measured in years, is modelled by the random variable $X \sim \mathrm{N}(\mu, \sigma^2)$. The reading ages of an i.i.d. sample of 20 children were measured, and the data obtained is summarised as follows:

$$\sum_{i=1}^{20} x_i = 232.6, \qquad \sum_{i=1}^{20} x_i^2 = 2756.22.$$

a) Calculate unbiased estimates for μ and σ^2.
b) Calculate a 95% confidence interval for μ.

Solution.

a) We have unbiased estimators for μ and σ from Theorem 19.2. These give the estimates

$$\hat{\mu} = \bar{x}_{20} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{232.6}{20} = 11.63$$

and

$$\begin{aligned}
\hat{\sigma}^2 = s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{20}(x_i - \bar{x}_n)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{20} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{20} x_i\right)^2\right) \\
&= \frac{1}{19}\left(2756.22 - \frac{(232.6)^2}{20}\right) \\
&\approx 2.689.
\end{aligned}$$

b) By Proposition 23.7 the 95% confidence interval is

$$\begin{aligned}&\left[\bar{x}_{20} - t_{19}(0.025)\frac{s_n}{\sqrt{20}}, \bar{x}_{20} + t_{19}(0.025)\frac{s_n}{\sqrt{20}}\right]\\ \approx &\left[11.63 - 2.093\frac{\sqrt{2.688}}{\sqrt{20}}, 11.63 - 2.093\frac{\sqrt{2.688}}{\sqrt{20}}\right]\\ \approx &[10.86, 12.40].\end{aligned}$$

The value $t_{19}(0.025) \approx 2.093$ can be found in statistical tables or can be calculated with the R command `qt(1-0.025, 19)`.

As the sample size gets bigger, the estimate s_n for σ gets more and more reliable, and thus the result from Proposition 23.7 should go towards that of Proposition 23.2. This is indeed the case because for sufficiently large n

$$t_{n-1,\alpha/2} \approx z_{\alpha/2}.$$

Now that we understand the method of deriving confidence intervals for the mean, we can apply the same method to other circumstances. In particular we can derive confidence intervals for the probabilities in our smarties example.

Recall that we had introduced the random variable Y that gives the number of yellow smarties in a box of n smarties. Under the assumption of independence between the smarties, we have that $Y \sim \text{Bin}(n, p)$, where p is the probability that an individual smartie is yellow. We already know, from the law of large numbers, that Y/n is an unbiased estimator for the probability p. We now want to derive a confidence interval for p.

We can write Y as the sum of independent and identically distributed random variables, $Y = \sum_{i=1}^{n} Y_i$, where Y_i is the indicator random variable for the event that the i-th smartie is yellow. Therefore we know from the central limit theorem rule of thumb that Y is approximately normally distributed. The mean and variance we can easily calculate, giving

$$Y \dot{\sim} \text{N}\left(np, np(1-p)\right).$$

We can transform this to give an approximately normally distributed random variable,

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \dot{\sim} \text{N}(0, 1).$$

So we know that

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \approx 1 - \alpha.$$

We now only have to translate the condition on Z into a condition on p to get our

approximate $100(1-\alpha)\%$ confidence interval for p.

$$\begin{aligned}
& -z_{\alpha/2} < Z < z_{\alpha/2} \\
\Leftrightarrow\ & Z^2 < z_{\alpha/2}^2 \\
\Leftrightarrow\ & \frac{(Y-np)^2}{np(1-p)} < z_{\alpha/2}^2 \\
\Leftrightarrow\ & (Y-np)^2 - np(1-p)z_{\alpha/2}^2 < 0 \\
\Leftrightarrow\ & L < p < U,
\end{aligned}$$

where L and P are the solutions of the quadratic equation $(Y-np)^2 - np(1-p)z_{\alpha/2}^2 = 0$:

$$U = \frac{Y + \frac{1}{2}z_{\alpha/2}^2 + \sqrt{\frac{1}{4}z_{\alpha/2}^4 + z_{\alpha/2}^2(Y - Y^2/n)}}{n + z_{\alpha/2}^2},$$

$$L = \frac{Y + \frac{1}{2}z_{\alpha/2}^2 - \sqrt{\frac{1}{4}z_{\alpha/2}^4 + z_{\alpha/2}^2(Y - Y^2/n)}}{n + z_{\alpha/2}^2}.$$

We have shown that the probability that the random interval (L, U) contains p approximately with probability $1-\alpha$. Hence the interval (l, u) obtained by putting in the observed value y for Y gives an approximate $100(1-\alpha)\%$ confidence interval for p.

For example in the case where 3 yellow smarties were observed in a box of 40, the point estimate for p is $3/40 = 0.075$ and the 95% confidence interval is approximately $(0.026, 0.199)$. Note how this is not symmetric about the point estimate.

If instead of a single box of 40 smarties we look at all the observed 11063 smarties of which 1409 were yellow, we obtain a point estimate for p of 0.127 and a 95% confidence interval of $(0.121, 0.134)$. It so happens that the value of $p = 1/8 = 0.125$ that we would have if all 8 colours were equally likely is contained in this confidence interval.