Math 5010
Introduction to Probability

Based on D. Stirzaker's book

Cambridge University Press

and

R.B. Ash's book

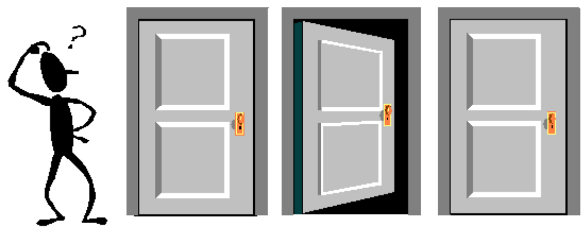
Dover Publications

Davar Khoshnevisan
University of Utah

Firas Rassoul-Agha
University of Utah

Last Scribed March 25, 2014

Lecture 1



Example 1.1 (The Monty Hall problem, Steve Selvin, 1975)**.** Three doors: behind one is a nice prize; behind the other two lie goats. You choose a door. The host (Monty Hall) knows where the prize is and opens a door that has a goat. He gives you the option of changing your choice to the remaining unopened door. Should you take his offer?

The answer is "yes." Indeed, if you do not change your mind, then to win you must choose the prize right from the start. This is 1 in 3. If you do change your mind, then you win if you choose a goat right from the start (for then the host opens the other door with the goat and when you switch you have the prize). This is 2 in 3. Your chances double if you switch.

1. The sample space, events, and outcomes

We need a math model for describing "random" events that result from performing an "experiment."

We cannot use "frequency of occurrence" as a model because it does not have the power of "prediction." For instance, if our definition of a fair coin is that the frequency of heads has to converge to $1/2$ as the number of tosses grows to infinity, then we have done things backwards. We have used our prediction to make a definition. What we should do is first define a model, then draw from it that prediction about the frequency of heads.

Here is how we will do things. First, we define a *state space* (or sample space) that we will denote by Ω. We think of the elements of Ω as *outcomes* of the experiment.

Then, we specify is a collection $\mathscr{F}$ of subsets of Ω. Each of these subsets is called an *event*. These events are the ones we are "allowed" to talk about the probability of. When Ω is finite, $\mathscr{F}$ can be taken to be the collection of all subsets of Ω.

The next step is to assign a *probability* $P(A)$ to every $A \in \mathscr{F}$. We will talk about this after the following examples.

Example 1.2. Roll a six-sided die; what is the probability of rolling a six? First, write a sample space. Here is a natural one:

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

In this case, Ω is finite and we want $\mathscr{F}$ to be the collection of all subsets of Ω. That is,

$$\mathscr{F} = \Big\{\varnothing, \{1\}, \dots, \{6\}, \{1, 2\}, \dots, \{1, 6\}, \dots, \{1, 2, \dots, 6\}\Big\}.$$

Example 1.3. Toss two coins; what is the probability that we get two heads? A natural sample space is

$$\Omega = \Big\{(H_1, H_2), (H_1, T_2), (T_1, H_2), (T_1, T_2)\Big\}.$$

Once we have readied a sample space Ω and an event-space $\mathscr{F}$, we need to assign a probability to every event. This assignment cannot be made at whim; it has to satisfy some properties.

2. Rules of probability

Rule 1. $0 \leqslant P(A) \leqslant 1$ for every event A.

Rule 2. $P(\Omega) = 1$. "Something will happen with probability one."

Rule 3 (Addition rule)**.** If A and B are disjoint events [i.e., $A \cap B = \varnothing$], then the probability that at least one of the two occurs is the sum of the individual probabilities. More precisely put,

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

Lemma 1.4. *Choose and fix an integer* $n \geqslant 1$. *If* $A_1, A_2, \dots, A_n$ *are pairwise disjoint events, then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P(A_1) + \cdots + P(A_n).$$

Proof. The proof uses *mathematical induction*.

Claim. If the assertion is true for $n - 1$, then it is true for n.

The assertion is clearly true for $n = 1$, and it is true for $n = 2$ by Rule 3. Because it is true for $n = 2$, the Claim shows that the assertion holds for $n = 3$. Because it holds for $n = 3$, the Claim implies that it holds for $n = 4$, etc.

Proof of Claim. We can write $A_1 \cup \cdots \cup A_n$ as $A_1 \cup B$, where $B = A_2 \cup \cdots \cup A_n$. Evidently, A_1 and B are disjoint. Therefore, Rule 3 implies that $P(A) = P(A_1 \cup B) = P(A_1) + P(B)$. But B itself is a disjoint union of $n - 1$ events. Therefore $P(B) = P(A_2) + \cdots + P(A_n)$, thanks to the assumption of the Claim ["the induction hypothesis"]. This ends the proof. $\square$

Homework Problems

Exercise 1.1. You ask a friend to choose an integer N between 0 and 9. Let $A = \{N \leqslant 5\}$, $B = \{3 \leqslant N \leqslant 7\}$ and $C = \{N \text{ is even and } > 0\}$. List the points that belong to the following events:

(a) $A \cap B \cap C$
(b) $A \cup (B \cap C^c)$
(c) $(A \cup B) \cap C^c$
(d) $(A \cap B) \cap ((A \cup C)^c)$

Exercise 1.2. Let A, B and C be events in a sample space Ω. Prove the following identities:

(a) $(A \cup B) \cup C = A \cup (B \cup C)$
(b) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
(c) $(A \cup B)^c = A^c \cap B^c$
(d) $(A \cap B)^c = A^c \cup B^c$

Exercise 1.3. Let A, B and C be arbitrary events in a sample space Ω. Express each of the following events in terms of A, B and C using intersections, unions and complements.

(a) A and B occur, but not C;
(b) A is the only one to occur;
(c) at least two of the events A, B, C occur;
(d) at least one of the events A, B, C occurs;
(e) exactly two of the events A, B, C occur;
(f) exactly one of the events A, B, C occurs;
(g) not more than one of the events A, B, C occur.

Exercise 1.4. Two sets are disjoint if their intersection is empty. If A and B are disjoint events in a sample space Ω, are A^c and B^c disjoint? Are $A \cap C$ and $B \cap C$ disjoint ? What about $A \cup C$ and $B \cup C$?

Exercise 1.5. We roll a die 3 times. Give a probability space $(\Omega, \mathcal{F}, P)$ for this experiment.

Exercise 1.6. An urn contains three chips: one black, one green, and one red. We draw one chip at random. Give a sample space Ω and a collection of events $\mathcal{F}$ for this experiment.

Exercise 1.7. If $A_n \subset A_{n-1} \subset \cdots \subset A_1$, show that $\cap_{i=1}^n A_i = A_n$ and $\cup_{i=1}^n A_i = A_1$.

Lecture 2

Let us recall this set-theoretical notation.

1. Algebra of events

Given two sets A and B that are subsets of some bigger set Ω:

- $A \cup B$ is the "union" of the two and consists of elements belonging to either set; i.e. $x \in A \cup B$ is equivalent to $x \in A$ *or* $x \in B$.
- $A \cap B$ is the "intersection" of the two and consists of elements shared by the two sets; i.e. $x \in A \cap B$ is equivalent to $x \in A$ *and* $x \in B$.
- A^c is the "complement" of A and consists of elements in Ω that are *not* in A.

We write $A \smallsetminus B$ for $A \cap B^c$; i.e. elements in A but not in B.

Clearly, $A \cup B = B \cup A$ and $A \cap B = B \cap A$. Also, $A \cup (B \cup C) = (A \cup B) \cup C$, which we simply write as $A \cup B \cup C$. Thus, it is clear what is meant by $A_1 \cup \cdots \cup A_n$. Similarly for intersections.

We write $A \subseteq B$ when A is inside B; i.e. $x \in A$ implies $x \in B$. It is clear that if $A \subseteq B$, then $A \cap B = A$ and $A \cup B = B$. Thus, if $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n$, then $\cap_{i=1}^n A_i = A_1$ and $\cup_{i=1}^n A_i = A_n$.

It is clear that $A \cap A^c = \varnothing$ and $A \cup A^c = \Omega$. It is also not very hard to see that $(A \cup B)^c = A^c \cap B^c$. (Not being in A or B is the same thing as not being in A and not being in B.) Similarly, $(A \cap B)^c = A^c \cup B^c$.

We say that $A_1, \cdots, A_n$ are disjoint if $\cap_{i=1}^n A_i = \varnothing$. We say they are pair-wise disjoint if $A_i \cap A_j = \varnothing$, for all $i \neq j$.

Example 2.1. The sets $\{1, 2\}$, $\{2, 3\}$, and $\{1, 3\}$ are disjoint but not pair-wise disjoint.

Example 2.2. If A and B are disjoint, then $A \cup C$ and $B \cup C$ are disjoint only when $C = \varnothing$. To see this, we write $(A \cup C) \cap (B \cup C) = (A \cap B) \cup C = \varnothing \cup C = C$. On the other hand, $A \cap C$ and $B \cap C$ are obviously disjoint.

Example 2.3. If A, B, C, and D are some events, then the event "B and at least A or C, but not D" is written as $B \cap (A \cup C) \smallsetminus D$ or, equivalently, $B \cap (A \cup C) \cap D^c$. Similarly, the event "A but not B, or C and D" is written $(A \cap B^c) \cup (C \cap D)$.

Example 2.4. To be more concrete, let $A = \{1, 3, 7, 13\}$, $B = \{2, 3, 4, 13, 15\}$, $C = \{1, 2, 3, 4, 17\}$, $D = \{13, 17, 30\}$. Then, $A \cup C = \{1, 2, 3, 4, 7, 13, 17\}$, $B \cap (A \cup C) = \{2, 3, 4, 13\}$, and $B \cap (A \cup C) \smallsetminus D = \{2, 3, 4\}$. Similarly, $A \cap B^c = \{1, 7\}$, $C \cap D = \{17\}$, and $(A \cap B^c) \cup (C \cap D) = \{1, 7, 17\}$.

Example 2.5. We want to write the solutions to $|x-5| + |x-3| \geqslant |x|$ as a union of disjoint intervals. For this, we first need to figure out what the absolute values are equal to. There are four cases. If $x \leqslant 0$, then the inequality becomes $5 - x + 3 - x \geqslant -x$, that is $8 \geqslant x$, which is always satisfied (when $x \leqslant 0$). Next, is the case $0 \leqslant x \leqslant 3$, and then we have $5 - x + 3 - x \geqslant x$, which means $8 \geqslant 3x$, and so $8/3 < x \leqslant 3$ is not allowed. The next case is $3 \leqslant x \leqslant 5$, which gives $5 - x + x - 3 \geqslant x$ and thus $2 \geqslant x$, which cannot happen (when $3 \leqslant x \leqslant 5$). Finally, $x \geqslant 5$ implies $x - 5 + x - 3 \geqslant x$ and $x \geqslant 8$, which rules out $5 \leqslant x < 8$. In short, the solutions to the above equation are the whole real line except the three intervals $(8/3, 3]$, $[3, 5]$, and $[5, 8)$. This is really the whole real line except the one interval $(8/3, 8)$. In other words, the solutions are the points in $(-\infty, 8/3] \cup [8, \infty)$.

We have the following distributive relation.

Lemma 2.6. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *and* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof. First, we show that $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$. Indeed, if $x \in A \cup (B \cap C)$, then either x is in A or it is in both B and C. Either way, x is in $A \cup B$ and in $A \cup C$.

Next, we show that $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$. Here too, if x is in $A \cup B$ and in $A \cup C$, then either $x \in A$, or x is not in A and hence it is in both B and C. Either way, it is in $A \cup (B \cap C)$.

To prove the second equality either proceed similarly to the above proof, or apply the first equality to A^c, B^c, and C^c, and take complements of both side to get

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

□

Recall that we say a set I is countable if there is a bijective (1-to-1 and onto) function from $\mathbb{N} = \{1, 2, 3, \dots\}$ onto I; in other words if we can "count" I. Examples of countable sets are $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}^2$, $\mathbb{Z}^3, \dots$, and $\mathbb{Q}$. An example of an uncountable set is the interval $[0, 1)$, or any nonempty interval for that matter.

One can form countable unions of sets by defining $\cup_{i \geqslant 1} A_i$ to be the set of elements that are in at least one of the sets A_i. Similarly, $\cap_{i \geqslant 1} A_i$ is the set of elements that are in all of the A_i's simultaneously. (If there are no such elements, then the intersection is simply the empty set.)

Example 2.7. Let $a < b - 1$. Then, $\cup_{n \geqslant 1}(a, b - 1/n) = (a, b)$. It is clear that $(a, b - 1/n) \subset (a, b)$, for all $n \geqslant 1$. Thus, $\cup_{n \geqslant 1}(a, b - 1/n) \subseteq (a, b)$. On the other hand, if $x \in (a, b)$, then there exists an $n \geqslant 1$ such that $x < b - 1/n$. For otherwise, $x \geqslant b - 1/n$ for all $n \geqslant 1$ and thus taking $n \to \infty$ we have $x \geqslant b$. We just proved that if $x \in (a, b)$, then there is an $n \geqslant 1$ such that $x \in (a, b - 1/n)$. Thus, $(a, b) \subseteq \cup_{n \geqslant 1}(a, b - 1/n)$.

Example 2.8. Similarly to the above we can show that if $a < b$, then $\cap_{n \geqslant 1}(a, b + 1/n) = (a, b]$. In particular, $\cap_{n \geqslant 1}(0, 1/n) = \varnothing$. (Even though this is a sequence of nonempty decreasing sets, their intersection is empty!)

Homework Problems

Exercise 2.1. A public opinion poll (fictional) consists of the following three questions:

(1) Are you a registered Democrat ?

(2) Do you approve of President Obama's performance in office ?

(3) Do you favor the Health Care Bill ?

A group of 1000 people is polled. Answers to the questions are either *yes* or *no*. It is found that 550 people answer yes to the third question and 450 answer no. 325 people answer yes exactly twice (i.e. their answers contain 2 yeses and one no). 100 people answer yes to all three questions. 125 registered Democrats approve of Obama's performance. How many of those who favor the Health Care Bill do not approve of Obama's performance and in addition are not registered Democrats? (Hint: use a Venn diagram.)

Exercise 2.2. Let A and B be events in a sample space Ω. We remind that $A \setminus B = A \cap B^c$. Prove the following:

(a) $A \cap (B \setminus C) = (A \cap B) \setminus (A \cap C)$

(b) $A \setminus (B \cup C) = (A \setminus B) \setminus C$

(c) Is it true that $(A \setminus B) \cup C = (A \cup C) \setminus B$?

Exercise 2.3. Let Ω be the reals. Establish

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right] = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b\right)$$

$$[a, b] = \bigcap_{n=1}^{\infty} \left[a, b + \frac{1}{n}\right) = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right]$$

Lecture 3

1. Rules of probability, continued

Recall rules 1–3 (from the previous lecture). Rules 1–3 suffice if we want to study only finite sample spaces. But infinite sample spaces are also interesting. This happens, for example, if we want to write a model that answers, "what is the probability that we toss a coin 12 times before we toss heads?" This leads us to the next, and final, rule of probability.

Rule 4 (Extended addition rule)**.** If $A_1, A_2, \ldots$ are (countably-many) pairewise disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

This rule will be extremely important to us soon. It looks as if we might be able to derive this as a consequence of Lemma 1.4, but that is not the case ... it needs to be assumed as part of our model of probability theory.

2. Properties of probability

Rules 1–4 have other consequences as well.

Example 3.1. Let $A \subseteq B$. Note that A and $B \setminus A$ are disjoint. Because $B = A \cup (B \setminus A)$ is a disjoint union, Rule 3 implies then that

$$\begin{aligned} P(B) &= P(A \cup (B \setminus A)) \\ &= P(A) + P(B \setminus A). \end{aligned}$$

Thus, we obtain the statement that

$$A \subseteq B \Longrightarrow P(B \setminus A) = P(B) - P(A).$$

As a special case, taking $B = \Omega$ and using Rule 2, we have the physically–appealing statement that

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

For instance, this yields $P(\varnothing) = 1 - P(\Omega) = 0$. "Chances are zero that nothing happens."

Example 3.2. Since $P(B \setminus A) \geqslant 0$, the above also shows another physically-appealing property:

$$A \subseteq B \Longrightarrow P(A) \leqslant P(B).$$

Example 3.3. Suppose $\Omega = \{\omega_1, \ldots, \omega_N\}$ has N distinct elements ("N distinct outcomes of the experiment"). One way of assigning probabilities to every subset of Ω is to just let

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{N},$$

where $|E|$ denotes the number of elements of E. Let us check that this probability assignment satisfies Rules 1–4. Rules 1 and 2 are easy to verify, and Rule 4 holds vacuously because Ω does not have infinitely-many disjoint subsets. It remains to verify Rule 3. If A and B are disjoint subsets of Ω, then $|A \cup B| = |A| + |B|$. Rule 3 follows from this. In this example, each outcome ω_i has probability $1/N$. Thus, this is the special case of "equally likely outcomes."

The following generalizes Rule 3, because $P(A \cap B) = 0$ when A and B are disjoint.

Lemma 3.4 (Another addition rule)**.** *If A and B are events (not necessarily disjoint), then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{3.1}$$

Proof. We can write $A \cup B$ as a disjoint union of three events:

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

By Rule 3,

$$P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B). \tag{3.2}$$

Similarly, write $A = (A \cap B^c) \cup (A \cap B)$, as a disjoint union, to find that

$$P(A) = P(A \cap B^c) + P(A \cap B). \tag{3.3}$$

There is a third identity that is proved the same way. Namely,

$$P(B) = P(A^c \cap B) + P(A \cap B). \tag{3.4}$$

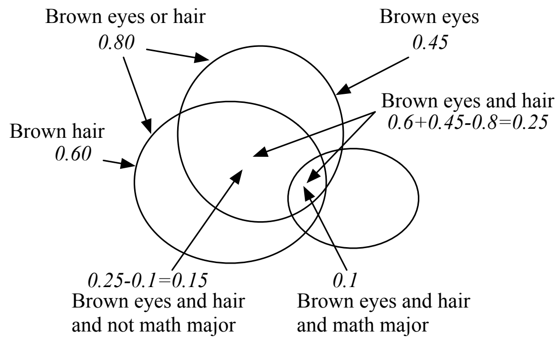


Figure 3.1. Venn diagram for Example 3.6.

Add (3.3) and (3.4) and solve to find that

$$P(A \cap B^c) + P(A^c \cap B) = P(A) + P(B) - 2P(A \cap B).$$

Plug this in to the right-hand side of (3.2) to finish the proof. □

As a corollary we have the following useful fact.

Lemma 3.5 (Boole's inequality). *If A_i, $i \geqslant 1$, are [countably many] events (not necessarily disjoint), then*

$$P(\cup_{i \geqslant 1} A_i) \leqslant \sum_{i \geqslant 1} P(A_i).$$

For finitely many such events, $A_1, \cdots, A_n$, the proof of the lemma goes by induction using the previous lemma. The proof of the general case of infinitely many events uses rule 4 and is omitted.

Example 3.6. The probability a student has brown hair is 0.6, the probability a student has brown eyes is 0.45, the probability a student has brown hair and eyes and is a math major is 0.1, and the probability a student has brown eyes or brown hair is 0.8. What is the probability of a student having brown eyes and hair, but not being a math major? We know that

$$\begin{aligned} &P\{\text{brown eyes or hair}\} \\ &\quad = P\{\text{brown eyes}\} + P\{\text{brown hair}\} - P\{\text{brown eyes and hair}\}. \end{aligned}$$

Thus, the probability of having brown eyes and hair is $0.45 + 0.6 - 0.8 = 0.25$. But then,

$$\begin{aligned} P\{\text{brown eyes and hair}\} &= P\{\text{brown eyes and hair and math major}\} \\ &\quad + P\{\text{brown eyes and hair and not math major}\}. \end{aligned}$$

Therefore, the probability we are seeking equals $0.25 - 0.1 = 0.15$. See Figure 3.1.

Formula 3.1 has a generalization. The following is called the "inclusion-exclusion" rule.

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{\substack{1 \leqslant j_1,\dots,j_i \leqslant n \\ j_1,\dots,j_i \text{ all different}}} P(A_{j_1} \cap \cdots \cap A_{j_i}).$$

For example,

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \end{aligned} \tag{3.5}$$

Proving the inclusion-exclusion formula is deferred to Exercise 33.2.

3. Topics of special interest: About the set of events $\mathscr{F}$

You can think of $\mathscr{F}$ as the set of events which you are allowed to ask about the probability of. We can be very general and choose it to be the set of all subsets of Ω, allowing ourselves to ask about anything. This turns out to be OK if the space is finite [or even if it is countably infinite]. However, this turns out to be too much to ask for if the space is, say, $\Omega = [0, 1]$. (We will see why shortly.)

In any case, the empty set must belong to $\mathscr{F}$; i.e. we should be able to ask about the probability that nothing happens. Here is another obvious property any choice of $\mathscr{F}$ must satisfy:

$$\text{if } A, B \in \mathscr{F}, \text{ then } A^c \in \mathscr{F} \text{ and } A \cup B \in \mathscr{F}.$$

In other words, if we can ask about the probability an event occurs, we should be able to ask about the probability it does not occur. Furthermore, if we can ask about the probabilities of two events, we should be able to ask about the probability at least one of them occurs. (Note that $A \cap B = (A^c \cup B^c)^c$ is then also in $\mathscr{F}$.)

Example 3.7. We can take $\mathscr{F} = \{\varnothing, \Omega\}$. This is the smallest possible $\mathscr{F}$. In this case, we are only allowed to ask about the probability of something happening and that of nothing happening.

Example 3.8. If A is a subset of Ω, we can take $\mathscr{F} = \{\varnothing, A, A^c, \Omega\}$. This is the smallest possible $\mathscr{F}$ containing A. In this case, we are only allowed to ask about the probability of something happening and that of nothing happening, as well as about the probabilies of A occurring or not.

Figure 3.2. Félix Édouard Justin Émile Borel (Jan 7, 1871 – Feb 3, 1956, France)

Example 3.9. If A and B are subsets of Ω, the smallest $\mathscr{F}$ containing them is

$$\{\varnothing, A, A^c, B, B^c, A\cap B, A\cap B^c, A^c\cap B, A^c\cap B^c, A\cup B, A\cup B^c,\\ A^c\cup B, A^c\cup B^c, (A\cap B^c)\cup(A^c\cap B), (A^c\cup B)\cap(A\cup B^c), \Omega\}.$$

Finally, if we have an infinite space, then $\mathscr{F}$ should also satisfy:

if A_i, $i \geqslant 1$, are in $\mathscr{F}$, then so is $\cup_{i\geqslant 1} A_i$.

(Note that, since $A_i^c \in \mathscr{F}$, the above automatically implies that $\cap_{i\geqslant 1} A_i = (\cup_{i\geqslant 1} A_i^c)^c \in \mathscr{F}$. So we do not need an extra "requirement".)

Example 3.10. It turns out that if $\Omega = [0,1]$, then there is a "smallest" $\mathscr{F}$ that satisfies the above requirements and contains all the intervals (a,b), $0 < a < b < 1$. This set turns out to be much smaller than the set of all subsets of $[0,1]$. It is called the Borel sigma-algebra and cannot be described in a simpler way than just "the smallest $\mathscr{F}$ that contains all the intervals"! Proving its existence requires quite a bit of mathematical analysis, which we do not go into in this class. It is noteworthy that one can also prove that any set in $\mathscr{F}$ can be written as intersections and unions of (countably many) intervals and complements of intervals of the form $(a,b]$.

3.1. Why $\mathscr{F}$ is not "everything". Now we can learn why the set of events $\mathscr{F}$ cannot be taken as the set of all subsets of Ω, if Ω is not countable; e.g. if $\Omega = [0,1]$. The reason is simply because then there are too many sets that one has to take into account and it is not clear if there is even one probability measure that can satisfy rules 1-4. If one instead uses the smallest $\mathscr{F}$ that contains the "sets of interest", e.g. the intervals, then one can prove that there are lots of probability measures.

In what follows, $\mathscr{F}$ will be in the background. We will not need it explicitly in this course. Any events we ask about the probability of will happen to be in this mysterious $\mathscr{F}$. Thus, we will not talk about it anymore (even though it is essential when doing more serious work in probability theory).

Homework Problems

Exercise 3.1. We toss a coin twice. We consider three steps in this experiment: 1. before the first toss; 2. after the first toss, but before the second toss; 3. after the two tosses.

(a) Give a sample space Ω for this experiment.

(b) Give the collection $\mathcal{F}_3$ of observable events at step 3.

(c) Give the collection $\mathcal{F}_2$ of observable events at step 2.

(d) Give the collection $\mathcal{F}_1$ of observable events at step 1.

Exercise 3.2. Aaron and Bill toss a coin one after the other until one of them gets a *head*. Aaron starts and the first one to get a head wins.

(a) Give a sample space for this experiment.

(b) Describe the events that correspond to "Aaron wins", "Bill wins" and "no one wins" ?

Exercise 3.3. Give an example to show that $P(A \setminus B)$ does not need to equal $P(A) - P(B)$.

Lecture 4

1. Modeling the experiment: Constructing state spaces

At this stage, we would like to construct probability measures for random experiments that satisfy Rules 1-4. The only example (so far) that we could verify these properties is is example 3.3, when all elements of the state space have equal mass.

Example 4.1. Let

$$\Omega = \Big\{ (H_1 , H_2) , (H_1 , T_2) , (T_1 , H_2) , (T_1 , T_2) \Big\}.$$

There are four possible outcomes. Suppose that they are equally likely. Then, by Rule 3,

$$\begin{aligned} P(\{H_1\}) &= P\Big(\{H_1 , H_2\} \cup \{H_1 , T_2\}\Big) \\ &= P(\{H_1 , H_2\}) + P(\{H_1 , T_2\}) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

In fact, in this model for equally-likely outcomes, $P(\{H_1\}) = P(\{H_2\}) = P(\{T_1\}) = P(\{T_2\}) = 1/2$. Thus, we are modeling two fair tosses of two fair coins.

Example 4.2. Let us continue with the sample space of the previous example, but assign probabilities differently. Here, we define $P(\{H_1 , H_2\}) = P(\{T_1 , T_2\}) = 1/2$ and $P(\{H_1 , T_2\}) = P(\{T_1 , H_2\}) = 0$. We compute, as we did before, to find that $P(\{H_1\}) = P(\{H_2\}) = P(\{T_1\}) = P(\{T_2\}) = 1/2$. But now the coins are not tossed fairly. In fact, the results of the two coin tosses are the same in this model; i.e. the first coin is a fair coin and once it is tossed and the result is known the second coin is simply flipped to match the result of the first coin. Thus, each of the two coins seems fair, but the second toss depends on the first one and is not hence a fair toss.

Example 4.3 (Word of caution)**.** One has to be careful when working out the state space. Consider, for example, tossing two identical fair coins and asking about the probability of the two coins landing with different faces; i.e. one heads and one tails. Since the two coins are identical and one cannot tell which is which, the state space can be taken as

$$\Omega = \{\text{“two heads”}, \text{“two tails”}, \text{“one heads and one tails”}\}.$$

A common mistake, however, is to assume these outcomes to be equally likely. To resolve the issue, let us paint one coin in red. Then, we can tell which coin is which and a natural state space is

$$\Omega = \{(H_1, H_2), (T_1, T_2), (H_1, T_2), (T_1, H_2)\}.$$

Now, these outcomes are equally likely. Since coins do not behave differently when they are painted, the probabilities assigned to the state space in the previous case of identical coins must be

$$P\{\text{two heads}\} = P\{\text{two tails}\} = 1/4 \text{ and } P\{\text{one heads and one tails}\} = 1/2.$$

We have seen already that $P(A) = |A|/|\Omega|$ for any event A when all outcomes of Ω are equally likely. Therefore, the first question we must address is, "how many items are in Ω?" In particular, we need to be able to (very carefully) count cardinalities of sets. we begin with two basic principles of counting that make sense intuitively.

Proposition 4.4 (The second principle of counting)**.** *If we have m distinct roads from village A to village B and n distinct roads from village B to village C, then we have mn ways to go from village A to village C.*

...not to be mistaken with ...

Proposition 4.5 (The first principle of counting)**.** *If we have m distinct roads from village A to village B and n distinct roads from village B to village C then we have $m + n$ distinct roads in the area.*

Example 4.6. Dice rolls

Roll two fair dice fairly; all possible outcomes are equally likely.

A good sample space is

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & \cdots & (1,6) \\ \vdots & \vdots & \ddots & \vdots \\ (6,1) & (6,2) & \cdots & (6,6) \end{Bmatrix}$$

We can think of Ω as a 6-by-6 table; so $|\Omega| = 6 \times 6 = 36$, by second-grade arithmetic.

1.1. What is the probability that we roll doubles? Let

$$A = \{(1,1), (2,2), \ldots, (6,6)\}.$$

We are asking to find $P(A) = |A|/36$. But there are 6 items in A; hence, $P(A) = 6/36 = 1/6$.

1.2. What are the chances that we roll a total of five pips? Let

$$A = \{(1,4), (2,3), (3,2), (4,1)\}.$$

We need to find $P(A) = |A|/36 = 4/36 = 1/9$.

1.3. What is the probability that we roll somewhere between two and five pips (inclusive)? Let

$$A = \left\{ \overbrace{(1,1)}^{\text{sum = 2}}, \underbrace{(1,2), (2,1)}_{\text{sum =3}}, \overbrace{(1,3), (2,2), (3,1)}^{\text{sum =4}}, \underbrace{(1,4), (4,1), (2,3), (3,2)}_{\text{sum=5}} \right\}.$$

We are asking to find $P(A) = 10/36$.

1.4. What are the odds that the product of the number of pips thus rolls is an odd number? The event in question is

$$A := \left\{ \begin{matrix} (1,1), & (1,3), & (1,5) \\ (3,1), & (3,3), & (3,5) \\ (5,1), & (5,3), & (5,5) \end{matrix} \right\}.$$

And $P(A) = 9/36 = 1/4$.

Example 4.7 (Easy cards)**.** There are 52 cards in a deck. You deal two cards, all pairs equally likely.

Math model: Ω is the collection of all pairs [drawn without replacement from an ordinary deck]. What is $|\Omega|$? To answer this note that $2|\Omega|$ is the number of all possible ways to give a pair out; i.e., $2|\Omega| = 52 \times 51$, by the principle of counting. Therefore,

$$|\Omega| = \frac{52 \times 51}{2} = 1326.$$

- The probability that exactly one card is an ace is $4 \times 48 = 192$ divided by 1326. This probability is $\simeq 0.1448$
- The probability that both cards are aces is $(4 \times 3)/2 = 6$ divided by 1326, which is $\simeq 0.0045$.
- The probability that both cards are the same is $P\{\text{ace and ace}\} + \cdots + P\{\text{king and king}\} = 13 \times 6/1326 \simeq 0.0588$.

Homework Problems

Exercise 4.1. A fair die is rolled 5 times and the sequence of scores recorded.

(a) How many outcomes are there?
(b) Find the probability that first and last rolls are 6.

Exercise 4.2. If a 3-digit number (000 to 999) is chosen at random, find the probability that exactly one digit will be larger than 5.

Exercise 4.3. A license plate is made of 3 numbers followed by 3 letters.

(a) What is the total number of possible license plates ?
(b) What is the number of license plates that start with an A ?

Exercise 4.4. An urn contains 3 red, 8 yellow and 13 green balls; another urn contains 5 red, 7 yellow and 6 green balls. We pick one ball from each urn at random. Find the probability that both balls are of the same color.

Lecture 5

1. The birthday problem

n people in a room; all birthdays are equally likely, and assigned at random. What are the chances that no two people in the room are born on the same day? You may assume that there are 365 days a years, and that there are no leap years.

Let $p(n)$ denote the probability in question.

To understand this consider finding $p(2)$ first. There are two people in the room.

The sample space is the collection of all pairs of the form (D_1, D_2), where D_1 and D_2 are birthdays. Note that $|\Omega| = 365^2$ [principle of counting].

In general, Ω is the collection of all "n-tuples" of the form $(D_1, \ldots, D_n)$ where the D_i's are birthdays; $|\Omega| = 365^n$. Let A denote the collection of all elements $(D_1, \ldots, D_n)$ of Ω such that all the D_i's are distinct. We need to find $|A|$.

To understand what is going on, we start with $n = 2$. In order to list all the elements of A, we observe that we have to assign two separate birthdays. [Forks = first birthday; knives = second birthday]. There are therefore 365×364 outcomes in A when $n = 2$. Similarly, when $n = 3$, there are $365 \times 364 \times 363$, and in general, $|A| = 365 \times \cdots \times (365 - n + 1)$. **Check this with induction!**

Thus,

$$p(n) = \frac{|A|}{|\Omega|} = \frac{365 \times \cdots \times (365 - n + 1)}{365^n}.$$

For example, check that $p(10) \simeq 0.88$ while $p(50) \simeq 0.03$. In fact, if $n \geqslant 23$, then $p(23) < 0.5$.

2. An urn problem

n purple and n orange balls are in an urn. You select two balls at random [without replacement]. What are the chances that they have different colors?

Here, Ω denotes the collection of all pairs of colors. Note that $|\Omega| = 2n(2n-1)$ [principle of counting].

$$P\{\text{two different colors}\} = 1 - P\{\text{the same color}\}.$$

Also,

$$P\{\text{the same color}\} = P(P_1 \cap P_2) + P(O_1 \cap O_2),$$

where O_j denotes the event that the jth ball is orange, and P_k the event that the kth ball is purple. The number of elements of $P_1 \cap P_2$ is $n(n-1)$; the same holds for $O_1 \cap O_2$. Therefore,

$$\begin{aligned} P\{\text{different colors}\} &= 1 - \left[\frac{n(n-1)}{2n(2n-1)} + \frac{n(n-1)}{2n(2n-1)}\right] \\ &= \frac{n}{2n-1}. \end{aligned}$$

In particular, regardless of the value of n, we always have

$$P\{\text{different colors}\} > \frac{1}{2}.$$

3. Combinatorics

Recall the two basic principles of counting [combinatorics]:

First principle: m distinct garden forks plus n distinct fish forks equals $m+n$ distinct forks.

Second principle: m distinct knives and n distinct forks equals mn distinct ways of taking a knife and a fork.

4. Ordered selection with replacement

Theorem 5.1. *Let $n \geqslant 1$ and $k \geqslant 0$ be integers. There are n^k ways to pick k balls from a bag containing n distinct (numbered 1 through n) balls, replacing the ball each time back in the bag.*

Proof. To prove this think of the case $k = 2$. Let B be the set of balls. Then, $B^2 = B \times B$ is the state space corresponding to picking two balls with replacement. The second principle of counting says $|B^2| = |B|^2 = n^2$. More generally, when picking k balls we have $|B^k| = |B|^k = n^k$ ways. $\square$

Note that the above theorem implies that the number of functions from a set A to a set B is $|B|^{|A|}$. (Think of $A = \{1, \dots, k\}$ and B being the set of balls. Each function from A to B corresponds to exactly one way of picking k balls from B, and vice-versa.)

Example 5.2. What is the probability that 10 persons, picked at random, are all born in May? Let us assume the year has 365 days and ignore leap years. There are 31 days in May and thus 31^{10} ways to pick 10 birthdays in May. In total, there are 365^{10} ways to pick 10 days. Thus, the probability in question is $\frac{31^{10}}{365^{10}}$.

Example 5.3. A PIN number is a four-symbol code word in which each entry is either a letter (A-Z) or a digit (0-9). Let A be the event that exactly one symbol is a letter. What is $P(A)$ if a PIN is chosen at random and all outcomes are equally likely? To get an outcome in A, one has to choose which symbol was the letter (4 ways), then choose that letter (26 ways), then choose the other three digits ($10 \times 10 \times 10$ ways). Thus,

$$P(A) = \frac{4 \times 26 \times 10 \times 10 \times 10}{36 \times 36 \times 36 \times 36} \simeq 0.0619.$$

Example 5.4. An experiment consists of rolling a fair die, drawing a card from a standard deck, and tossing a coin. Then, the probability that the die score is even, the card is a heart, and the coin is heads is equal to $\frac{3 \times 13 \times 1}{6 \times 52 \times 2} = 1/16$.

Example 5.5. We roll a fair die then toss a coin the number of times shown on the die. What is the probability of the event A that all coin tosses result in heads? One could use the state space

$$\Omega = \{(1, H), (1, T), (2, H, H), (2, T, T), (2, T, H), (2, H, T), \cdots\}.$$

However, the outcomes are then not all equally likely. Instead, we continue tossing the coin up to 6 times regardless of the outcome of the die. Now, the state space is $\Omega = \{1, \cdots, 6\} \times \{H, T\}^6$ and the outcomes are equally likely. Then, the event of interest is $A = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6$, where

A_i is the event that the die came up i and the first i tosses of the coin came up heads. There is one way the die can come up i and 2^{6-i} ways the first i tosses come up heads. Then,

$$P(A_i) = \frac{2^{6-i}}{6 \times 2^6} = \frac{1}{6 \times 2^i}.$$

These events are clearly disjoint and

$$P(A) = \frac{1}{6}\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64}\right) = \frac{21}{128}.$$

Homework Problems

Exercise 5.1. Suppose that there are 5 duck hunters, each a perfect shot. A flock of 10 ducks fly over, and each hunter selects one duck at random and shoots. Find the probability that 5 ducks are killed.

Exercise 5.2. Suppose that 8 rooks are randomly placed on a chessboard. Show that the probability that no rook can capture another is $8!8!/(64 \times 63 \times \cdots \times 57)$.

Exercise 5.3. A conference room contains m men and w women. These people seat at random in $m + w$ seats arranged in a row. Find the probability that all the women will be adjacent.

Exercise 5.4. If a box contains 75 good light bulbs and 25 defective bulbs and 15 bulbs are removed, find the probability that at least one will be defective.

Exercise 5.5. A lottery is played as follows: the player picks six numbers out of $\{1,2,\ldots,54\}$. Then, six numbers are drawn at random out of the 54. You win the first price of you have the 6 correct numbers and the second prize if you get 5 of them.

(a) What is the probability to win the first prize ?

(b) What is the probability to win the second prize ?

Exercise 5.6. Another lottery is played as follows: the player picks five numbers out of $\{1,2,\ldots,50\}$ and two other numbers from the list $\{1,\ldots,9\}$. Then, five numbers are drawn at random from the first list and two from the random list.

(a) You win the first prize if all numbers are correct. What is the probability to win the first prize ?

(b) Which lottery would you choose to play between this one and the one from the previous problem ?

Exercise 5.7. Find the probability that a five-card poker hand (i.e. 5 out of a 52-card deck) will be :

(a) *Four of a kind*, that is four cards of the same value and one other card of a different value (xxxxy shape).

(b) *Three of a kind*, that is three cards of the same value and two other cards of different values (xxxyz shape).

(c) A *straight flush*, that is five cards in a row, of the same suit (ace may be high or low).

(d) A *flush*, that is five cards of the same suit, but not a straight flush.

(e) A *straight*, that is five cards in a row, but not a straight flush (ace may be high or low).

Lecture 6

1. Ordered selection without replacement: Permutations

The following follows directly from the second principle of counting.

Theorem 6.1. *Let $1 \leqslant k \leqslant n$ be integers. There are $n(n-1)\cdots(n-k+1)$ ways to pick k balls out of a bag of n distinct (numbered 1 through n) balls, without replacing the balls back in the bag.*

As a special case one concludes that there are $n(n-1)\cdots(2)(1)$ ways to put n objects in order. (This corresponds to picking n balls out of a bag of n balls, without replacement.)

Definition 6.2. If $n \geqslant 1$ is an integer, then we define "n factorial" as the following integer:

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

For consistency of future formulas, we define also

$$0! = 1.$$

Note that the number in the above theorem can be written as

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

Example 6.3. 6 dice are rolled. What is the probability that they all show different faces?

$\Omega =?$

$|\Omega| = 6^6$.

If A is the event in question, then $|A| = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

Example 6.4. Five rolls of a fair die. What is $P(A)$, where A is the event that all five show different faces? Note that $|A|$ is equal to 6 [which face is left out] times 5!. Thus,

$$P(A) = \frac{6 \cdot 5!}{6^5} = \frac{6!}{6^5}.$$

Example 6.5. The number of permutations of cards in a regular 52-card deck is $52! > 8 \times 10^{68}$. If each person on earth shuffles a deck per second and even if each of the new shuffled decks gives a completely new permutation, it would still require more than 3×10^{50} years to see all possible decks! The currently accepted theory says Earth is no more than 5×10^9 years old and our Sun will collapse in about 7×10^9 years. The Heat Death theory places 3×10^{50} years from now in the Black Hole era. The matter that stars and life was built of no longer exists.

Example 6.6. Eight persons, consisting of four couples are to be seated in a row of eight chairs. What is the probability that significant others in each couple sit together? Since we have 4 couples, there are 4! ways to arrange them. Then, there are 2 ways to arrange each couple. Thus, there are $4! \times 2^4$ ways to seat couples together. The probability is thus $\frac{4! \times 2^4}{8!} = 1/105$.

2. Unordered selection without replacement: Combinations

Theorem 6.7. *The number of ways to choose k balls from a bag of n identical (unnumbered) balls is "n choose k." Its numerical value is*

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

More generally, let $k_1, \ldots, k_r \geqslant 0$ be integers such that $k_1 + \cdots + k_r = n$. Then, the number of ways we can choose k_1 balls, mark them 1, k_2 balls, mark them 2, ... , k_r balls, mark them r, out of a bag of n identical balls, is equal to

$$\binom{n}{k_1, \ldots, k_r} = \frac{n!}{k_1! \cdots k_r!}.$$

Before we give the proof, let us do an example that may shed a bit of light on the situation.

Example 6.8. If there are n people in a room, then they can shake hands in $\binom{n}{2}$ many different ways. Indeed, the number of possible hand shakes is the same as the number of ways we can list all pairs of people, which is clearly $\binom{n}{2}$. Here is another, equivalent, interpretation. If there are n vertices in a "graph," then there are $\binom{n}{2}$ many different possible "edges" that can be formed between distinct vertices. The reasoning is the same. Another way to reason is to say that there are n ways to pick the first vertex of the edge and $n-1$ way to pick the second one. But then we would count each edge twice (once from the point of view of each end of the edge) and thus the number of edges is $n(n-1)/2 = \binom{n}{2}$.

Example 6.9. Roll 4 dice; let A denote the event that all faces are different. Then,

$$|A| = \binom{6}{4} 4! = \frac{6!}{2!} = \frac{6!}{2}.$$

The 6-choose-4 is there because that is how many ways we can choose the different faces. Note that another way to count is via permutations. We are choosing 4 distinct faces out of 6. In any case,

$$P(A) = \frac{6!}{2 \times 6^4}.$$

Example 6.10. A poker hand consists of 5 cards dealt without replacement and without regard to order from a standard 52-cards deck. There are

$$\binom{52}{5} = 2,598,960$$

different standard poker hands possible.

Example 6.11. The number of different "pairs" $\{a, a, b, c, d\}$ in a poker hand is

$$\underbrace{13}_{\text{choose the } a} \times \underbrace{\binom{4}{2}}_{\text{deal the two } a\text{'s}} \times \underbrace{\binom{12}{3}}_{\text{choose the } b,\, c, \text{ and } d} \times \underbrace{4^3}_{\text{deal } b, c, d}.$$

The last 4^3 corresponds to an ordered choice because once b, c, and d are chosen, they are distinct and the order in which the suites are assigned does matter. Also, it is a choice with replacement because in each case all 4 suites are possible.

From the above we conclude that

$$P(\text{pairs}) = \frac{13 \times \binom{4}{2} \times \binom{12}{3} \times 4^3}{\binom{52}{5}} \approx 0.42.$$

We also can compute this probability by imposing order. Then, the number of ways to get one pair is

$$\underbrace{13}_{\text{choose the } a} \times \underbrace{4 \times 3}_{\text{deal the two } a\text{'s}} \times \underbrace{\binom{5}{2}}_{\text{choose where the } a\text{'s go}} \times \underbrace{12 \times 11 \times 10}_{\text{choose the } b,\, c, \text{ and } d} \times \underbrace{4^3}_{\text{deal } b, c, d}.$$

Then

$$P(\text{pairs}) = \frac{13 \times 4 \times 3 \times \binom{5}{2} \times 12 \times 11 \times 10 \times 4^3}{52 \times 51 \times 50 \times 49 \times 48}.$$

Check this is exactly the same as the above answer.

Example 6.12. Let A denote the event that we get two pairs $[a, a, b, b, c]$ in a poker hand. Then,

$$|A| = \underbrace{\binom{13}{2}}_{\text{choose } a, b} \times \underbrace{\binom{4}{2}^2}_{\text{deal the } a, b} \times \underbrace{11}_{\text{choose } c} \times \underbrace{4}_{\text{deal } c}.$$

Therefore,

$$P(\text{two pairs}) = \frac{\binom{13}{2} \times \binom{4}{2}^2 \times 11 \times 4}{\binom{52}{5}} \approx 0.06.$$

Proof of Theorem 6.7. Let us first consider the case of n distinct balls. Then, there is no difference between, on the one hand, ordered choices of k_1 balls, k_2 balls, etc, and on the other hand, putting n balls in order. There are $n!$ ways to do so. Now, each choice of k_1 balls out of n identical balls corresponds to $k_1!$ possible choices of k_1 balls out of n distinct balls. Hence, if the number of ways of choosing k_1 balls, marking them 1, then k_2 balls, marking them 2, etc, out of n identical balls is N, we can write $k_1! \cdots k_r! N = n!$. Solve to finish. $\square$

Homework Problems

Exercise 6.1. Suppose that n people are to be seated at a round table. Show that there are $(n-1)!$ distinct seating arrangements. Hint: the mathematical significance of a round table is that there is no dedicated first chair.

Exercise 6.2. An experiment consists of drawing 10 cards from an ordinary 52-card deck.

(a) If the drawing is made *with* replacement, find the probability that no two cards have the same face value.

(b) If the drawing is made *without* replacament, find the probability that at least 9 cards will have the same suit.

Exercise 6.3. An urn contains 10 balls numbered from 1 to 10. We draw five balls from the urn, *without* replacement. Find the probability that the second largest number drawn is 8.

Exercise 6.4. Eight cards are drawn without replacement from an ordinary deck. Find the probability of obtaining exactly three aces or exactly three kings (or both).

Exercise 6.5. How many possible ways are there to seat 8 people (A,B,C,D,E,F,G and H) in a row, if:

(a) No restrictions are enforced;

(b) A and B want to be seated together;

(c) assuming there are four men and four women, men should be only seated between women and the other way around;

(d) assuming there are five men, they must be seated together;

(e) assuming these people are four married couples, each couple has to be seated together.

Exercise 6.6. John owns six discs: 3 of classical music, 2 of jazz and one of rock (all of them different). How many possible ways does John have if he wants to store these discs on a shelf, if:

(a) No restrictions are enforced;

(b) The classical discs and the jazz discs have to be stored together;

(c) The classical discs have to be stored together, but the jazz discs have to be separated.

Exercise 6.7. How many (not necessarily meaningful) *words* can you form by shuffling the letters of the following words: (a) bike; (b) paper; (c) letter; (d) minimum.

Lecture 7

1. Properties of combinations

Clearly, n choose 0 and n choose n are both equal to 1. The following is also clear from the definition.

Lemma 7.1. *For any integers* $0 \leqslant k \leqslant n$

$$\binom{n}{k} = \binom{n}{n-k}.$$

Recall that n choose k is the number of ways one can choose k elements out of a set of n elements. Thus, the above formula is obvious: choosing which k balls we remove from a bag is equivalent to choosing which $n-k$ balls we keep in the basket. This is called a *combinatorial proof*.

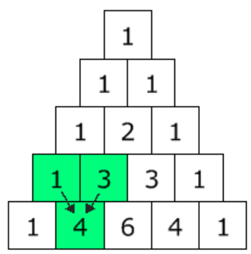
Lemma 7.2. *For* $1 \leqslant k \leqslant n-1$ *integers we have*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof. We leave the algebraic proof to the student and give instead the combinatorial proof. Consider a set of n identical balls and mark one of them, say with a different color. Any choice of k balls out of the n will either include or exclude the marked ball. There are $n-1$ choose k ways to choose k elements that exclude the ball and $n-1$ choose $k-1$ ways to choose k elements that include the ball. The formula now follows from the first principle of counting. $\square$

Figure 7.1. Blaise Pascal (Jun 19, 1623 – Aug 19, 1662, France)

This allows to easily generate the so-called Pascal's triangle [Chandas Shastra (5th?-2nd? century BC), Al-Karaji (953-1029), Omar Khayyám (1048-1131), Yang Hui (1238-1298), Petrus Apianus (1495-1552), Niccolò Fontana Tartaglia (1500-1577), Blaise Pascal (1653)]:



Example 7.3. How many subsets does $\{1, \ldots, n\}$ have? Assign to each element of $\{1, \ldots, n\}$ a zero ["not in the subset"] or a one ["in the subset"]. Thus, the number of subsets of a set with n distinct elements is 2^n.

Example 7.4. Choose and fix an integer $r \in \{1, \ldots, n\}$. The number of subsets of $\{1, \ldots, n\}$ that have size r is $\binom{n}{r}$. This, and the preceding proves the following amusing combinatorial identity:

$$\sum_{r=0}^{n} \binom{n}{r} = 2^n.$$

You may need to also recall the first principle of counting.

The preceding example has a powerful generalization.

Theorem 7.5 (The binomial theorem)**.** *For all integers $n \geqslant 0$ and all real numbers x and y,*

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

Remark 7.6. When $n = 2$, this yields the familiar algebraic identity

$$(x+y)^2 = x^2 + 2xy + y^2.$$

For $n = 3$ we obtain

$$\begin{aligned}(x+y)^3 &= \binom{3}{0}x^0y^3 + \binom{3}{1}x^1y^2 + \binom{3}{2}x^2y^1 + \binom{3}{3}x^3y^0 \\ &= y^3 + 3xy^2 + 3x^2y + x^3.\end{aligned}$$

Proof. This is obviously correct for $n = 0, 1, 2$. We use induction. Induction hypothesis: True for $n - 1$.

$$\begin{aligned}(x+y)^n &= (x+y)\cdot(x+y)^{n-1} \\ &= (x+y)\sum_{j=0}^{n-1}\binom{n-1}{j}x^j y^{n-j-1} \\ &= \sum_{j=0}^{n-1}\binom{n-1}{j}x^{j+1}y^{n-(j+1)} + \sum_{j=0}^{n-1}\binom{n-1}{j}x^j y^{n-j}.\end{aligned}$$

Change variables [$k = j + 1$ for the first sum, and $k = j$ for the second] to deduce that

$$\begin{aligned}(x+y)^n &= \sum_{k=1}^{n}\binom{n-1}{k-1}x^k y^{n-k} + \sum_{k=0}^{n-1}\binom{n-1}{k}x^k y^{n-k} \\ &= \sum_{k=1}^{n-1}\left\{\binom{n-1}{k-1} + \binom{n-1}{k}\right\}x^k y^{n-k} + x^n + y^n \\ &= \sum_{k=1}^{n-1}\binom{n}{k}x^k y^{n-k} + x^n + y^n.\end{aligned}$$

The binomial theorem follows. □

Remark 7.7. A combinatorial proof of the above theorem consists of writing

$$(x+y)^n = \underbrace{(x+y)(x+y)\cdots(x+y)}_{n\text{-times}}.$$

Then, one observes that to get the term $x^k y^{n-k}$ one has to choose k of the above n multiplicands and pick x from them, then pick y from the $n - k$ remaining multiplicands. There are n choose k ways to do that.

Example 7.8. The coefficient in front of x^3y^4 in $(2x - 4y)^7$ is $\binom{7}{3}2^3(-4)^4 = 71680$.

2. Conditional Probabilities

Example 7.9. There are 5 women and 10 men in a room. Three of the women and 9 of the men are employed. You select a person at random from the room, all people being equally likely to be chosen. Clearly, Ω is the collection of all 15 people, and

$$P\{\text{male}\} = \frac{2}{3}, \quad P\{\text{female}\} = \frac{1}{3}, \quad P\{\text{employed}\} = \frac{4}{5}.$$

Also,

$$P\{\text{male and employed}\} = \frac{9}{15}, \quad P\{\text{female and employed}\} = \frac{1}{5}.$$

Someone has looked at the result of the sample and tells us that the person sampled is employed. Let $P(\text{female} \,|\, \text{employed})$ denote the conditional probability of "female" given this piece of information. Then,

$$P(\text{female} \,|\, \text{employed}) = \frac{|\text{female among employed}|}{|\text{employed}|} = \frac{3}{12} = \frac{1}{4}.$$

Definition 7.10. If A and B are events and $P(B) > 0$, then the *conditional probability of* A *given* B is

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

For the previous example, this amounts to writing

$$P(\text{Female} \,|\, \text{employed}) = \frac{|\text{female and employed}|/|\Omega|}{|\text{employed}|/|\Omega|} = \frac{1}{4}.$$

The above definition is consistent with the frequentist intuition about probability. Indeed, if we run an experiment n times and observe that an event B occurred n_B times, then probabilistic intuition tells us that $P(B) \simeq n_B/n$. If among these n_B times an event A occurred n_{AB} times, then $P(A \,|\, B)$ should be about n_{AB}/n_B. Dividing through by n one recovers the above definition of conditional probability.

Example 7.11. If we deal two cards fairly from a standard deck, the probability of $K_1 \cap K_2$ [$K_j = \{\text{King on the j draw}\}$] is

$$P(K_1 \cap K_2) = P(K_1)P(K_2 \,|\, K_1) = \frac{4}{52} \times \frac{3}{51}.$$

This agrees with direct counting: $|K_1 \cap K_2| = 4 \times 3$, whereas $|\Omega| = 52 \times 51$.

Similarly,

$$\begin{aligned} P(K_1 \cap K_2 \cap K_3) &= P(K_1) \times \frac{P(K_1 \cap K_2)}{P(K_1)} \times \frac{P(K_3 \cap K_1 \cap K_2)}{P(K_1 \cap K_2)} \\ &= P(K_1)P(K_2 \mid K_1)P(K_3 \mid K_1 \cap K_2) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}. \end{aligned}$$

Or for that matter,

$$P(K_1 \cap K_2 \cap K_3 \cap K_4) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}. \text{ (Check!)}$$

3. Topics of special interest: Multinomial theorem and coefficient

One can similarly work out the coefficients in the multinomial theorem.

Theorem 7.12 (The multinomial theorem)**.** *For all integers $n \geqslant 0$ and $r \geqslant 2$, and all real numbers $x_1, \ldots, x_r$,*

$$(x_1 + \cdots + x_r)^n = \sum_{\substack{0 \leqslant k_1,\ldots,k_r \leqslant n \\ k_1+\cdots+k_r=n}} \binom{n}{k_1, \cdots, k_r} x_1^{k_1} \cdots x_r^{k_r},$$

where $\binom{n}{k_1,\cdots,k_r}$ was defined in Theorem 6.7.

The sum in the above display is over r-tuples $(k_1, \ldots, k_r)$ such that each k_i is an integer between 0 and n, and the k_i's add up to n. In the case $r = 2$, these are simply $(k, n-k)$ where k runs from 0 to n. So there are $n+1$ terms. The following theorem gives the number of terms for more general r.

Theorem 7.13. *The number of terms in the expansion of $(x_1 + \cdots + x_r)^n$ is $\binom{n+r-1}{r-1}$.*

For example, the number of terms in the expansion of $(a+b+c)^5$ is $\binom{5+3-1}{3-1} = \binom{7}{2} = 21$ terms.

Proof of Theorem 7.13. To prove the above theorem imagine we have a collection of n indistinguishable balls that we want to split among r friends. (Friend number 1 then gets k_1 balls, etc.) We want to compute the number of ways we can do this.

To split the balls among the friend, put the n balls in a row and insert $r-1$ indistinguishable stones in between the balls. This will break the balls into exactly r groups. The first group goes to friend number 1, and so on.

Now we see that the problem amounts to just putting $n+r-1$ objects (the balls and the stones) in a row, in any order. However, (as was done in the proof of Theorem 6.7) since the balls are indistinguishable, we need to divide by $n!$. Similarly, since all stones are indistinguishable, we need to divide by $(r-1)!$. Hence, the number of ways we can split n identical balls into r groups is

$$\frac{(n+r-1)!}{n!(r-1)!} = \binom{n+r-1}{r-1}. \qquad \square$$

Homework Problems

Exercise 7.1. Find the coefficient of x^5 in $(2+3x)^8$.

Exercise 7.2 (The game of rencontre)**.** An urn contains n tickets numbered $1, 2, \ldots, n$. The tickets are shuffled thoroughly and then drawn one by one without replacement. If the ticket numbered r appears in the r-th drawing, this is denoted as a *match* (French: *rencontre*). Show that the probability of at least one match is

$$1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!} \to 1 - e^{-1} \quad \text{as } n \to \infty.$$

Exercise 7.3. Show that

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0},$$

where $0 \leqslant r \leqslant \min(n, m)$, $r, m, n \in \mathbb{N}$. Try to find a combinatorial proof and an algebraic proof.

Exercise 7.4. (a) Prove the equality

$$\sum_{k=1}^{n} k \binom{n}{k} = n\, 2^{n-1}$$

by computing in two different ways the number of possible ways to form a team with a captain out of n people. (The size of the team can be anything.)

(b) Similarly as in (a), prove that

$$\sum_{k=1}^{n} k(k-1) \binom{n}{k} = n(n-1)\, 2^{n-2}.$$

(c) Find again the results of (a) and (b) by applying the binomial theorem to $(1+x)^n$ and taking derivatives with respect to x.

Exercise 7.5. We are interested in 4-digit numbers. (The number 0013 is a 2-digit number, not a 4-digit number.)

(a) How many of them have 4 identical digits?

(b) How many of them are made of two pairs of 2 identical digits?

(c) How many of them have 4 different digits?

(d) How many of them have 4 different digits, in increasing order (from left to right)?

(e) What are the answers to (a), (c) and (d) if we replace 4 by n?

Lecture 8

1. Conditional Probabilities, continued

Sometimes, to compute the probability of some event A, it turns out to be helpful if one knew something about another event B. This then can be used as follows.

Theorem 8.1 (Law of total probability). *For all events A and B,*

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

If, in addition, $0 < P(B) < 1$, then

$$P(A) = P(A \,|\, B)P(B) + P(A \,|\, B^c)P(B^c).$$

Proof. For the first statement, note that $A = (A \cap B) \cup (A \cap B^c)$ is a disjoint union. For the second, write $P(A \cap B) = P(A \,|\, B)P(B)$ and $P(A \cap B^c) = P(A \,|\, B^c)P(B^c)$. □

Example 8.2. Once again, we draw two cards from a standard deck. The probability $P(K_2)$ (second draw is a king, regardless of the first) is best computed by splitting it into the two disjoint cases: $K_1 \cap K_2$ and $K_1^c \cap K_2$. Thus,

$$\begin{aligned} P(K_2) &= P(K_2 \cap K_1) + P(K_2 \cap K_1^c) = P(K_1)P(K_2 \,|\, K_1) + P(K_1^c)P(K_2 \,|\, K_1^c) \\ &= \frac{4}{52} \times \frac{3}{51} + \frac{48}{52} \times \frac{4}{51}. \end{aligned}$$

In the above theorem what mattered was that B and B^c partitioned the space Ω into two disjoint parts. The same holds if we partition the space into any other number of disjoint parts (even countably many).

Figure 8.1. Thomas Bayes (1702 – Apr 17, 1761, England)

Example 8.3. There are three types of people: 10% are poor (π), 30% have middle-income (μ), and the rest are rich (ρ). 40% of all π, 45% of μ, and 60% of ρ are over 25 years old (Θ). Find $P(\Theta)$. The result of Theorem 8.1 gets replaced with

$$\begin{aligned} P(\Theta) &= P(\Theta \cap \pi) + P(\Theta \cap \mu) + P(\Theta \cap \rho) \\ &= P(\Theta \,|\, \pi)P(\pi) + P(\Theta \,|\, \mu)P(\mu) + P(\Theta \,|\, \rho)P(\rho) \\ &= 0.4P(\pi) + 0.45P(\mu) + 0.6P(\rho). \end{aligned}$$

We know that $P(\rho) = 0.6$ (why?), and thus

$$P(\Theta) = (0.4 \times 0.1) + (0.45 \times 0.3) + (0.6 \times 0.6) = 0.535.$$

Example 8.4. Let us recall the setting of Example 5.4. We can now use the state space

$$\{(D1, H_1), (D1, T_1), (D2, T_1, T_2), (D2, T_1, H_2), (D2, H_1, T_2), (D2, H_1, H_2), \cdots\},$$

even though we know the outcomes are not equally likely. We can then compute

$$\begin{aligned} P(A) &= P\{(D1, H_1)\} + P\{(D_2, H_1, H_2)\} + \cdots + P\{(D6, H_1, H_2, H_3, H_4, H_5, H_6)\} \\ &= P(D1)P(H_1 \,|\, D1) + P(D_2)P\{(H_1, H_2) \,|\, D_2\} + \cdots \\ &= P(D1)P(H_1 \,|\, D1) + P(D2)P(H_1 \,|\, D2)P(H_2 \,|\, D1, H_1) + \cdots . \end{aligned}$$

We will finish this computation once we learn about independence in the next lecture.

2. Bayes' Theorem

The following question arises from time to time: Suppose A and B are two events of positive probability. If we know $P(B \,|\, A)$ but want $P(A \,|\, B)$, then

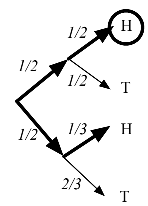


Figure 8.2. Boldface arrows indicated paths giving heads. The path going to the boldface circle corresponds to choosing the first coin and getting heads. Probabilities multiply along paths by Bayes' formula.

we can proceed as follows:

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

If we know only the conditional probabilities, then we can write $P(B)$, in turn, using Theorem 8.1, and obtain

Theorem 8.5 (Bayes's Rule)**.** *If* A, A^c *and* B *are events of positive probability, then*

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

Example 8.6. As before, deal two cards from a standard deck. Then, $P(K_1 \mid K_2)$ seems complicated to compute. But Bayes' rule says:

$$\begin{aligned} P(K_1 \mid K_2) &= \frac{P(K_1 \cap K_2)}{P(K_2)} = \frac{P(K_1)P(K_2 \mid K_1)}{P(K_1)P(K_2 \mid K_1) + P(K_1^c)P(K_2 \mid K_1^c)} \\ &= \frac{\frac{4}{52} \times \frac{3}{51}}{\frac{4}{52} \times \frac{3}{51} + \frac{48}{52} \times \frac{4}{51}}. \end{aligned}$$

Example 8.7. There are two coins on a table. The first tosses heads with probability 1/2, whereas the second tosses heads with probability 1/3. You select one at random (equally likely) and toss it. Say you got heads. What are the odds that it was the first coin that was chosen?

Let C denote the event that you selected the first coin. Let H denote the event that you tossed heads. We know: $P(C) = 1/2$, $P(H \mid C) = 1/2$, and $P(H \mid C^c) = 1/3$. By Bayes's formula (see Figure 8.2),

$$P(C \mid H) = \frac{P(H \mid C)P(C)}{P(H \mid C)P(C) + P(H \mid C^c)P(C^c)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{3} \times \frac{1}{2}\right)} = \frac{3}{5}.$$

Remark 8.8. The denominator in Bayes' rule simply computes $P(B)$ using the law of total probability. Sometimes, partitioning the space Ω into A and A^c is not the best way to go (e.g. when the event A^c is complicated).

In that case, one can apply the law of total probability by partitioning the space Ω into more than just two parts (as was done in Example 8.3 to compute the probability $P(\Theta)$). The corresponding diagram (analogous to Figure 8.2) could then have more than two branches out of each node. But the methodology is the same. See Exercise 8.10 for an example of this.

3. Conditional probabilities as probabilities

Suppose B is an event of positive probability. Consider the conditional probability distribution, $Q(\cdots) = P(\cdots \mid B)$.

Theorem 8.9. Q *is a probability on the new sample space* B*. [It is also a probability on the larger sample space* Ω*, why?]*

Proof. Rule 1 is easy to verify: For all events A,

$$0 \leqslant Q(A) = \frac{P(A \cap B)}{P(B)} \leqslant \frac{P(B)}{P(B)} = 1,$$

because $A \cap B \subseteq B$ and hence $P(A \cap B) \leqslant P(B)$.

For Rule 2 we check that

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

Next suppose $A_1, A_2, \ldots$ are disjoint events. Then,

$$Q\left(\bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)} P\left(\bigcup_{n=1}^{\infty} A_n \cap B\right).$$

Note that $\cup_{n=1}^{\infty} A_n \cap B = \cup_{n=1}^{\infty}(A_n \cap B)$, and $(A_1 \cap B), (A_2 \cap B), \ldots$ are disjoint events. Therefore,

$$Q\left(\bigcup_{n=1}^{\infty} A_n\right) = \frac{1}{P(B)} \sum_{n=1}^{\infty} P\left(A_n \cap B\right) = \sum_{n=1}^{\infty} Q(A_n).$$

This verifies Rule 4, and hence Rule 3. $\square$

Homework Problems

Exercise 8.1. In 10 Bernoulli trials find the conditional probability that all successes will occur consecutively (i.e., no two successes will be separated by one or more failures), given that the number of successes is between four and six.

Exercise 8.2. We toss a fair coin n times. What is the probability that we get *at least* 3 heads given that we get at least one.

Exercise 8.3. A fair die is rolled. If the outcome is odd, a fair coin is tossed repeatedly. If the outcome is even, a biased coin (with probability of heads $p \neq \frac{1}{2}$) is tossed repeatedly. If the first n throws result in heads, what is the probability that the fair coin is being used?

Exercise 8.4. We select a positive integer I with $P\{I = n\} = \frac{1}{2^n}$. If $I = n$, we toss a coin with probability of heads $p = e^{-n}$. What is the probability that the result is *heads*?

Exercise 8.5. A bridge player and his partner are known to have six spades between them. Find the probability that the spades will be split (a) 3-3, (b) 4-2 or 2-4, (c) 5-1 or 1-5, (d) 6-0 or 0-6.

Exercise 8.6. An urn contains 30 white and 15 black balls. If 10 balls are drawn with (respectively without) replacement, find the probability that the first two balls will be white, given that the sample contains exactly six white balls.

Exercise 8.7. In a certain village, 20% of the population has some disease. A test is administered which has the property that if a person is sick, the test will be positive 90% of the time and if the person is not sick, then the test will still be positive 30% of the time. All people tested positive are prescribed a drug which always cures the disease but produces a rash 25% of the time. Given that a random person has the rash, what is the probability that this person had the disease to start with?

Exercise 8.8. An insurance company considers that people can be split in two groups : those who are likely to have accidents and those who are not. Statistics show that a person who is likely to have an accident has probability 0.4 to have one over a year; this probability is only 0.2 for a person who is not likely to have an accident. We assume that 30% of the population is likely to have an accident.

(a) What is the probability that a new customer has an accident over the first year of his contract?

(b) A new customer has an accident during the first year of his contract. What is the probability that he belongs to the group likely to have an accident?

Exercise 8.9. A transmitting system transmits 0's and 1's. The probability of a correct transmission of a 0 is 0.8, and it is 0.9 for a 1. We know that 45% of the transmitted symbols are 0's.

(a) What is the probability that the receiver gets a 0?

(b) If the receiver gets a 0, what is the probability the the transmitting system actually sent a 0?

Exercise 8.10. 46% of the electors of a town consider themselves as independent, whereas 30% consider themselves democrats and 24% republicans. In a recent election, 35% of the independents, 62% of the democrats and 58% of the republicans voted.

(a) What proportion of the total population actually voted?

(b) A random voter is picked. Given that he voted, what is the probability that he is independent? democrat? republican?

Exercise 8.11. To go to the office, John sometimes drives - and he gets late once every other time - and sometimes takes the train - and he gets late only once every other four times. When he get on time, he always keeps the same transportation the day after, whereas he always changes when he gets late. Let p be the probability that John drives on the first day.

(a) What is the probability that John drives on the n^{th} day?

(b) What is the probability that John gets late on the n^{th} day?

(c) Find the limit as $n \to \infty$ of the results in (a) and (b).

Lecture 9

1. Independence

It is reasonable to say that A is *independent* of B if

$$P(A \mid B) = P(A),\ P(A^c \mid B) = P(A^c),\ P(A \mid B^c) = P(A),\ \text{and}\ P(A^c \mid B^c) = P(A^c);$$

i.e. "knowledge of B tells us nothing new about A." It turns out that the first equality above implies the other three. (Check!) It also is equivalent to the definition that we will actually use: A and B are independent if and only if

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

Note that this is now a symmetric formula and thus B is also independent of A. Note also that the last definition makes sense even if $P(B) = 0$ or $P(A) = 0$.

Example 9.1. In fact, if $P(A) = 0$ or $P(A) = 1$, then A is independent of any other event B. Indeed, if $P(A) = 0$ then $P(A \cap B) \leqslant P(A)$ implies that $P(A \cap B) = 0 = P(A)P(B)$. Also, if $P(A) = 1$, then $P(A^c) = 0$ and

$$P(B^c) \leqslant P(A^c \cup B^c) \leqslant P(A^c) + P(B^c)$$

implies that $P(A^c \cup B^c) = P(B^c)$ and thus $P(A \cap B) = P(B) = P(A)P(B)$.

Example 9.2. Conversely, if A is independent of any other event B (and so in particular A is independent of itself!), then it must be the case that $P(A)$ is 0 or 1. To see this observe that $P(A \cap A) = P(A)P(A)$.

It is noteworthy that being independent and being disjoint have nothing to do with each other.

Example 9.3. Roll a die and let A be the event of an even outcome and B that of an odd outcome. The two are obviously dependent. Mathematically, $P(A \cap B) = 0$ while $P(A) = P(B) = 1/2$. On the other hand the two are disjoint. Conversely, let C be the event of getting a number less than 2. Then, $P(A \cap C) = P\{2\} = 1/6$ and $P(A)P(C) = 1/2 \times 1/3 = 1/6$. So even though A and C are not disjoint, they are independent.

Two experiments are *independent* if A_1 and A_2 are independent for all choices of events A_j of experiment j.

Example 9.4. Toss two fair coins; all possible outcomes are equally likely. Let H_j denote the event that the jth coin landed on heads, and $T_j = H_j^c$. Then,

$$P(H_1 \cap H_2) = \frac{1}{4} = P(H_1)P(H_2).$$

In fact, the two coins are independent because $P(T_1 \cap T_2) = P(T_1 \cap H_2) = P(H_1 \cap H_2) = 1/4$ also. Conversely, if two fair coins are tossed independently, then all possible outcomes are equally likely to occur. What if the coins are not fair, say $P(H_1) = P(H_2) = 1/4$?

Similarly to the above reasoning, three events A_1, A_2, and A_3 are independent if any combination of two is independent of both the third and of its complement; e.g. A_1 and $A_2^c \cap A_3$ are independent as well as are A_2^c and $A_1 \cap A_3$ and so on. It turns out that all these relations follow simply from saying that any two of the events are independent and that also

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \tag{9.1}$$

For example, then

$$\begin{aligned} P(A_1 \cap A_2^c \cap A_3) &= P(A_1 \cap A_3) - P(A_1 \cap A_2 \cap A_3) \\ &= P(A_1)P(A_3) - P(A_1)P(A_2)P(A_3) \\ &= P(A_1)(1 - P(A_2))P(A_3) \\ &= P(A_1)P(A_2^c)P(A_3). \end{aligned}$$

Note that (9.1) by itself is not enough for independence. It is essential that on top of that every two events are independent.

Example 9.5. Roll two dice and let A be the event of getting a number less than 3 on the first die, B the event of getting 3, 4, or 5, on the first die, and C the event of the two faces adding up to 9. Then, $P(A \cap B \cap C) = 1/36 = P(A)P(B)P(C)$ but $P(A \cap B) = 1/6 \neq 1/4 = P(A)P(B)$.

Also, it could happen that any two are independent but (9.1) does not hold and hence A_1, A_2, and A_3 are not independent.

Example 9.6. Roll two dice and let A be the event of getting a number less than 3 on the first die, B the event of getting a number larger than 4 on the second die, and C the event of the two faces adding up to 7. Then, each two of these are independent (check), while

$$P(A \cap B \cap C) = P\{(1,6),(2,5),(3,4)\} = \frac{1}{12}$$

but $P(A)P(B)P(C) = 1/24$.

More generally, having defined independence of $n-1$ events, then $A_1, A_2, A_3, \ldots, A_n$ are independent if any $n-1$ of them are and

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)\cdots P(A_n).$$

n experiments are independent if $A_1, \cdots, A_n$ are, for any events A_j of experiment j.

Example 9.7. In 10 fair tosses of a coin that comes up heads with probability p, the conditional probability that all heads will occur consecutively, given that the number of heads is between four and six, is equal to the ratio of the probability of getting exactly four, five, or six consecutive heads (and the rest tails), by the probability of getting between four and six heads. That is,

$$\frac{7p^4(1-p)^6 + 6p^5(1-p)^5 + 5p^6(1-p)^4}{\binom{10}{4}p^4(1-p)^6 + \binom{10}{5}p^5(1-p)^5 + \binom{10}{6}p^6(1-p)^4}.$$

(7 ways to get 4 heads in a row and the rest tails, etc.)

Example 9.8. We can now finish the computation from Example 8.4 which gives an alternative solution to Example 5.4. Indeed, the die and the coins are independent. Hence,

$$\begin{aligned} P(A) &= P(D1)P(H_1 \mid D1) + P(D2)P(H_1 \mid D2)P(H_2 \mid D2, H_1) + \cdots \\ &= P(D1)P(H_1) + P(D2)P(H_1)P(H_2) + \cdots \\ &= \frac{1}{6}\Big(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^6}\Big). \end{aligned}$$

2. Gambler's ruin formula (Huygens)

You, the "Gambler," are playing independent repetitions of a fair game against the "House." When you win, you gain a dollar; when you lose, you lose a dollar. You start with k dollars, and the House starts with K dollars. What is the probability that the House is ruined before you?

Observe that if you reach $k+K$ dollars, the house is ruined, while if you reach 0 dollars, you are ruined. In either case the game ends. Let us think slightly more generally and define P_j to be the conditional probability that when the game ends you have $K + k$ dollars (i.e. you win), given that you start with j dollars initially. We want to find P_k.

Two easy cases are: $P_0 = 0$ and $P_{k+K} = 1$.

By direct use of the definitions we see that

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

[This is the conditional version of $P(A \cap B) = P(A \,|\, B)P(B)$.]

Let H be the event "the house is ruined", W be the event we win the next \$1 and L be the event we lose our next \$1. By Theorem 8.1 we then have for $j = 1, \ldots, k + K - 1$

$$\begin{aligned} P_j &= P(H \,|\, \text{start with \$j}) \\ &= P(H \cap W \,|\, \text{start with \$j}) + P(H \cap L \,|\, \text{start with \$j}) \\ &= P(W \,|\, \text{start with \$j}) P(H \,|\, W \text{ and start with \$j}) \\ &\qquad\qquad + P(L \,|\, \text{start with \$j}) P(H \,|\, L \text{ and start with \$j}). \end{aligned}$$

Since winning or losing \$1 is independent of how much we start with, and the probability of each is just 1/2, and since starting with \$j and winning \$1 results in us having $\$(j + 1)$, and similarly for losing \$1, we have

$$\begin{aligned} P_j &= \frac{1}{2} P(H \,|\, \text{start with \$}(j+1)) + \frac{1}{2} P(H \,|\, \text{start with \$}(j-1)) \\ &= \frac{1}{2} P_{j+1} + \frac{1}{2} P_{j-1}. \end{aligned}$$

In order to solve this, write $P_j = \frac{1}{2} P_j + \frac{1}{2} P_j$, so that

$$\frac{1}{2} P_j + \frac{1}{2} P_j = \frac{1}{2} P_{j+1} + \frac{1}{2} P_{j-1} \qquad \text{for } 0 < j < k + K.$$

Multiply both side by two and solve:

$$P_{j+1} - P_j = P_j - P_{j-1} \qquad \text{for } 0 < j < k + K.$$

In other words,

$$P_{j+1} - P_j = P_1 \qquad \text{for } 0 < j < k + K.$$

This is the simplest of all possible "difference equations." In order to solve it you note that, since $P_0 = 0$,

$$\begin{aligned} P_{j+1} &= (P_{j+1} - P_j) + (P_j - P_{j-1}) + \cdots + (P_1 - P_0) \qquad \text{for } 0 < j < k + K \\ &= (j + 1) P_1 \qquad \text{for } 0 < j < k + K. \end{aligned}$$

Apply this with $j = k + K - 1$ to find that

$$1 = P_{k+K} = (k + K) P_1, \quad \text{and hence} \quad P_1 = \frac{1}{k + K}.$$

Therefore,

$$P_{j+1} = \frac{j+1}{k+K} \qquad \text{for } 0 < j < k+K.$$

Set $j = k - 1$ to find the following:

Theorem 9.9 (Gambler's ruin formula)**.** *If you start with* k *dollars, then the probability that you end with* $k + K$ *dollars before losing all of your initial fortune is* $k/(k+K)$ *for all* $k, K \geqslant 1$.

Homework Problems

Read: section 1.7 of Ash's book carefully.

Exercise 9.1. A single card is drawn from a standard 52-card deck. Give examples of events A and B that are:

(a) Disjoint but not independent;

(b) Independent but not disjoint;

(c) Independent and disjoint;

(d) Neither independent nor disjoint.

Exercise 9.2. Of the 100 people in a certain village, 50 always tell the truth, 30 always lie, and 20 always refuse to answer. A sample of size 30 is taken with replacement.

(a) Find the probability that the sample will contain 10 people of each category.

(b) Find the probability that there will be exactly 12 liars.

Exercise 9.3. Six fair dice are rolled independently. Find the probability that the number of 1's minus the number of 2's is equal to 3.

Exercise 9.4. Prove the following statements.

(a) If an event A is independent of itself, then $P(A) = 0$ or 1.

(b) If $P(A) = 0$ or 1, then A is independent of any event B.

Exercise 9.5. We toss a fair coin three times. Let G_1 be the event "the second and third tosses give the same outcome", G_2 the event "tosses 1 and 3 give the same outcome" and G_3 the event "tosses 1 and 2 give the same outcome". Prove that these events are pairwise independent but not independent.

Exercise 9.6. We assume that the gender of a child is independent of the gender of the other children of the same couple and that the probability to get a boy is 0.52. Compute, for a 4-child family, the probabilities of the following events:

(a) all children have the same gender;

(b) the three oldest children are boys and the youngest is a girl;

(c) there are exactly 3 boys;

(d) the two oldest are boys;

(e) there is at least a girl.

Lecture 10

Figure 10.1. Christiaan Huygens (Apr 14, 1629 – Jul 8, 1695, Netherlands)

1. Random Variables

We often want to measure certain characteristics of an outcome of an experiment; e.g. we pick a student at random and measure their height. Assigning a value to each possible outcome is what a random variable does.

Definition 10.1. A D-valued random variable is a function X from Ω to D. The set D is usually [for us] a subset of the real line $\mathbf{R}$, or d-dimensional space $\mathbf{R}^d$.

We use capital letters (X, Y, Z, etc) for random variables.

Example 10.2. Define the sample space,

$$\Omega = \left\{\text{⚀},\text{⚁},\text{⚂},\text{⚃},\text{⚄},\text{⚅}\right\}.$$

Then, the random variable $X(\text{⚀}) = 1$, $X(\text{⚁}) = 2, \ldots, X(\text{⚅}) = 6$ models the number of pips in a roll of a fair six-sided die.

The random variable $Y(\text{⚀}) = Y(\text{⚂}) = 5$, $Y(\text{⚁}) = Y(\text{⚄}) = Y(\text{⚅}) = 2$, and $Y(\text{⚃}) = -1$ models the game where you roll a die and win \$5 if you get 1 or 3, win \$2 if you get 2, 5 or 6, and lose \$1 if you get 4.

Now if, say, we picked John and he was 6 feet tall, then there is nothing random about 6 feet! What is random is how we picked the student; i.e. the procedure that led to the 6 feet. Picking a different student is likely to lead to a different value for the height. This is modeled by giving a probability P on the state space Ω.

Example 10.3. In the previous example assume the die is fair; i.e. all outcomes are equally likely. This corresponds to the probability P on Ω that gives each outcome a probability of 1/6. As a result, for all $k = 1, \ldots, 6$,

$$P\left(\{\omega \in \Omega : X(\omega) = k\}\right) = P(\{k\}) = \frac{1}{6}. \tag{10.1}$$

This probability is zero for other values of k, since X does not take such values. Usually, we write $\{X \in A\}$ in place of the set $\{\omega \in \Omega : X(\omega) \in A\}$. In this notation, we have

$$P\{X = k\} = \begin{cases} \frac{1}{6} & \text{if } k = 1, \ldots, 6, \\ 0 & \text{otherwise.} \end{cases} \tag{10.2}$$

This is a math model for the result of rolling a fair die. Similarly,

$$P\{Y = 5\} = \frac{1}{3},\ P\{Y = 2\} = \frac{1}{2},\ \text{and } P\{Y = -1\} = \frac{1}{6}. \tag{10.3}$$

This is a math model for the the game mentioned in the previous example.

Observe that we could have chosen our state space as

$$\Omega = \left\{ \text{𓄿}, \text{𓆑}, \text{𓅓}, \text{𓆛}, \text{𓀭}, \text{𓅞} \right\}.$$

If we then define the random variable as $Y(\text{𓄿}) = Y(\text{𓅓}) = 5$, $Y(\text{𓆑}) = Y(\text{𓀭}) = Y(\text{𓅞}) = 2$, and $Y(\text{𓆛}) = -1$, then we are still modeling the same game and (10.3) still holds. In fact, if we change the weights of our die so that 𓄿, 𓆑, 𓅓, 𓆛, 𓀭, and 𓅞 come up with probabilities, respectively, 1/12, 5/24, 3/12, 1/6, 1/24, and 1/4, then (10.3) still holds and we are once again modeling the same game even though we are using a different die! The point is that what matters are the values X takes and the corresponding probabilities.

Homework Problems

Read: section 2.2 of Ash's book.

Exercise 10.1. We toss a fair coin 3 times. Let X be the number of tails we obtain. Give a sample space Ω, a probability measure P and a random variable $X : \Omega \to \mathbb{R}$ corresponding to this experiment.

Exercise 10.2. We roll a fair die 3 times. Let X be the product of the outcomes. Give a sample space Ω, a probability measure P and a random variable $X : \Omega \to \mathbb{R}$ corresponding to this experiment.

Lecture 11

1. Random Variables, continued

Suppose X is a random variable, defined on some state space Ω. Let P be a probability on Ω (with events-set $\mathscr{F}$). By the *distribution* (or the *law*) of X *under* P we mean the collection of probabilities $P\{X \in A\}$, as A ranges over all Borel sets in $\mathbb{R}$. The law of a random variable determines all its statistical properties and hence characterizes the random variable.

Example 11.1. In the previous example 10.3, (10.2) gave the law of X and (10.3) gave the law of Y.

If we define $P_X(A) = P\{X \in A\}$, then one can check that this collection satisfies the rules of probability and is thus a probability on the set D. In other words, the law of a D-valued random variable is itself a probability on D. [In fact, this leads to a subtle point. Since we are now talking about a probability on D, we need an events set, say $\mathscr{G}$. But then we need $\{\omega : X(\omega) \in A\}$ to be in $\mathscr{F}$, for all $A \in \mathscr{G}$. This is actually another condition that we should require when defining a random variable, but we are overlooking this (important) technical point in this course.]

From now on we will focus on the study of two types of random variables: discrete random variables and continuous random variables.

2. Discrete Random Variables

If X takes values in a finite, or countably-infinite set, then we say that X is a *discrete random variable*. Its distribution is called a *discrete distribution*. The function

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

Figure 11.1. Jacob Bernoulli (also known as James or Jacques) (Dec 27, 1654 – Aug 16, 1705, Switzerland)

is then called the *mass function* of X. The values x for which $f(x) > 0$ are called the *possible values* of X.

Note that in this case knowledge of the mass function is sufficient to determine the law of X. Indeed, for any subset $A \subset D$,

$$P\{X \in A\} = \sum_{x \in A} P\{X = x\}, \tag{11.1}$$

since $A = \cup_{x \in A}\{x\}$ and this is a countable union of disjoint sets.

Here are two important properties of mass functions:

- $0 \leqslant f(x) \leqslant 1$ for all x. [Easy]
- $\sum_{x \in \Omega} f(x) = 1$. [Use $A = D$ in (11.1)]

In fact, given a countable set D and a function f satisfying the above two properties, we can reverse engineer a random variable with mass function f. Just take $\Omega = D$, $P(\{x\}) = f(x)$ for $x \in D$, and $X(x) = x$.

The upshot is that to describe a discrete random variable it is enough to give a formula for its mass function.

3. The Bernoulli distribution

Suppose we perform a trial once. Let $p \in [0, 1]$ be the probability of "success". So the state space is $\Omega = \{\text{success}, \text{failure}\}$. Let $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. Then, X is said to have a Bernoulli distribution with parameter p [$X \sim \text{Bernoulli}(p)$]. The mass function is simple:

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

A nice and useful way to rewrite this is as $f(x) = p^x(1-p)^{1-x}$, if $x \in \{0,1\}$, and $f(x) = 0$ otherwise.

4. The binomial distribution (Bernoulli)

Suppose we perform n independent trials; each trial leads to a "success" or a "failure"; and the probability of success per trial is the same number $p \in [0,1]$. This is like fairly tossing n coins that each give heads with probability p, and calling heads a success.

Let X denote the total number of successes in this experiment. This is a discrete random variable with possible values $0, \dots, n$. We say then that X is a binomial random variable [$X \sim \text{Binomial}(n,p)$].

Math modelling questions:

- Construct an Ω: $\Omega = \{0,1\}^n$ (with 1 being a success).
- Construct X on this Ω: $X(\omega_1, \dots, \omega_n) = \sum_{i=1}^n \omega_i$.

Let us find the mass function of X. We seek to find $f(x)$, where $x = 0, \dots, n$. For all other values of x, $f(x) = 0$.

Now suppose x is an integer between zero and n. Note that $f(x) = P\{X = x\}$ is the probability of getting exactly x successes and $n-x$ failures. There are $\binom{n}{x}$ ways to choose which x trials were successes. Moreover, by independence, the probability of getting any specific combination of x successes (e.g. first x trials were successes and the rest were failures) is $p^x(1-p)^{n-x}$. Therefore,

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

Note that $\sum_x f(x) = (p + 1 - p)^n = 1$ by the binomial theorem.

Remark 11.2. Observe that if $B_1, \dots, B_n$ are independent Bernoulli(p) random variables (i.e. outcomes of n fair coin tosses), then $X = \sum_{k=1}^n B_k$ is Binomial(n, p).

Example 11.3. Ten percent of a certain (large) population smoke. If we take a random sample [without replacement] of 5 people from this population, what are the chances that at least 2 people smoke in the sample?

Let X denote the number of smokers in the sample. Then $X \sim \text{Binomial}(n\,,p)$, with $p = 0.1$ and $n = 5$ ["success" = "smoker"]. Therefore,

$$\begin{aligned} P\{X \geqslant 2\} &= 1 - P\{X \leqslant 1\} = 1 - [f(0) + f(1)] \\ &= 1 - \left[\binom{5}{0}(0.1)^0(1-0.1)^{5-0} + \binom{5}{1}(0.1)^1(1-0.1)^{5-1}\right] \\ &= 1 - (0.1)^5 - 5(0.1)(0.9)^4. \end{aligned}$$

Alternatively, we can follow the longer route and write

$$P\{X \geqslant 2\} = P\left(\{X = 2\} \cup \cdots \{X = 5\}\right) = f(2) + f(3) + f(4) + f(5).$$

Homework Problems

Exercise 11.1. Consider a sequence of five Bernoulli trials. Let X be the number of times that a head is followed immediately by a tail. For example, if the outcome is $\omega = HHTHT$ then $X(\omega) = 2$ since a head is followed directly by a tail at trials 2 and 3, and also at trials 4 and 5. Find the probability mass function of X.

Exercise 11.2. We roll a fair die three times. Let X be the number of times that we roll a 6. What is the probability mass function of X?

Exercise 11.3. We roll two fair dice.

(a) Let X be the product of the two outcomes. What is the probability mass function of X?

(b) Let X be the maximum of the two outcomes. What is the probability mass function of X?

Exercise 11.4. Let $\Omega = \{1, \dots, 6\}^2 = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \dots, 6\}\}$ and P the probability measure given by $P\{\omega\} = \frac{1}{36}$, for all $\omega \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be the number of dice that rolled even. Give the probability mass function of X.

Exercise 11.5. An urn contains 5 balls numbered from 1 to 5. We draw 3 of them at random without replacement.

(a) Let X be the largest number drawn. What is the probability mass function of X?

(b) Let X be the smallest number drawn. What is the probability mass function of X?

Lecture 12

1. The geometric distribution

Suppose we now do not fix the number of independent trials at n. Instead, we keep running the trials until the first success. Another way to think about this is as follows. A p-*coin* is a coin that tosses heads with probability p and tails with probability $1-p$. Suppose we toss a p-coin until the first time heads appears. Let X denote the number of tosses made. Then X is a so-called geometric random variable ["$X \sim \text{Geometric}(p)$"].

Evidently, if n is an integer greater than or equal to one, then $P\{X = n\} = (1-p)^{n-1}p$. Therefore, the mass function of X is given by

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

1.1. The tail of the distribution.

Example 12.1. A couple has children until their first son is born. Suppose the genders of their children are independent from one another, and the probability of girl is 0.6 every time. Let X denote the number of their children to find then that $X \sim \text{Geometric}(0.4)$. In particular,

$$\begin{aligned} P\{X \leqslant 3\} &= f(1) + f(2) + f(3) \\ &= p + p(1-p) + p(1-p)^2 \\ &= p\left[1 + 1 - p + (1-p)^2\right] \\ &= p\left[3 - 3p + p^2\right] \\ &= 0.784. \end{aligned}$$

This gives $P\{X \geqslant 4\} = 1 - 0.784 = 0.216$.

More generally, consider the tail of the distribution of $X \sim \text{Geometric}(p)$ (probability of large values). Namely, the probability that $X \geqslant n$. This is the same as the probability of failing in all of the first $n - 1$ experiments. That is,

$$P\{X \geqslant n\} = (1-p)^{n-1}.$$

In the above couples example, $P\{X \geqslant 4\} = 0.6^3$.

Example 12.2. Let $X \sim \text{Geometric}(p)$. Fix an integer $k \geqslant 1$. Then, the conditional probability of $X - k = x$, given $X \geqslant k+1$ equals

$$\begin{aligned} P\{X - k = x \,|\, X \geqslant k+1\} &= \frac{P\{X = k + x \text{ and } X \geqslant k+1\}}{P\{X \geqslant k+1\}} \\ &= \frac{P\{X = k+x\}}{P\{X \geqslant k+1\}} = \frac{p(1-p)^{x+n-1}}{(1-p)^k} = p(1-p)^{x-1}. \end{aligned}$$

This says that if we know we have not gotten heads by the k-th toss (i.e. $X \geqslant k+1$), then the distribution of when we will get the first head, from that moment on (i.e. $X - k$), is again geometric with the same parameter. This, of course, makes sense: we are still using the same coin, still waiting for the first heads to come, and the future tosses are independent of the first k we made so far; i.e. we might as well consider we are starting afresh! This fact is usually stated as: "the geometric distribution forgets the past."

2. The negative binomial (or Pascal) distribution

Suppose we are tossing a p-coin, where $p \in [0\,,1]$ is fixed, until we obtain r heads. Let X denote the number of tosses needed. Then, X is a discrete random variable with possible values $r, r+1, r+2, \ldots$. When $r = 1$, then X is $\text{Geometric}(p)$. In general,

$$f(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{if } x = r, r+1, r+2, \ldots\,, \\ 0 & \text{otherwise.} \end{cases}$$

This X is said to have a *negative binomial distribution with parameters* r *and* p. Note that our definition differs slightly from that of Stirzaker's text (p. 117).

We can think of the whole process as follows: first toss a p-coin until the first head is obtained, then toss an independent p-coin, until the second head appears, and so on. This shows that the negative binomial(r, p) is in fact the sum of r independent $\text{Geometric}(p)$. We will prove this rigorously when we study moment generating functions.

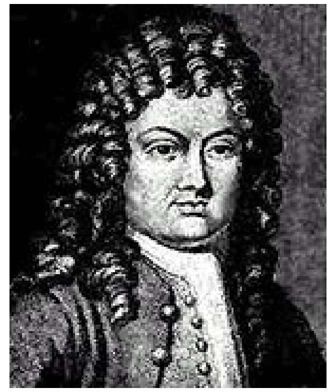



Figure 12.1. Left: Siméon Denis Poisson (Jun 21, 1781 – Apr 25, 1840, France). Right: Sir Brook Taylor (Aug 18, 1685 – Nov 30, 1731, England)

3. The Poisson distribution (Poisson, 1838)

Choose and fix a number $\lambda > 0$. A random variable X is said to have the *Poisson distribution with parameter* λ ($X \sim \text{Poisson}(\lambda)$) if its mass function is

$$f(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases} \tag{12.1}$$

In order to make sure that this makes sense, it suffices to prove that $\sum_x f(x) = 1$, but this is an immediate consequence of the Taylor expansion of e^λ, viz.,

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

Poisson random variables are often used to model the length of a waiting list or a queue (e.g. the number of people ahead of you when you stand in line at the supermarket). The reason this makes a good model is made clear in the following section.

3.1. Law of rare events. Is there a physical manner in which $\text{Poisson}(\lambda)$ arises naturally? The answer is "yes." Let $X = \text{Binomial}(n, \lambda/n)$. For instance, X could denote the total number of sampled people who have a rare disease (population percentage $= \lambda/n$) in a large sample of size n. Or, the total number of people, in a population of size n, who decide to stand in line in the supermarket, with each of them making an independent decision to join the queue with a small chance of λ/n [in order to make the "average length" of the line about $n \times \lambda/n = \lambda$]. Then, for all fixed integers $k = 0, \ldots, n$,

$$f_X(k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}. \tag{12.2}$$

Poisson's "law of rare events" states that if n is large, then the distribution of X is approximately Poisson(λ). This explains why Poisson random variables make good models for queue lengths.

In order to deduce this we need two computational lemmas.

Lemma 12.3. *For all $z \in \mathbf{R}$,*

$$\lim_{n\to\infty}\left(1+\frac{z}{n}\right)^n = e^z.$$

Proof. Because e^x is continuous, it suffices to prove that

$$\lim_{n\to\infty} n \ln\left(1+\frac{z}{n}\right) = z. \tag{12.3}$$

By Taylor's expansion,

$$\ln\left(1+\frac{z}{n}\right) = \frac{z}{n} + \frac{\theta^2}{2},$$

where θ lies between 0 and z/n. Equivalently,

$$\frac{z}{n} \leqslant \ln\left(1+\frac{z}{n}\right) \leqslant \frac{z}{n} + \frac{z^2}{2n^2}.$$

Multiply all sides by n and take limits to find (12.3), and thence the lemma.

Alternatively, one can set $h = z/n$ and write (12.3) as

$$z\lim_{h\to 0}\frac{\ln(1+h)}{h} = z\lim_{h\to 0}\frac{\ln(1+h)-\ln(1)}{h} = z(\ln x)'\big|_{x=0} = z\,. \qquad \square$$

Lemma 12.4. *If $k \geqslant 0$ is a fixed integer, then*

$$\binom{n}{k} \sim \frac{n^k}{k!} \qquad \textit{as } n \to \infty.$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty}(a_n/b_n) = 1$.

Proof. If $n \geqslant k$, then

$$\begin{aligned}\frac{n!}{n^k(n-k)!} &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\\ &= \frac{n}{n}\times\frac{n-1}{n}\times\cdots\times\frac{n-k+1}{n}\\ &\to 1 \qquad \text{as } n\to\infty.\end{aligned}$$

The lemma follows upon writing out $\binom{n}{k}/\frac{n^k}{k!}$ and applying the above. $\square$

Thanks to Lemmas 12.3 and 12.4, and to (12.2),

$$f_X(k) \sim \frac{n^k}{k!}\frac{\lambda^k}{n^k}e^{-\lambda} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

That is, when n is large, X behaves like a Poisson(λ), and this proves our assertion.

Homework Problems

Exercise 12.1. Solve the following.

(a) Let X be a geometric random variable with parameter p. Prove that $\sum_{n=1}^{\infty} P\{X = n\} = 1$.

(b) Let Y be a Poisson random variable with parameter λ. Prove that $\sum_{n=0}^{\infty} P\{Y = n\} = 1$.

Exercise 12.2. Some day, 10,000 cars are travelling across a city ; one car out of 5 is gray. Suppose that the probability that a car has an accident this day is 0.002. Using the approximation of a binomial distribution by a Poisson distribution, compute:

(a) the probability that exactly 15 cars have an accident this day;

(b) the probability that exactly 3 gray cars have an accident this day.

Lecture 13

1. (Cumulative) distribution functions

Let X be a [real-valued] random variable. The (cumulative) *distribution function* (CDF) F of X *under* P is defined by

$$F(x) = P\{X \leqslant x\}.$$

Here are some basic properties distribution functions have to satisfy.

(a) $F(x) \leqslant F(y)$ whenever $x \leqslant y$; i.e. F is non-decreasing.

(b) $\lim_{b\to\infty} F(b) = 1$ and $\lim_{a\to-\infty} F(a) = 0$.

(c) F is right-continuous. That is, $\lim_{y\searrow x} F(y) = F(x)$ for all x.

Property (a) just follows from the fact that $(-\infty, x] \subset (-\infty, y]$. The other two properties follow from the facts that $\cup_{n\geqslant1}(-\infty, n] = (-\infty, \infty)$, $\cap_{n\geqslant1}(-\infty, -n] = \varnothing$, $\cup_{n\geqslant1}(-\infty, x + 1/n] = (-\infty, x]$, and the following lemma.

Lemma 13.1. *Let P be a probability. Let A_n be an increasing set of events: $A_1 \subset A_2 \subset A_3 \subset \cdots$. Then,*

$$\lim_{n\to\infty} P(A_n) = P\big(\cup_{n\geqslant1} A_n\big).$$

Similarly, if A_n is decreasing, i.e. $A_1 \supset A_2 \supset A_3 \supset \cdots$, then

$$\lim_{n\to\infty} P(A_n) = P\big(\cap_{n\geqslant1} A_n\big).$$

Proof. Let us start with the first statement. The proof uses Rule 4 of probability. To do this we write

$$\cup_{n\geqslant1} A_n = A_1 \cup (A_2 \smallsetminus A_1) \cup (A_3 \smallsetminus A_2) \cup \cdots .$$

Note that the sets on the right-hand-side are disjoint. Hence,

$$\begin{aligned} P\big(\cup_{n\geqslant 1} A_n\big) &= P(A_1) + \sum_{i\geqslant 2} P(A_i \smallsetminus A_{i-1}) \\ &= P(A_1) + \lim_{n\to\infty} \sum_{i=2}^{n} (P(A_i) - P(A_{i-1})) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

The other statement follows from taking complements. □

In fact, the converse is also true. Any function F that satisfies the above properties (a)-(c) is a distribution function of some random variable X. This is because of the following property. If X has distribution function F under P, then

(d) $F(b) - F(a) = P\{a < X \leqslant b\}$ for $a < b$.

Now, say we have a function F satisfying (a)-(c) and we want to reverse engineer a random variable X with distribution function F. Let $\Omega = (-\infty, \infty)$ and for $a < b$ define

$$P((a, b]) = F(b) - F(a). \tag{13.1}$$

Recall at this point the Borel sets from Example 3.10. It turns out that properties (a)-(c) are exactly what is needed to be able to extend (13.1) to a collection $\{P(B) : B \text{ is a Borel set}\}$ that satisfies the rules of probability. This fact has a pretty sophisticated proof that we omit here. But then, consider the random variable $X(\omega) = \omega$. Its distribution function under P is equal to

$$P\{X \leqslant x\} = P((-\infty, x]) = P\big(\cap_{n\geqslant 1} (-n, x]\big) = \lim_{n\to\infty} (F(x) - F(-n)) = F(x).$$

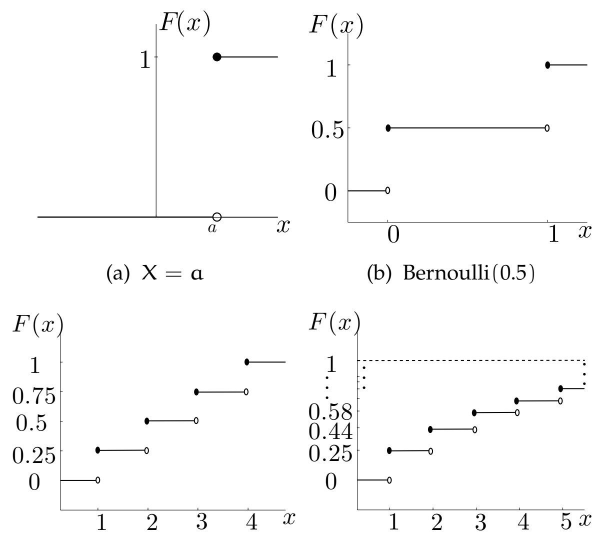
The upshot is that it is in general (whether X is discrete, continuous, or neither) enough to specify the CDF in order to fully describe a random variable.

Here are two more useful properties of distribution functions.

(e) $P\{X > x\} = 1 - F(x)$.

(f) $P\{X = x\} = F(x) - \lim_{y \nearrow x} F(y)$ is the size of the jump [if any] at x.

The last property is proved again using Lemma 13.1. It shows that for a discrete random variable the distribution function is a step function that jumps precisely at the possible values of X. The size of the jump at x is



(a) $X = a$ (b) Bernoulli(0.5) (c) Uniform$(1,2,3,4)$ (d) Geometric(0.25)

Figure 13.1. CDFs for some discrete distributions

exactly the mass function $f(x)$. In fact, in this case

$$F(x) = \sum_{y:\, y \leqslant x} f(y). \tag{13.2}$$

In particular, the CDF of a discrete random variable is piecewise constant.

Example 13.2. Let X be nonrandom. That is, $P\{X = a\} = 1$ for some number a. Such a random variable is called "deterministic." Then (see Figure 13.1(a)),

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ 1 & \text{if } x \geqslant a. \end{cases}$$

Example 13.3. Let X be Bernoulli with parameter $p \in [0, 1]$. Then (see Figure 13.1(b)),

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - p & \text{if } 0 \leqslant x < 1, \\ 1 & \text{if } x \geqslant 1. \end{cases}$$

Example 13.4. Let $\Omega = \{1, \ldots, n\}$ and let $X(k) = k$ for all $k \in \Omega$. Let P the probability on Ω corresponding to choosing an element, equally likely;

$P\{k\} = 1/n$ for all $k \in \Omega$. Then (see Figure 13.1(c)),

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{k}{n} & \text{if } k \leqslant x < k+1, \, k \in \{1, \cdots, n-1\}, \\ 1 & \text{if } x \geqslant n. \end{cases}$$

Example 13.5. Let X be binomial with parameters n and p. Then,

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} & \text{if } k \leqslant x < k+1, \, 0 \leqslant k < n, \\ 1 & \text{if } x \geqslant n. \end{cases}$$

Example 13.6. Let X be geometric with parameter p. Then (see Figure 13.1(d)),

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1-(1-p)^n & \text{if } n \leqslant x < n+1, \, n \geqslant 1. \end{cases}$$

Here, we used the fact that

$$f(0) + f(1) + \cdots + f(n) = p + (1-p)p + \cdots + (1-p)^{n-1}p = 1-(1-p)^n.$$

Homework Problems

Exercise 13.1. Let F be the function defined by:

$$\mathsf{F}(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x^2}{3} & \text{if } 0 \leqslant x < 1, \\ \frac{1}{3} & \text{if } 1 \leqslant x < 2, \\ \frac{1}{6}x + \frac{1}{3} & \text{if } 2 \leqslant x < 4, \\ 1 & \text{if } x \geqslant 4. \end{cases}$$

Let X be a random variable which corresponds to F.

(a) Verify that F is a cumulative distribution function.

(b) Is X discrete or continuous?

(c) Compute $P\{X = 2\}$.

(d) Compute $P\{X < 2\}$.

(e) Compute $P\{X = 2 \text{ or } \frac{1}{2} \leqslant X < \frac{3}{2}\}$.

(f) Compute $P\{X = 2 \text{ or } \frac{1}{2} \leqslant X \leqslant 3\}$.

Lecture 14

1. Continuous Random Variables

We say that X is a *continuous* random variable with (probability) *density function* (pdf) f if f is a piecewise continuous nonnegative function, and for all real numbers x,

$$F(x) = P\{X \leqslant x\} = \int_{-\infty}^{x} f(y)\, dy.$$

As a consequence, F is a continuous function when X is a continuous random variable.

It is noteworthy at this point that if F is not continuous nor piecewise constant then it is not the CDF of a discrete nor of a continuous random variable. (Draw a CDF like that!)

Comparing the above to (13.2) shows that f is playing the role of the mass function that was used in the discrete case. However, note that when X is continuous,

$$P\{X = x\} = F(x) - \lim_{y \nearrow x} F(y) = 0,$$

for all x. Hence, $f(x)$ is <u>not</u> a mass function. A good way to think about it is as the "likelihood" of getting outcome x, instead of as the probability of getting x.

If f is continuous at x, then by the fundamental theorem of calculus,

$$F'(x) = f(x).$$

And since F is non-decreasing, we have that $f(x) \geqslant 0$, for all x where f is continuous.

Conversely, *any* piecewise continuous f such that $\int_{-\infty}^{\infty} f(y)\,dy = 1$ and $f(x) \geqslant 0$ corresponds to a continuous random variable X. Simply define $F(x) = \int_{-\infty}^{x} f(y)\,dy$ and check that properties (a)-(c) of distribution functions are satisfied!

In fact, if X has pdf f, then

$$P\{X \in A\} = \int_A f(x)\,dx.$$

In particular,

$$P\{a \leqslant X \leqslant b\} = P\{a < X \leqslant b\} = P\{a \leqslant X < b\} = P\{a < X < b\} = \int_a^b f(x)\,dx.$$

Example 14.1. Say X is a continuous random variable with probability density function

$$f(x) = \frac{1}{4x^2}, \text{ if } |x| > 1/2.$$

Then, to find $P\{X^4 - 2X^3 - X^2 + 2X > 0\}$ we need to write the set in question as a union of disjoint intervals and then integrate f over each interval and add up the results. So we observe that

$$X^4 - 2X^3 - X^2 + 2X = (X+1)X(X-1)(X-2)$$

and thus the region in question is $(-\infty, -1) \cup (0, 1) \cup (2, \infty)$. Note that X is never in $(0, 1/2)$, since f_X vanishes there. The probability of X being in this region is then

$$\int_{-\infty}^{-1} \frac{1}{4x^2}\,dx + \int_{1/2}^{1} \frac{1}{4x^2}\,dx + \int_{2}^{\infty} \frac{1}{4x^2}\,dx = \frac{5}{8}.$$

Example 14.2. Say X is a continuous random variable with probability density function $f(x) = \frac{1}{4}\min(1, \frac{1}{x^2})$. Then $f(x) = \frac{1}{4}$ for $-1 \leqslant x \leqslant 1$ and $f(x) = \frac{1}{4x^2}$, for $|x| \geqslant 1$. Thus,

$$P\{-2 \leqslant X \leqslant 4\} = \int_{-2}^{-1} \frac{1}{4x^2}\,dx + \int_{-1}^{1} \frac{1}{4}\,dx + \int_{1}^{4} \frac{1}{4x^2}\,dx = \frac{13}{16}.$$

Homework Problems

Exercise 14.1. Let X be a random variable with probability density function given by

$$f(x) = \begin{cases} c(4-x^2) & \text{if } -2 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is the value of c?

(b) Find the cumulative distribution function of X.

Exercise 14.2. Let X be a random variable with probability density function given by

$$f(x) = \begin{cases} c\cos^2(x) & \text{if } 0 < x < \frac{\pi}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is the value of c?

(b) Find the cumulative distribution function of X.

Exercise 14.3. Let X be a random variable with probability density function given by

$$f(x) = \frac{1}{2}\exp(-|x|).$$

Compute the probabilities of the following events:

(a) $\{|X| \leqslant 2\}$,

(b) $\{|X| \leqslant 2 \text{ or } X \geqslant 0\}$,

(c) $\{|X| \leqslant 2 \text{ or } X \leqslant -1\}$,

(d) $\{|X| + |X-3| \leqslant 3\}$,

(e) $\{X^3 - X^2 - X - 2 \geqslant 0\}$,

(f) $\{e^{\sin(\pi X)} \geqslant 1\}$,

(g) $\{X \in \mathbb{N}\}$.

Exercise 14.4. Solve the following.

(a) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \dfrac{c}{\sqrt{x}} & \text{if } x \geqslant 1, \\ 0 & \text{if } x < 1. \end{cases}$$

Does there exist a value of c such that f becomes a probability density function?

(b) Let $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(x) = \begin{cases} e^{-\frac{1}{x}} & \text{if } x > 0, \\ 0 & \text{if } x \leqslant 0. \end{cases}$$

Is F a cumulative distribution function? If yes, what is the associated probability density function?

Exercise 14.5. (a) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \dfrac{c}{1+x^2} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Does there exist a value of c such that f becomes a probability density function?

(b) Let $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(x) = \frac{1}{2}\left(1 + \frac{x}{\sqrt{1+x^2}}\right), \qquad x \in \mathbb{R}.$$

Is F a cumulative distribution function ? If yes, what is the associated probability density function?

Lecture 15

1. Continuous Random Variables, continued

Here are some standard examples of continuous random variables.

Example 15.1 (Uniform density)**.** If $a < b$ are fixed, then the uniform density on $(a\,,b)$ is the function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 0 & \text{otherwise;} \end{cases}$$

see Figure 15.1(a). In this case, we can compute the distribution function as follows:

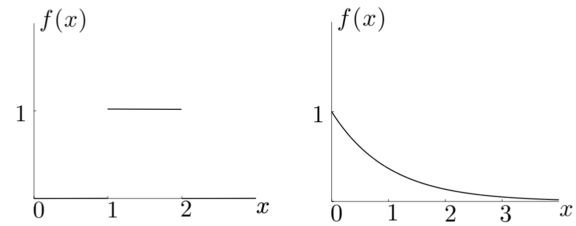
$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{x-a}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 1 & \text{if } x > b. \end{cases}$$

A random variable with this density ($X \sim \text{Uniform}(a, b)$) takes any value in $[a, b]$ "equally likely" and has 0 likelihood of taking values outside $[a, b]$.

Note that if $a < c < d < b$, then

$$P\{c \leqslant X \leqslant d\} = F(d) - F(c) = \frac{d-c}{b-a}.$$

This says that the probability we will pick a number in $[c, d]$ is equal to the ratio of $d - c$ "the number of desired outcomes" by $b - a$ "the total number of outcomes."



(a) Uniform(1, 2) (b) Exponential(1)

Figure 15.1. pdf for certain continuous distributions

Example 15.2 (Exponential densities)**.** Let $\lambda > 0$ be fixed. Then

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

is a density, and is called the *exponential density with parameter* λ. See Figure 15.1(b). It is not hard to see that

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

We write $X \sim \text{Exponential}(\lambda)$ to say that X is distributed exponentially with parameter λ.

The exponential distribution is the continuous analogue of the geometric distribution. In fact, just as Poisson's law of rare events explains how binomial random variables approximate Poisson random variables, there is also a sense in which geometric random variables "approximate" exponential ones. This explains why the latter are used to model waiting times; e.g. the time it takes to be served when you are first in line at the supermarket.

To see this, imagine the cashier operates as follows: they flip a coin every $1/n$ seconds and serve you only when the coin falls heads. The coin, however, is balanced to give heads with a small probability of λ/n. So on average, it will take about n/λ coin flips until you get heads, and you will be served in about $1/\lambda$ seconds.

Now, let n be large (i.e. decisions whether to serve you or not are made very often). Let X be the time when you get served. Then, the probability you get served by time x ($P\{X \leqslant x\}$) is the same as the probability the coin lands heads in the first nx tosses. This is equal to the distribution function at nx of a geometric variable with parameter λ/n. That is, $1 - (1 - \lambda/n)^{nx}$. By Lemma 12.3 this converges to $1 - e^{-\lambda x}$, which is the distribution function of an exponential random variable with parameter λ.

Remark 15.3. A familiar situation that may be helpful to have in mind while making sense of the above is how continuously compound interest arrises. Recall that if the interest is compounded n times a year, with interest rate r, then an initial amount of A dollars becomes $A(1 + r/n)^n$. Compounding interest continuously simply means $n \to \infty$ and thus the amount becomes Ae^r. Note that if we compound n times, then the interest rate for each time is r/n, not r. Do you see how this is similar to the above derivation of the exponential distribution?

Example 15.4. Just as we have seen for a geometric random variable, an exponential random variable does not recall history; see Example 12.2. Indeed, if $X \sim \text{Exponential}(\lambda)$, then for $a \geqslant 0$ and $x \geqslant 0$ we have

$$P\{X - a \leqslant x \,|\, X > a\} = \frac{P\{a < X \leqslant a + x\}}{P\{X > a\}} = \frac{e^{-\lambda a} - e^{-\lambda(a+x)}}{e^{-\lambda a}} = 1 - e^{-\lambda x};$$

i.e. given that you have not been served by time a, the distribution of the remaining waiting time is again exponential with the same parameter λ. Makes sense, no?

Lecture 16

Figure 16.1. Baron Augustin-Louis Cauchy (Aug 21, 1789 – May 23, 1857, France)

1. Continuous Random Variables, continued

Example 16.1 (The Cauchy density (Cauchy, 1827))**.** Define for all real numbers x,

$$f(x) = \frac{1}{\pi}\,\frac{1}{1+x^2}.$$

Because

$$\frac{d}{dx}\arctan x = \frac{1}{1+x^2},$$

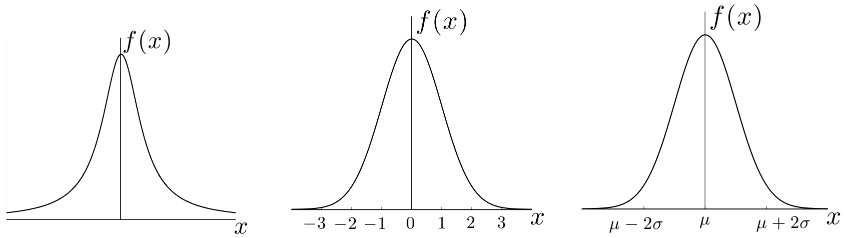
we have

$$\int_{-\infty}^{\infty} f(y)\,dy = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+y^2}\,dy = \frac{1}{\pi}\left[\arctan(\infty) - \arctan(-\infty)\right] = 1.$$

Hence, f is a density; see Figure 16.2(a). Also,

$$\begin{aligned} F(x) &= \frac{1}{\pi}\int_{-\infty}^{x} f(y)\,dy = \frac{1}{\pi}\left[\arctan(x) - \arctan(-\infty)\right] \\ &= \frac{1}{\pi}\arctan(x) + \frac{1}{2} \qquad \text{for all real } x. \end{aligned}$$

Note that f decays rather slowly as $|x| \to \infty$ (as opposed to an exponential distribution, for example). This means that a Cauchy distributed random variable has a "good chance" of taking large values. For example, it turns



(a) Cauchy (b) Standard Normal (c) Normal(μ,σ^2)

Figure 16.2. pdf for certain continuous distributions

out that it is a good model of the distance for which a certain type of squirrels carries a nut before burring it. The fat tails of the distribution then explain the vast spread of certain types of trees in a relatively short time period!

Example 16.2 (Standard normal density)**.** I claim that

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

defines a density function; see Figure 2(b). Clearly, $\phi(x) \geqslant 0$ and is continuous at all points x. So it suffices to show that the area under ϕ is one. Define

$$A = \int_{-\infty}^{\infty} \phi(x)\,dx.$$

Then,

$$A^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy.$$

Changing to polar coordinates ($x = r\cos\theta$, $y = r\sin\theta$ gives a Jacobian of r) one has

$$A^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}}\,r\,dr\,d\theta.$$

Let $s = r^2/2$ to find that the inner integral is $\int_0^\infty e^{-s}\,ds = 1$. Therefore, $A^2 = 1$ and hence $A = 1$, as desired. [Why is A not -1?]

The distribution function of ϕ is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,dx.$$

Of course, we know that $\Phi(z) \to 0$ as $z \to -\infty$ and $\Phi(z) \to 1$ as $z \to \infty$. Due to symmetry, we also know that $\Phi(0) = 1/2$. (Check that!) Unfortunately, a theorem of Liouville tells us that $\Phi(z)$ cannot be computed (in terms of other "nice" functions). In other words, $\Phi(z)$ cannot be computed

Figure 16.3. Johann Carl Friedrich Gauss (Apr 30, 1777 – Feb 23, 1855, Germany)

exactly for any value of z other than $z = 0, \pm\infty$. Therefore, people have approximated and tabulated $\Phi(z)$ for various choices of z, using standard methods used for approximating integrals; see the table in Appendix C.

Here are some consequences of that table [check!!]:

$$\Phi(0.09) \approx 0.5359, \quad \Phi(0.90) \approx 0.8159, \quad \Phi(3.35) \approx 0.9996.$$

And because ϕ is symmetric, $\Phi(-z) = 1 - \Phi(z)$. Therefore [check!!],

$$\Phi(-0.09) = 1 - \Phi(0.09) \approx 1 - 0.5359 = 0.4641, \quad \text{etc.}$$

Of course, nowadays one can also use software to compute $\Phi(z)$ very accurately. For example, in Excel one can use the command `NORMSDIST(0.09)` to compute $\Phi(0.09)$.

Example 16.3 (Normal or Gaussian density (Gauss, 1809))**.** Given two numbers $-\infty < \mu < \infty$ and $\sigma > 0$, the *normal curve* (Normal(μ,σ^2)) is described by the density function

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \qquad \text{for } -\infty < x < \infty;$$

see Figure 16.2. Using a change of variables, one can relate this distribution to the standard normal one, denoted N(0,1). Indeed, for all

$-\infty < a \leqslant b < \infty$,

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\,dx \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \qquad [z=(x-\mu)/\sigma] \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z)\,dz \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).\end{aligned} \tag{16.1}$$

One can take $a \to -\infty$ or $b \to \infty$ to compute, respectively,

$$\int_{-\infty}^b f(x)dx = \Phi\left(\frac{b-\mu}{\sigma}\right) \text{ and } \int_a^\infty f(x)dx = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Note at this point that taking both $a \to -\infty$ and $b \to \infty$ proves that f is indeed a density curve (i.e. has area 1 under it). The operation $x \mapsto z = (x-\mu)/\sigma$ is called *standardization*. Thus, the above calculation shows that the area between a and b under the Normal(μ,σ^2) curve is the same as the one between the *standard scores* of a and b but under the standard Normal(0,1) curve. One can now use the standard normal table to estimate these areas.

Lecture 17

1. Continuous Random Variables, continued

Example 17.1 (Gamma densities)**.** Choose and fix two numbers (parameters) $\alpha, \lambda > 0$. The *gamma density* with parameters α and λ is the probability density function that is proportional to

$$\begin{cases} x^{\alpha-1}e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

The above is nonnegative, but does not necessarily integrate to 1. Thus, to make it into a density function we have to divide it by its integral (from 0 to ∞). Now,

$$\int_0^\infty x^{\alpha-1}e^{-\lambda x}\,dx = \frac{1}{\lambda^\alpha}\int_0^\infty y^{\alpha-1}e^{-y}\,dy.$$

Define the *gamma function* as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy \qquad \text{for all } \alpha > 0.$$

One can prove that there is "no nice formula" that "describes" $\Gamma(\alpha)$ for all α (theorem of Liouville). Thus, the best we can do is to say that the following is a Gamma density with parameters $\alpha, \lambda > 0$:

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases}$$

You can probably guess by now (and correctly!) that $F(x) = \int_{-\infty}^x f(y)\,dy$ cannot be described by nice functions either. Nonetheless, let us finish by making the observation that $\Gamma(\alpha)$ is computable for some reasonable values of $\alpha > 0$. The key to unraveling this remark is the following "reproducing property":

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \qquad \text{for all } \alpha > 0. \tag{17.1}$$

The proof uses integration by parts:

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= \int_0^\infty u(x)v'(x)\,dx,\end{aligned}$$

where $u(x) = x^\alpha$ and $v'(x) = e^{-x}$. Integration by parts states that[1]

$$\int uv' = uv - \int v'u \qquad \text{for indefinite integrals.}$$

Evidently, $u'(x) = \alpha x^{\alpha-1}$ and $v(x) = -e^{-x}$. Hence,

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x}\,dx \\ &= uv\Big|_0^\infty - \int_0^\infty v'u \\ &= \left(-\alpha x^{\alpha-1}e^{-x}\right)\Big|_0^\infty + \alpha\int_0^\infty x^{\alpha-1}e^{-x}\,dx.\end{aligned}$$

The first term is zero, and the second (the integral) is $\alpha\Gamma(\alpha)$, as claimed. Now, it easy to see that $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$. Therefore, $\Gamma(2) = 1 \times \Gamma(1) = 1$, $\Gamma(3) = 2 \times \Gamma(2) = 2, \ldots$, and in general,

$$\Gamma(n) = (n-1)! \qquad \text{for all integers } n \geqslant 1.$$

It is also not too hard to see that

$$\Gamma(1/2) = \int_0^\infty x^{-1/2}e^{-x}\,dx = \sqrt{2}\int_0^\infty e^{-y^2/2}\,dy = \sqrt{2} \times \frac{\sqrt{2\pi}}{2} = \sqrt{\pi}.$$

Thus,

$$\Gamma(n+1/2) = (n-1/2)(n-3/2)\cdots(1/2)\sqrt{\pi} \qquad \text{for all integers } n \geqslant 1.$$

In other words, even though $\Gamma(\alpha)$ is usually hard to compute, for a general α, it is quite easy to compute for α's that are are half nonnegative integers.

[1]This follows immediately from integrating the product rule: $(uv)' = u'v + uv'$.

2. Functions of a discrete random variable

Example 17.2. Suppose X has the mass function

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = -1, \\ \frac{1}{3} & \text{if } x = 0, \\ \frac{1}{2} & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Define a new random variable $Y = 2X^2 + 1$. Then, Y takes the values 1 and 3. The mass function of Y is

$$\begin{aligned} f_Y(y) &= P\{Y = y\} = P\{2X^2 + 1 = y\} = P\{X^2 = (y-1)/2\} \\ &= P\left\{X = \sqrt{(y-1)/2}\right\} + P\left\{X = -\sqrt{(y-1)/2}\right\} \\ &= f_X\left(\sqrt{(y-1)/2}\right) + f_X\left(-\sqrt{(y-1)/2}\right) \\ &= \begin{cases} \frac{1}{6} + \frac{1}{2} = \frac{2}{3} & \text{if } y = 3, \\ \frac{1}{3} & \text{if } y = 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The procedure of this example actually produces a theorem.

Theorem 17.3. *Let X be a discrete random variable with the set of possible values being D. If $Y = g(X)$ for a function g, then the set of possible values of Y is $g(D)$ and*

$$f_Y(y) = \begin{cases} \sum_{x:\, g(x)=y} f_X(x) & \text{if } y \in g(D), \\ 0 & \text{otherwise.} \end{cases}$$

When g is one-to-one and has inverse h (i.e. $x = h(y)$) then the formula simplifies to

$$f_Y(y) = f_X(h(y)). \tag{17.2}$$

In the above example, solving for x in terms of y gives

$$x = \begin{cases} -1 \text{ or } 1 & \text{if } y = 3, \\ 0 & \text{if } y = 1. \end{cases}$$

Thus,

$$f_Y(y) = \begin{cases} f_X(-1) + f_X(1) & \text{if } y = 3, \\ f_X(0) & \text{if } y = 1. \end{cases}$$

Lecture 18

1. Functions of a continuous random variable

The basic problem: If $Y = g(X)$, then how can we compute f_Y in terms of f_X? One way is to first compute F_Y from F_X and then take its derivative.

Example 18.1. Suppose X is uniform on $(0\,,1)$, and $Y = -\ln X$. Then, we compute f_Y by first computing F_Y, and then using $f_Y = F_Y'$. Here are the details:

$$F_Y(y) = P\{Y \leqslant y\} = P\{-\ln X \leqslant y\} = P\{\ln X \geqslant -y\}.$$

Now, the exponential function is an increasing function. Therefore, $\ln X \geqslant -y$ if and only if $X \geqslant e^{-y}$. Recalling that $F_X(x) = x$ for $x \in [0,1]$ we have

$$F_Y(y) = P\left\{X \geqslant e^{-y}\right\} = 1 - F_X(e^{-y}) = 1 - e^{-y}, \text{ for } y > 0.$$

We know, of course, that Y does not take negative values and so $F_Y(y) = 0$ for $y \leqslant 0$. Consequently, $f_Y(y) = 0$ for $y < 0$ and for $y > 0$ we have

$$f_Y(y) = F_Y'(y) = (1 - e^{-y})' = e^{-y}.$$

Let us make the observation that $X = e^{-Y}$ and

$$f_Y(y) = F_Y'(y) = (1 - F_X(e^{-y}))' = -f_X(e^{-y})(e^{-y})'.$$

This is not a coincidence.

Theorem 18.2. *Suppose X is a continuous random variable with density function f_X supported on a set $D \subset \mathbf{R}$. Let $g : D \to \mathbf{R}$ be a one-to-one function with inverse h and let $Y = g(X)$. Then,*

$$f_Y(y) = \begin{cases} f_X(h(y))\,|h'(y)| & \textit{for } y \in g(D), \\ 0 & \textit{otherwise.} \end{cases}$$

[Compare the above formula with the one for the discrete case (17.2)!]

Proof. We have two cases. If g is increasing, then so is h and we have

$$F_Y(y) = P\{Y \leqslant y\} = P\{g(X) \leqslant y\} = P\{X \leqslant h(y)\} = F_X(h(y)).$$

Thus,

$$f_Y(y) = f_X(h(y))\, h'(y) = f_X(h(y))\, |h'(y)|.$$

If, on the other hand, g is decreasing, then so is h and we have

$$F_Y(y) = P\{Y \leqslant y\} = P\{g(X) \leqslant y\} = P\{X \geqslant h(y)\} = 1 - F_X(h(y)).$$

[We have used the fact that X is continuous.] Thus,

$$f_Y(y) = -f_X(h(y))\, h'(y) = f_X(h(y))\, |h'(y)|. \qquad \square$$

Example 18.3. Suppose $\mu \in \mathbb{R}$ and $\sigma > 0$ are fixed constants, and define $Y = \mu + \sigma X$. Find the density of Y in terms of that of X. Since the transformation is one-to-one and its inverse is $x = (y - \mu)/\sigma$, we have

$$f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right| = \frac{1}{\sigma} f_X\left(\frac{y-\mu}{\sigma}\right).$$

For example, if X is standard normal, then

$$f_{\mu+\sigma X}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

In other words, $Y \sim N(\mu\,,\sigma^2)$.

Example 18.4. Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$. Then, $y = e^x > 0$, $x = \ln y$, and

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{\frac{-(\ln y-\mu)^2}{2\sigma^2}}, \text{ for } y > 0.$$

This is called the log-normal distribution. It is often encountered in medical and financial applications. By the central limit theorem, normally distributed random variables appear when a huge number of small independent errors are added. In chemistry, for example, concentrations are multiplied. So in huge reactions the logarithms of concentrations add up and give a normally distributed random variable. The concentration is then the exponential of this variable and is, therefore, a log-normal random variable.

Now what if g is not one-to-one?

The solution: First compute F_Y, by hand, in terms of F_X, and then use the fact that $F_Y' = f_Y$ and $F_X' = f_X$.

Example 18.5. Suppose X has density f_X. Then let us find the density function of $Y = X^2$. Again, we seek to first compute F_Y. Now, for all $y > 0$,

$$F_Y(y) = P\{X^2 \leqslant y\} = P\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Differentiate $[d/dy]$ to find that

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

On the other hand, $F_Y(y) = 0$ if $y \leqslant 0$ and so $f_Y(y) = 0$ as well.

For example, consider the case that X is standard normal. Then,

$$f_{X^2}(y) = \begin{cases} \dfrac{e^{-y}}{\sqrt{2\pi y}} & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0. \end{cases}$$

Or if X is Cauchy, then

$$f_{X^2}(y) = \begin{cases} \dfrac{1}{\pi\sqrt{y}(1+y)} & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0. \end{cases}$$

Example 18.6. If X is uniform $(0,1)$ and $Y = X^2$, then $X^2 = Y$ has one solution: $X = \sqrt{Y}$. Thus, we are in the 1:1 situation and

$$f_{X^2}(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

But what happens if X is uniform on $(-1,2)$ and $Y = X^2$? Well, then $X^2 = Y$ has two solutions when $0 < Y < 1$ and only one solution when $1 < Y < 4$. Repeat the above method to get that

$$f_{X^2}(y) = \begin{cases} \dfrac{1}{6\sqrt{y}} + \dfrac{1}{6\sqrt{y}} = \dfrac{1}{3\sqrt{y}} & \text{if } 0 < y < 1, \\ \dfrac{1}{6\sqrt{y}} & \text{if } 1 < y < 4, \\ 0 & \text{otherwise.} \end{cases}$$

(We leave this as an exercise but will show how it is done in the next example.)

Example 18.7. Suppose X is exponential with parameter $\lambda = 3$. Let $Y = (X-1)^2$. Then,

$$F_Y(y) = P\{1 - \sqrt{y} \leqslant X \leqslant 1 + \sqrt{y}\}.$$

Now, one has to be careful. If $0 \leqslant y \leqslant 1$, then

$$F_Y(y) = \int_{1-\sqrt{y}}^{1+\sqrt{y}} 3e^{-3x}\,dx$$

and

$$f_Y(y) = \frac{3e^{-3(1+\sqrt{y})} + 3e^{-3(1-\sqrt{y})}}{2\sqrt{y}}.$$

This formula cannot be true for y large. Indeed $e^{-3(1-\sqrt{y})}/\sqrt{y}$ goes to ∞ as $y \to \infty$, while f_Y integrates to 1.

In fact, if $y > 1$, then

$$F_Y(y) = \int_0^{1+\sqrt{y}} 3e^{-3x}\,dx$$

and

$$f_Y(y) = \frac{3e^{-3(1+\sqrt{y})}}{2\sqrt{y}}.$$

Another way to see the above is to write

$$x = \begin{cases} 1 - \sqrt{y} \text{ or } 1 + \sqrt{y} & \text{if } 0 < y < 1, \\ 1 + \sqrt{y} & \text{if } y > 1. \end{cases}$$

(The second solution is rejected when $y \geqslant 1$ because X is an exponential and is thus always nonnegative.) Now,

$$f_Y(y) = \begin{cases} f_X(1 - \sqrt{y})|(1 - \sqrt{y})'| + f_X(1 + \sqrt{y})|(1 + \sqrt{y})'| & \text{if } 0 < y < 1, \\ f_X(1 + \sqrt{y})|(1 + \sqrt{y})'| & \text{if } y > 1. \end{cases}$$

Finish the computation and check you get the same answer as before.

Example 18.8. Another common transformation is $g(x) = |x|$. In this case, let $Y = |X|$ and note that if $y > 0$, then

$$F_Y(y) = P\{-y < X < y\} = F_X(y) - F_X(-y).$$

Else, $F_Y(y) = 0$. Therefore,

$$f_Y(y) = \begin{cases} f_X(y) + f_X(-y) & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0. \end{cases}$$

For instance, if X is standard normal, then

$$f_{|X|}(y) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-y^2/2} & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0. \end{cases}$$

Or if X is Cauchy, then

$$f_{|X|}(y) = \begin{cases} \frac{2}{\pi}\,\frac{1}{1+y^2} & \text{if } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Can you guess $f_{|X|}$ when X is uniform $(-1\,,1)$?

Example 18.9. As you can see, it is best to try to work on these problems on a case-by-case basis. Here is another example where you need to do that. Let Θ be uniformly distributed between $-\pi/2$ and $\pi/2$. Let $Y = \tan\Theta$. Geometrically, Y is obtained by picking a line, in the xy-plane, passing through the origin so that the angle of this line with the x-axis is uniformly distributed. The y-coordinate of the intersection between this line and the line $x = 1$ is our random variable Y. What is the pdf of Y? The transformation is $y = \tan\theta$ and thus the pdf of Y is

$$f_Y(y) = f_\Theta(\arctan(y))\,|\arctan'(y)| = \frac{1}{\pi(1+y^2)}.$$

That is, Y is Cauchy distributed.

Homework Problems

Exercise 18.1. Let X be a uniform random variable on $[-1, 1]$. Let $Y = e^{-X}$. What is the probability density function of Y ?

Exercise 18.2. Let X be an exponential random variable with parameter $\lambda > 0$. What is the probability density function of $Y = X^2$?

Exercise 18.3. Solve the following.

(a) (Log-normal distribution) Let X be a standard normal random variable. Find the probability density function of $Y = e^X$.

(b) Let X be a standard normal random variable and Z a random variable solution of $Z^3 + Z + 1 = X$. Find the probability density function of Z.

Exercise 18.4. Solve the following.

(a) Let X be an exponential random variable with parameter $\lambda > 0$. Find the probability density function of $Y = \ln(X)$.

(b) Let X be a standard normal random variable and Z a random variable with values in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ solution of $Z + \tan(Z) = X$. Find the density function of Z.

Exercise 18.5. Let X be a continuous random variable with probability density function given by $f_X(x) = \frac{1}{x^2}$ if $x \geqslant 1$ and 0 otherwise. A random variable Y is given by

$$Y = \begin{cases} 2X & \text{if } X \geqslant 2, \\ X^2 & \text{if } X < 2. \end{cases}$$

Find the probability density function of Y.

Exercise 18.6. Solve the following.

(a) Let f be the probability density function of a continuous random variable X. Find the probability density function of $Y = X^2$.

(b) Let X be a standard normal random variable. Show that $Y = X^2$ has a Gamma distribution and find the parameters.

Exercise 18.7. We throw a ball from the origin with velocity v_0 and an angle θ with respect to the x-axis. We assume v_0 is fixed and θ is uniformly distributed on $[0, \frac{\pi}{2}]$. We denote by R the distance at which the object *lands*, i.e. hits the x-axis again. Find the probability density function of R. *Hint :* we remind you that the laws of mechanics show that the distance is given by $R = \frac{v_0^2 \sin(2\theta)}{g}$, where g is the gravity constant.

Lecture 19

1. Generating random variables from Uniform$(0,1)$

Theorem 19.1. *If X is any random variable with a continuous CDF $F(x)$, then $U = F(X) \sim \text{Uniform}(0,1)$.*

Proof. Clearly, $0 \leqslant U \leqslant 1$. So $F(u) = 0$ for $u < 0$ and $F(u) = 1$ for $u \geqslant 1$.

If F is one-to-one, then we can simply write, for $0 < u < 1$,

$$F_U(u) = P\{U \leqslant u\} = P\{F(X) \leqslant u\} = P\{X \leqslant F^{-1}(u)\} = F(F^{-1}(u)) = u.$$

This is the CDF of a Uniform(0,1). If F is not one-to-one, then one needs to be more careful.

Fix $u \in (0,1)$. Fix $\varepsilon > 0$ such that $\varepsilon < u$ and $\varepsilon < 1-u$. Then, $(u, u+\varepsilon) \subset (0,1)$ and, by continuity of F and the fact that F goes to 0 at $-\infty$ and to 1 at ∞, its graph must pass between u and $u+\varepsilon$; i.e. there exists a b such that $F(b) \in (u, u+\varepsilon]$. Similarly, there exists an a such that $F(a) \in [u-\varepsilon, u]$.

Now note that $X \leqslant a$ implies $F(X) \leqslant F(a)$, because F is nondecreasing. This then implies that $F(X) \leqslant u$. Thus,

$$\begin{aligned} F_U(u) &= P\{U \leqslant u\} = P\{F(X) \leqslant u\} \geqslant P\{F(X) \leqslant F(a)\} \\ &\geqslant P\{X \leqslant a\} = F(a) \geqslant u - \varepsilon. \end{aligned}$$

Moreover, since $F(X) \leqslant u$ implies $F(X) < F(b)$ which implies $X < b$, we have

$$\begin{aligned} F_U(u) &= P\{U \leqslant u\} = P\{F(X) \leqslant u\} \leqslant P\{F(X) < F(b)\} \\ &\leqslant P\{X < b\} = P\{X \leqslant b\} = F(b) \leqslant u + \varepsilon. \end{aligned}$$

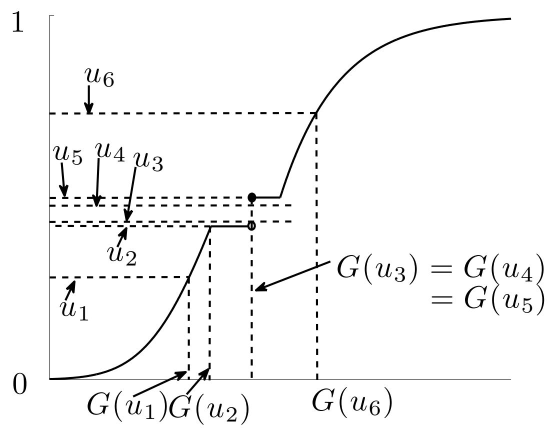


Figure 19.1. The function $G(u)$.

We have thus shown that $|F_U(u) - u| \leqslant \varepsilon$. Now take $\varepsilon \to 0$ to conclude that $F_U(u) = u$. □

It is noteworthy that the above does not work if F is not continuous. Take for example $X \sim \text{Bernoulli}(0.5)$. Then, $F(X) = F(0) = 0.5$, with probability 0.5, and $F(X) = F(1) = 1$ with probability 0.5. This is certainly not a $\text{Uniform}(0, 1)$!

The converse to the above theorem is quite useful.

Theorem 19.2. *Let F be a strictly increasing continuous function such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Let $U \sim \text{Uniform}(0, 1)$. Then, $X = F^{-1}(U)$ has CDF $F(x)$.*

Proof. Simply write

$$P\{X \leqslant x\} = P\{F^{-1}(U) \leqslant x\} = P\{U \leqslant F(x)\} = F(x). \qquad \square$$

Example 19.3. To generate an exponential random variable with parameter λ we solve

$$u = F(x) = 1 - e^{-\lambda x}$$

to get $x = -\frac{\log(1-u)}{\lambda}$. Thus, $-\lambda^{-1} \log(1-U)$ has an exponential distribution with parameter λ, where $U \sim \text{Uniform}(0, 1)$. [In this special case, $1 - U$ is also $\text{Uniform}(0, 1)$, and thus we can use $-\lambda^{-1} \log U$ as well.]

In fact, we can prove a much more general version of Theorem 19.2.

Theorem 19.4. *Let F be any nondecreasing right-continuous function such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$; i.e. F is any candidate for a CDF.*

Define

$$G(u) = \inf\{x : u \leqslant F(x)\} = \min\{x : u \leqslant F(x)\};$$

see Figure 19.1. [Note that $G(u) = F^{-1}(u)$ *wherever the latter exists.] Let* $U \sim \text{Uniform}(0,1)$. *Then,* $X = G(U)$ *has CDF* $F(x)$.

Proof. First, let us explain why the infimum in the definition of G is attained (and is thus a minimum). This is a consequence of the right-continuity of F. Indeed, if $x \searrow G(u)$ with $F(x) \searrow u$, then $F(G(u)) = u$.

One consequence of the above equation is that

$$P\{X \leqslant a\} = P\{G(U) \leqslant a\} \leqslant P\{F(G(U)) \leqslant F(a)\} = P\{U \leqslant F(a)\} = F(a).$$

Next, we observe that the definition of G implies that if $u \leqslant F(a)$, then $G(u) \leqslant a$. Thus,

$$P\{X \leqslant a\} = P\{G(U) \leqslant a\} \geqslant P\{U \leqslant F(a)\} = F(a).$$

We conclude that $P\{X \leqslant a\} = F(a)$, which means that X has CDF F. □

This theorem allows us to generate any random variable we can compute the CDF of, if we simply have a random number generator that generates numbers between 0 and 1 "equally likely."

Example 19.5. How do we flip a coin that gives heads with probability 0.6, using the random number generator on our calculator? The intuitive answer is: generate a number and call it tails if it is less than 0.4 and heads otherwise. Does the above theorem give the same answer?

Since the CDF of a Bernoulli(0.6) is not one-to-one, we need to compute G. This turns out not to be too hard. Recall that

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.4 & \text{if } 0 \leqslant x < 1, \\ 1 & \text{if } x \geqslant 1. \end{cases}$$

Then,

$$G(u) = \begin{cases} 0 & \text{if } 0 \leqslant u \leqslant 0.4, \\ 1 & \text{if } 0.4 < u \leqslant 1. \end{cases}$$

Just as our intuition had indicated.

Notice, by the way, that the above shows that one can start with a continuous random variable and transform it into a discrete random variable! Of course, the transformation G is not continuous.

Lecture 20

1. Mathematical Expectation: Discrete random variables

The *mathematical expectation* (or just the expectation, or mean, or average) $E[X]$ of a discrete random variable X with mass function f is defined formally as the average of the possible values of X, weighted by their corresponding probabilities:

$$E[X] = \sum_x x\, f(x). \tag{20.1}$$

When X has finitely many possible values the above sum is well defined. It corresponds to the physical notion of center of gravity of point masses placed at positions x with weights $f(x)$.

Example 20.1. We toss a fair coin and win \$1 for heads and lose \$1 for tails. This is a fair game since the average winnings equal \$0. Mathematically, if X equals the amount we won, then $E[X] = 1 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 0$.

Example 20.2. We roll a die that is loaded as follows: it comes up 6 with probability 0.4, 1 with probability 0.2, and the rest of the outcomes come up with probability 0.1 each. Say we lose \$2 if the die shows a 2, 3, 4, or 5, while we win \$1 if it shows a 1 and \$2 if it shows a 6. On average we win

$$-2 \times 4 \times .1 + 1 \times 0.2 + 2 \times 0.4 = 0.2;$$

that is we win 20 cents. In a simple case like this one, where X has a finite amount of possible values, one can use a table:

x	-2	1	2
$f(x) = P\{X = x\}$	4×0.1	0.2	0.4
$xf(x)$	-0.8	0.2	0.8

$E[X] = 0.2$ is then the sum of the elements in the last row. Intuitively, this means that if we play, say, 1000 times, we expect to win about \$200. Making this idea more precise is what we mean by "connecting the mathematical and the intuitive definitions of probability." This also gives a fair price to the game: 20 cents is a fair participation fee for each attempt.

Example 20.3. You role a fair die and lose as many dollars as pips shown on the die. Then, you fairly toss an independent fair coin a number of times equal to the outcome of the die. Each head wins you \$2 and each tail loses you \$1. Is this a winning or a losing game? Let X be the amount of dollars you win after having played the game. Let us compute the average winning. First, we make a table of all the outcomes.

Outcome	1H	1T	2H	1H1T	2T
x	$-1+2$	$-1-1$	$-2+4$	$-2+2-1$	$-2-2$
$f(x)$	$\frac{1}{6}\times\frac{1}{2}$	$\frac{1}{6}\times\frac{1}{2}$	$\frac{1}{6}\times\frac{1}{4}$	$2\times\frac{1}{6}\times\frac{1}{4}$	$\frac{1}{6}\times\frac{1}{4}$
Outcome	3H	2H1T	1H2T	3T	4H
x	$-3+6$	$-3+4-1$	$-3+2-2$	$-3-3$	$-4+8$
$f(x)$	$\frac{1}{6}\times\frac{1}{8}$	$3\times\frac{1}{6}\times\frac{1}{8}$	$3\times\frac{1}{6}\times\frac{1}{8}$	$\frac{1}{6}\times\frac{1}{8}$	$\frac{1}{6}\times\frac{1}{16}$
Outcome	3H1T	2H2T	1H3T	4T	5H
x	$-4+6-1$	$-4+4-2$	$-4+2-3$	$-4-4$	$-5+10$
$f(x)$	$4\times\frac{1}{6}\times\frac{1}{16}$	$6\times\frac{1}{6}\times\frac{1}{16}$	$4\times\frac{1}{6}\times\frac{1}{16}$	$\frac{1}{6}\times\frac{1}{16}$	$\frac{1}{6}\times\frac{1}{32}$
Outcome	4H1T	3H2T	2H3T	1H4T	5T
x	$-5+8-1$	$-5+6-2$	$-5+4-3$	$-5+2-4$	$-5-5$
$f(x)$	$5\times\frac{1}{6}\times\frac{1}{32}$	$10\times\frac{1}{6}\times\frac{1}{32}$	$10\times\frac{1}{6}\times\frac{1}{32}$	$5\times\frac{1}{6}\times\frac{1}{32}$	$\frac{1}{6}\times\frac{1}{32}$
Outcome	6H	5H1T	4H2T	3H3T	2H4T
x	$-6+12$	$-6+10-1$	$-6+8-2$	$-6+6-3$	$-6+4-4$
$f(x)$	$\frac{1}{6}\times\frac{1}{64}$	$6\times\frac{1}{6}\times\frac{1}{64}$	$15\times\frac{1}{6}\times\frac{1}{64}$	$20\times\frac{1}{6}\times\frac{1}{64}$	$15\times\frac{1}{6}\times\frac{1}{64}$
Outcome	1H5T	6T			
x	$-6+2-5$	$-6-6$			
$f(x)$	$6\times\frac{1}{6}\times\frac{1}{64}$	$\frac{1}{6}\times\frac{1}{64}$			

Then,

$$E[X] = \sum x\, f(x) = -\frac{7}{4} = -1.75.$$

In conclusion, the game is a losing game. In fact, I would only play if they pay me a dollar and 75 cents each time!

Example 20.4. If $X \sim \text{Bernoulli}(p)$, then $E[X] = p \times 1 + (1-p) \times 0 = p$. More generally, if $X \sim \text{Binomial}(n\,, p)$, then I claim that $E[X] = np$. Here is why:

$$
\begin{aligned}
E[X] &= \sum_{k=0}^{n} k \overbrace{\binom{n}{k} p^k (1-p)^{n-k}}^{f(k)} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j} (1-p)^{(n-1)-j} \\
&= np(p + (1-p))^{n-1} = np,
\end{aligned}
$$

thanks to the binomial theorem.

Lecture 21

1. Mathematical Expectation: Discrete random variables, continued

If X has infinitely-many possible values, then the sum in (20.1) must be defined. If $P\{X \geqslant 0\} = 1$ (i.e. all possible values of X are nonnegative), then the sum in question is that of nonnegative numbers and is thus always defined [though could be ∞]. Similarly, if $P\{X \leqslant 0\} = 1$, then the sum is that of nonpositive numbers and is always defined [though could be $-\infty$].

Example 21.1. Suppose $X \sim \text{Poisson}(\lambda)$. Then, I claim that $E[X] = \lambda$. Indeed,

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

because $e^\lambda = \sum_{j=0}^\infty \lambda^j/j!$, thanks to Taylor's expansion. So when modeling the length of a waiting line, the parameter λ is the average length of the line.

Example 21.2. Suppose X is negative binomial with parameters r and p. Then, $E[X] = r/p$ because

$$
\begin{aligned}
E[X] &= \sum_{k=r}^{\infty} k\binom{k-1}{r-1} p^r(1-p)^{k-r} \\
&= \sum_{k=r}^{\infty} \frac{k!}{(r-1)!(k-r)!} p^r(1-p)^{k-r} \\
&= r\sum_{k=r}^{\infty} \binom{k}{r} p^r(1-p)^{k-r} \\
&= \frac{r}{p}\sum_{k=r}^{\infty} \binom{k}{r} p^{r+1}(1-p)^{(k+1)-(r+1)} \\
&= \frac{r}{p}\sum_{j=r+1}^{\infty} \underbrace{\binom{j-1}{(r+1)-1} p^{r+1}(1-p)^{j-(r+1)}}_{P\{\text{Negative binomial }(r+1,p)=j\}} \\
&= \frac{r}{p}.
\end{aligned}
$$

Thus, for example, $E[\text{Geometric}(p)] = 1/p$.

Example 21.3 (St.-Petersbourg paradox)**.** Here is an example of a random variable with infinite expectation. Let us toss a fair coin until we get heads. The first toss wins us \$2, and then each consecutive toss doubles the winnings. So if X is the amount we win, then it has the mass function $f(2^n) = 1/2^n$, for $n \geqslant 1$; i.e. $X = 2^n$ with probability $1/2^n$. This is a nonnegative random variable and thus the expectation must be defined. However, $2^n f(2^n) = 1$ for all $n \geqslant 1$. Thus, the sum of these terms is indeed infinite. This means that the game has an infinite price and you should be willing to play regardless of the fee. The paradox is that this contradicts our instincts. For example, if you are asked to pay \$4 to play the game, then you will probably agree since all you need is to get tails on your first toss, which you assess as being quite likely. On the other hand, if you are asked to pay \$32, then you may hesitate. In this case, you need to get 4 tails in a row to break even, which you estimate as being quite unlikely. But what if you get 5 tails in a row? Then you get the \$32 back and get \$32 more! This is what is hard to grasp. The unrealistic part of this paradox is that it assumes the bank has infinite supplies and that in the unlikely event of you getting 265 tails in a row, they will have $\$2^{266}$ to pay you! (This is more than 10^{80}, which is the estimated number of atoms in the observable universe!)

If X has infinitely-many possible values but can take both positive and negative values, then we have to be careful with the definition of the sum $E[X] = \sum x\, f(x)$. We can always add the positive and negative parts separately. So, formally, we can write

$$E[X] = \sum_{x>0} x\, f(x) + \sum_{x<0} x\, f(x).$$

Now, we see that if one of these two sums is finite then, even if the other were infinite, $E[X]$ would be well defined. Moreover, $E[X]$ is finite if, and only if, both sums are finite; i.e. if

$$\sum |x|\, f(x) < \infty.$$

Example 21.4. Say X has the mass function $f(2^n) = f(-2^n) = 1/2^n$, for $n \geqslant 2$. (Note that the probabilities do add up to one: $2\sum_{n\geqslant 2} \frac{1}{2^n} = 1$.) Then, the positive part of the sum gives

$$\sum_{n\geqslant 2} 2^n \times \frac{1}{2^n} = \infty,$$

and the negative part of the sum gives

$$\sum_{n\geqslant 2} (-2^n) \times \frac{1}{2^n} = -\infty.$$

This implies that $E[X]$ is not defined. In fact, if we compute

$$\sum_{n=2}^{N} 2^n \times \frac{1}{2^n} = N - 1 \text{ and } \sum_{n=2}^{M} (-2^n) \times \frac{1}{2^n} = -M + 1,$$

then, in principle, to get $E[X]$ we need to add the two and take N and M to infinity. But we now see that the sum equals $N - M$ and so depending on how we take N and M to infinity, we get any value we want for $E[X]$. Indeed, if we take $N = 2M$, then we get ∞. If we take $M = 2N$ we get $-\infty$. And if we take $N = M + a$, we get a, for any integer a.

Homework Problems

Exercise 21.1. In Las Vegas, a roulette is made of 38 boxes, namely 18 black boxes, 18 red boxes, a box '0' and a box '00'. If you bet \$1 on 'black', you get \$2 if the ball stops in a black box and \$0 otherwise. Let X be your profit. Compute $E[X]$.

Exercise 21.2. In the game *Wheel of Fortune*, you have 52 possible outcomes: one "0", one "00", two "20", four "10", seven "5", fifteen "2" et twenty-two "1". If you bet \$1 on some number, you receive this amount of money if the wheel stops on this number. If you bet \$1 on "0" or "00", you receive \$40 if the wheel stops on this number.

(a) Assume you bet \$1 on each of the seven possible numbers or symbols (for a total of \$7), what is the expectation of your profit?

(b) Assume you want to bet \$1 on only one of these numbers or symbols, which has the best (resp. worst) profit expectation?

Exercise 21.3. Let X be a Geometric r.v. with parameter $p \in [0, 1]$. Compute $E[X]$.

Lecture 22

1. Mathematical Expectation: Continuous random variables

When X is a continuous random variable with density $f(x)$, we can repeat the same reasoning as for discrete random variables and obtain the formula

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

The same issues as before arise: if $\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty$, then the above integral is well defined and finite. If, on the other hand, $\int_0^{\infty} x\, f(x)\, dx < \infty$ but $\int_{-\infty}^{0} x\, f(x)\, dx = -\infty$, then the integral is again defined but equals $-\infty$. Conversely, if $\int_0^{\infty} x\, f(x)\, dx = \infty$ but $\int_{-\infty}^{0} x\, f(x)\, dx > -\infty$, then $E[X] = \infty$. Finally, if both integrals are infinite, then $E[X]$ is not defined.

Example 22.1 (Uniform)**.** Suppose X is uniform on $(a\,, b)$. Then,

$$E[X] = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{2}\frac{b^2 - a^2}{b-a} = \frac{1}{2}\frac{(b-a)(b+a)}{b-a} = \frac{b+a}{2}.$$

N.B.: The formula of the first example on page 303 of Stirzaker's text is wrong.

Example 22.2 (Gamma)**.** If X is Gamma(α, λ), then for all positive values of x we have $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, and $f(x) = 0$ for $x < 0$. Therefore,

$$\begin{aligned}
E[X] &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x}\, dx \\
&= \frac{1}{\lambda \Gamma(\alpha)} \int_0^\infty z^\alpha e^{-z}\, dz \qquad (z = \lambda x) \\
&= \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \\
&= \frac{\alpha}{\lambda}.
\end{aligned}$$

In the special case that $\alpha = 1$, $\frac{1}{\lambda}$ is the expectation of an exponential random variable with parameter λ. So when modeling a waiting time, the parameter of the exponential is *one over the average waiting time*. The parameter λ is thus equal to the serving rate: the number of people served per unit time. Now, you should understand a bit better the derivation of the exponential distribution that came after Exercise 15.2. Namely, if we recall from Exercise 21.2 that the average of a geometric with parameter p is $1/p$, we see that if we use $p = \lambda/n$, the average will be n/λ. If each coin flip takes $1/n$ seconds, then the average serving time is $1/\lambda$, as desired.

Another observation is that E[Gamma(α, λ)] $= \alpha/\lambda$ the same way as E[Negative Binomial(r, p)] $= r/p$. This is not a coincidence and one can derive the Gamma distribution from the negative binomial similarly to how the exponential was derived from a geometric.

Example 22.3 (Normal)**.** Suppose $X \sim N(\mu, \sigma^2)$; i.e. $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Then,

$$\begin{aligned}
E[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^\infty x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty (\mu + \sigma z) e^{-z^2/2}\, dz \qquad (z = (x-\mu)/\sigma) \\
&= \mu \underbrace{\int_{-\infty}^\infty \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz}_{1} + \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^\infty z e^{-z^2/2}\, dz}_{0,\text{ by symmetry}} \\
&= \mu.
\end{aligned}$$

Example 22.4 (Cauchy)**.** In this example, $f(x) = \pi^{-1}(1+x^2)^{-1}$. Note that the expectation is defined only if the following limit exists regardless of how we let n and m tend to ∞:

$$\frac{1}{\pi^2} \int_{-m}^n \frac{y}{1+y^2}\, dy.$$

Now I argue that the limit does not exist; I do so by showing two different choices of $(n\,,m)$ which give rise to different limiting "integrals."

Suppose $m = e^{\pi^2 a} n$, for some fixed number a. Then,

$$\begin{aligned}
\frac{1}{\pi^2}\int_{-e^{-\pi^2 a}n}^{n} \frac{y}{1+y^2}\,dy &= \frac{1}{\pi^2}\int_0^n \frac{y}{1+y^2}\,dy - \frac{1}{\pi^2}\int_0^{e^{-\pi^2 a}n} \frac{y}{1+y^2}\,dy \\
&= \frac{1}{2\pi^2}\int_1^{1+n^2} \frac{dz}{z} - \frac{1}{2\pi^2}\int_1^{1+e^{-2\pi^2 a}n^2} \frac{dz}{z} \qquad (z = 1+y^2) \\
&= \frac{1}{2\pi^2}\ln\left(\frac{1+n^2}{1+e^{-2\pi^2 a}n^2}\right) \\
&\to \frac{1}{2\pi^2}\ln e^{2\pi^2 a} = a \quad \text{as } n\to\infty.
\end{aligned}$$

Thus, we can make the limit converge to any number a we want. In fact, taking $m = n^2$ and repeating the above calculation allows us to make the limit converge to $-\infty$, while taking $m = \sqrt{n}$ makes the limit equal to ∞. The upshot is that the Cauchy density does not have a well-defined expectation. [That is not to say that the expectation is well defined, but infinite.] In particular, we conclude that $E[|X|] = \infty$.

Theorem 22.5. *If X is a positive random variable with density f, then*

$$E[X] = \int_0^\infty P\{X > x\}\,dx = \int_0^\infty (1 - F(x))\,dx.$$

Proof. The second identity is a consequence of the fact that $1 - F(x) = P\{X > x\}$. In order to prove the first identity note that $P\{X > x\} = \int_x^\infty f(y)\,dy$. Therefore, if $A = \{(x,y) : y > x > 0\}$ then

$$\begin{aligned}
\int_0^\infty P\{X > x\}\,dx &= \int_0^\infty \left(\int_x^\infty f(y)\,dy\right) dx = \iint_A f(y)\,dx\,dy \\
&= \int_0^\infty f(y)\left(\int_0^y dx\right) dy = \int_0^\infty y\,f(y)\,dy \\
&= E[X].
\end{aligned}$$

□

If X is a negative random variable, then $-X$ is positive and we have

$$E[X] = -E[(-X)] = -\int_0^\infty P\{X < -x\}\,dx = -\int_{-\infty}^0 P\{X < x\}\,dx.$$

If X takes negative and positive values, and at least one of $\int_0^\infty P\{X > x\}\,dx$ and $\int_{-\infty}^0 P\{X < x\}\,dx$ is finite, then $E[X]$ equals their difference:

$$E[X] = \int_0^\infty P\{X > x\}\,dx - \int_{-\infty}^0 P\{X < x\}\,dx.$$

Note that the above formula does not involve the density function f. It turns out (and we omit the math) that we can define the expectation of any positive random variable (discrete, continuous, or other) using that formula. That is to say the notion of mathematical expectation (or average value) is general and applies to any real-valued random variable.

Lecture 23

1. Some properties of expectations

The following theorem is useful when computing averages of transformations of random variables.

Theorem 23.1. *If* X *has mass function* $f(x)$ *and* $\sum_x g(x)\,f(x)$ *is well defined, i.e.* $\sum_{x:g(x)\geqslant 0} g(x)\,f(x) < \infty$ *or* $\sum_{x:g(x)\leqslant 0} g(x)\,f(x) > -\infty$*, then*

$$E[g(X)] = \sum_x g(x)f(x).$$

Proof. Let $Y = g(X)$. Then, by definition

$$\begin{aligned} E[g(X)] = E[Y] &= \sum_y y\,P\{Y = y\} = \sum_y y \sum_{x:g(x)=y} P\{X = x\} \\ &= \sum_y \sum_{x:g(x)=y} g(x)\,P\{X = x\} = \sum_x g(x)\,P\{X = x\} \end{aligned}$$

as desired. □

Now, we can prove the following natural properties.

Theorem 23.2. *Let* X *and* Y *be any random variables (discrete, continuous, or other) with well defined expectations* $E[X]$ *and* $E[Y]$*. Let* a *be any (nonrandom) number. Then:*

(1) $E[aX] = aE[X]$;
(2) *If either* $E[X]$ *or* $E[Y]$ *is finite, or if they are both* ∞ *or both* $-\infty$*, then* $E[X + Y] = E[X] + E[Y]$.
(3) $E[a] = a$;

(4) *If* $P\{X \leqslant Y\} = 1$, *then* $E[X] \leqslant E[Y]$;

(5) *If* $P\{X \geqslant 0\} = 1$ *and* $E[X] = 0$, *then* $P\{X = 0\} = 1$;

(6) *If* X *and* Y *are independent and are either both nonnegative, both nonpositive, or both have finite expectations, then* $E[XY] = E[X]E[Y]$.

Proof. We show the proofs in the discrete case. The proofs in the continuous case are similar, and the proofs in the general case are omitted. To prove (1) let $x_1, x_2, \ldots$ be the possible values of X. Then, $ax_1, ax_2, \ldots$ are the possible values of aX. Moreover,

$$E[aX] = \sum_i ax_i f(x_i) = a \sum_i x_i f(x_i) = aE[X].$$

Let us now prove (2). We will only treat the case where both variables are nonnegative. Let $x_1, x_2, \ldots$ be the possible (nonnegative) values of X and $y_1, y_2, \ldots$ the possible (nonnegative) values of Y. Then, the possible values of $X + Y$ are $\{x_i + y_j : i = 1, 2, \ldots,\ j = 1, 2, \ldots\}$ and are nonnegative. Thus,

$$\begin{aligned} E[X+Y] &= \sum_{i,j} (x_i + y_j) P\{X = x_i, Y = y_j\} \\ &= \sum_{i,j} x_i P\{X = x_i, Y = y_j\} + \sum_{i,j} y_j P\{X = x_i, Y = y_j\} \\ &= \sum_i x_i \sum_j P\{X = x_i, Y = y_j\} + \sum_j y_j \sum_i P\{X = x_i, Y = y_j\} \\ &= \sum_i x_i P\{X = x_i\} + \sum_j y_j P\{Y = y_j\} \\ &= E[X] + E[Y]. \end{aligned}$$

In the second-to-last equality we used the fact that the sets $\{X = x_i\}$ are disjoint and their union is everything, and the same for the sets $\{Y = y_j\}$.

If now X and Y are both nonpositive, then $-X$ and $-Y$ are nonnegative and we can use property (1) to write

$$E[X] + E[Y] = -(E[-X] + E[-Y]) = -E[-(X+Y)] = E[X+Y].$$

If at least one of the variables takes positive and negative values, then one needs to use slightly more involved arguments requiring facts about infinite series. We omit the proof in this case.

Next, we prove (3). The only value the random variable a takes is a and it takes it with probability 1. Thus, its mathematical expectation simply equals a itself. To prove (4) observe that $Y - X$ is a nonnegative random variable; i.e. its possible values are all nonnegative. Thus, it has a nonnegative average. But by (1) and (2) we have $0 \leqslant E[Y - X] = E[Y] + E[-X] = E[Y] - E[X]$. Property (5) is obvious since if there existed an $x_0 > 0$

for which $f(x_0) > 0$, then we would have had $E[X] = \sum xf(x) \geqslant x_0 f(x_0) > 0$, since the sum is over $x \geqslant 0$.

Finally, we prove (6). Again, let x_i and y_j be the possible values of X and Y, respectively. Then, the possible values of XY are given by the set $\{x_i y_j : i = 1, 2, \ldots,\ j = 1, 2, \ldots\}$. Thus,

$$\begin{aligned} E[XY] &= \sum_{i,j} x_i y_j P\{X = x_i, Y = y_j\} \\ &= \sum_{i,j} x_i y_j P\{X = x_i\} P\{Y = y_j\} \quad \text{(by independence)} \\ &= \sum_i x_i P\{X = x_i\} \sum_j y_j P\{Y = y_j\} \\ &= E[X]E[Y]. \end{aligned}$$

The third equality was simply the result of summing over j first and then over i. We can sum in any order because the terms are either of the same sign, or are summable (if the expectations of X and Y are finite). $\square$

As a consequence of the above theorem we have that $E[aX + b] = aE[X] + b$ for all constants a and b. Also, if $P\{X = Y\} = 1$ (i.e. X and Y are *almost-surely* equal), then $E[X] = E[Y]$.

Example 23.3. If $X \sim \text{Binomial}(n, p)$, then we found in Example 20.4 that $E[X] = np$. Here is a quick way to recover this. Recall that if $B_1, \ldots, B_n$ are independent Bernoulli random variables with parameter p, then $X = B_1 + \cdots + B_n \sim \text{Binomial}(n, p)$. Now, recall that $E[\text{Bernoulli}(p)] = p$ and apply property (2) in Theorem 23.2 n times to get that $E[X] = E[B_1] + \cdots + E[B_n] = np$.

Example 23.4 (Bernoulli random variables)**.** Suppose $X \sim \text{Bernoulli}(p)$. Recall that $E[X] = (1-p) \times 0 + p \times 1 = p$. Now let us compute $E[X^2]$:

$$E[X^2] = (1-p) \times 0^2 + p \times 1^2 = p.$$

Two observations:

(1) This is obvious because $X = X^2$ *in this particular example*; and
(2) $E[X^2] \neq (E[X])^2$. In fact, the difference between $E(X^2)$ and $(EX)^2$ is an important quantity, called the *variance of* X. We will return to this topic later.

Example 23.5. If $X = \text{Binomial}(n, p)$, then what is $E[X^2]$? It may help to recall that $E[X] = np$. We have

$$E[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \,\frac{n!}{(k-1)!(n-k)!}\, p^k (1-p)^{n-k}.$$

The question is, "how do we reduce the factor k further"? If we had $k-1$ instead of k, then this would be easy to answer. So let us first solve a related problem.

$$\begin{aligned}
E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} k(k-1)\ \frac{n!}{k!(n-k)!}\ p^k(1-p)^{n-k} \\
&= n(n-1)\sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!\big([n-2]-[k-2]\big)!}\ p^k q^{n-k} \\
&= n(n-1)\sum_{k=2}^{n} \binom{n-2}{k-2} p^k(1-p)^{n-k} \\
&= n(n-1)p^2\sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{[n-2]-[k-2]} \\
&= n(n-1)p^2\sum_{\ell=0}^{n-2} \binom{n-2}{\ell} p^{\ell}(1-p)^{[n-2]-\ell}.
\end{aligned}$$

The summand is the probability that Binomial$(n-2\,,p)$ is equal to ℓ. Since that probability is added over all of its possible values, the sum is one. Thus, we obtain $E[X(X-1)] = n(n-1)p^2$. But $X(X-1) = X^2 - X$. Therefore, we can apply Theorem 23.2 to find that

$$E[X^2] = E[X(X-1)] + E[X] = n(n-1)p^2 + np = (np)^2 + np(1-p).$$

Example 23.6. Suppose $X \sim$ Poisson(λ). We saw in Example 21.1 that $E[X] = \lambda$. In order to compute $E[X^2]$, we first compute $E[X(X-1)]$ and find that

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

The sum is equal to one; change variables ($j = k-2$) and recognize the jth term as the probability that Poisson$(\lambda) = j$. Therefore,

$$E[X(X-1)] = \lambda^2.$$

Because $X(X-1) = X^2 - X$, the left-hand side is $E[X^2] - E[X] = E[X^2] - \lambda$. Therefore,

$$E[X^2] = \lambda^2 + \lambda.$$

Homework Problems

Exercise 23.1. Let X be an Exponential r.v. with parameter $\lambda > 0$. Compute $E[X]$ and $E[X^2]$.

Exercise 23.2. Let X be a random variable with $N(0,1)$ distribution. Show that

$$E[X^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ (n-1)(n-3)\cdots 3\cdot 1 & \text{if } n \text{ is even.} \end{cases}$$

Exercise 23.3. We assume that the length of a telephone call is given by a random variable X with probability density function

$$f(x) = \begin{cases} xe^{-x} & \text{if } x \geqslant 0 \\ 0 & \text{otherwise.} \end{cases}$$

The cost of a call is given as a function of the length by

$$c(X) = \begin{cases} 2 & \text{if } 0 < X \leqslant 3 \\ 2 + 6(X-3) & \text{if } X > 3. \end{cases}$$

Find the average cost of a call.

Lecture 24

1. Variance

When E[X] is well-defined, the *variance* of X is defined as

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

If $\mathrm{E}[X] = \infty$ or $-\infty$ the above is just infinite and does not carry any information. Thus, the variance is a useful notion when E[X] is finite. The next theorem says that this is the same as asking for E[|X|] to be finite. (Think of absolute summability or absolute integrability in calculus.)

Theorem 24.1 (Triangle inequality)**.** E[X] *is well defined and finite if, and only if,* $\mathrm{E}[|X|] < \infty$. *In that case,*

$$|\mathrm{E}[X]| \leqslant \mathrm{E}[|X|].$$

Proof. Observe that $-|X| \leqslant X \leqslant |X|$ and apply (4) of Theorem 23.2. □

This of course makes sense: the average of X must be smaller than the average of |X|, since there are no cancellations when averaging the latter.

Note that the triangle inequality that we know is a special case of the above: $|a + b| \leqslant |a| + |b|$. Indeed, let X equal a or b, equally likely. Now apply the above theorem and see what happens!

Thus, when E[|X|] is finite:

(1) We predict the as-yet-unseen value of X by the nonrandom number E[X] (its average value);
(2) Var(X) is the expected squared-error in this prediction. Note that Var(X) is also a nonrandom number.

The variance measures the amount of variation in the random variable. It vanishes if, and only if, there is no variation at all.

Theorem 24.2. *If* $\mathrm{Var}(X) = 0$, *then* X *is almost-surely constant. That is, there exists a constant* m *such that* $P\{X = m\} = 1$.

Proof. The constant m has to be the average of X. So we will prove that if the variance vanishes, then $P\{X = E[X]\} = 1$. But this follows from property (5) in Theorem 23.2. Indeed, $0 = \mathrm{Var}(X) = E[(X - E[X])^2]$ implies that $P\{(X - E[X])^2 = 0\} = 1$. This is what we wanted to prove. □

Here are some useful (and natural) properties of the variance.

Theorem 24.3. *Let* X *be such that* $E[X^2] < \infty$ *and let* a *be a nonrandom number.*

(1) $\mathrm{Var}(X) \geqslant 0$;
(2) $\mathrm{Var}(a) = 0$;
(3) $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$;
(4) $\mathrm{Var}(X + a) = \mathrm{Var}(X)$.

The proofs go by direct computation and are left to the student. Note that (2) says that nonrandom quantities have no variation. (4) says that shifting by a nonrandom amount does not change the amount of variation in the random variable.

Let us now compute the variance of a few random variables. But first, here is another useful way to write the variance

$$E\left[(X - E[X])^2\right] = E\left[X^2 - 2XE[X] + (E[X])^2\right] = E[X^2] - 2E[X]E[X] + (E[X])^2$$
$$= E[X^2] - (E[X])^2.$$

Example 24.4. We have seen in the previous lecture that if $X \sim \mathrm{Poisson}(\lambda)$, then $E[X^2] = \lambda^2 + \lambda$. We have also seen in Example 21.1 that $E[X] = \lambda$. Thus, in this case, $\mathrm{Var}(X) = \lambda$.

Example 24.5. Suppose $X \sim \mathrm{Bernoulli}(p)$. Then, $X^2 = X$ and $E[X^2] = E[X] = p$. But then, $\mathrm{Var}(X) = p - p^2 = p(1 - p)$.

Example 24.6. If $X = \mathrm{Binomial}(n\,, p)$, then what is $\mathrm{Var}(X)$? We have seen that $E[X] = np$ and $E[X^2] = (np)^2 + np(1 - p)$. Therefore, $\mathrm{Var}(X) = np(1 - p)$.

It is not a coincidence that the variance of $\mathrm{Binomial}(n, p)$ is n times the variance of $\mathrm{Bernoulli}(p)$. It is a consequence of the following fact.

Theorem 24.7. *Let* X *and* Y *be two independent random variables with both* $E[|X|] < \infty$ *and* $E[|Y|] < \infty$. *Then* $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. Observe first that $E[|XY|] = E[|X|]E[|Y|] < \infty$. Thus by the triangle inequality $E[XY]$ is well defined and finite. Now, the proof of the theorem follows by direct computation:

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

In the third equality we used property (6) in Theorem 23.2. □

Example 24.8. Since a Binomial(n, p) is the sum of n independent Bernoulli(p), each of which has variance $p(1-p)$, the variance of a Binomial(n, p) is simply $np(1-p)$, as already observed by direct computation.

Example 24.9. Suppose $X \sim$ Geometric(p) distribution. We have seen already that $E[X] = 1/p$ (Example 21.2). Let us find a new computation for this fact, and then go on and find also the variance.

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

In the above computation, we used that the derivative of the sum is the sum of the derivatives. This is OK when we have finitely many terms. Since we have infinitely many terms, one does need a justification that comes from facts in real analysis. We will overlook this issue...

Next we compute $E[X^2]$ by first finding

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

Because $E[X(X-1)] = E[X^2] - E[X] = E[X^2] - (1/p)$, this proves that

$$E[X^2] = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2}.$$

Consequently,

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

For a different solution, see Example (13) on page 124 of Stirzaker's text.

As a consequence of Theorem 24.7 we have the following.

Example 24.10. Let X be a negative binomial with parameters n and p. Then, we know that X is a sum of n independent Geometric(p) random variables. We conclude that $\mathrm{Var}(X) = n(1-p)/p^2$. Can you do a direct computation to verify this?

Example 24.11 (Variance of Uniform$(a\,,b)$)**.** If X is Uniform$(a\,,b)$, then $E[X] = \frac{a+b}{2}$ and

$$E[X^2] = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{b^2+ab+a^2}{3}.$$

In particular, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Example 24.12 (Moments of $N(0\,,1)$)**.** Compute $E[X^n]$, where $X \sim N(0\,,1)$ and $n \geqslant 1$ is an integer:

$$\begin{aligned} E[X^n] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^n e^{-x^2/2}\,dx \\ &= 0 \quad \text{if } n \text{ is odd, by symmetry.} \end{aligned}$$

If n is even (or even when n is odd but we are computing $E[|X|^n]$ instead of $E[X^n]$), then

$$\begin{aligned} E[X^n] &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} x^n e^{-x^2/2}\,dx = \sqrt{\frac{2}{\pi}}\int_0^{\infty} x^n e^{-x^2/2}\,dx \\ &= \sqrt{\frac{2}{\pi}}\int_0^{\infty} (2z)^{n/2} e^{-z} \underbrace{\left((2z)^{-1/2}\,dz\right)}_{dx} \qquad \left(z = x^2/2 \;\Leftrightarrow\; x = \sqrt{2z}\right) \\ &= \frac{2^{n/2}}{\sqrt{\pi}}\int_0^{\infty} z^{(n-1)/2} e^{-z}\,dz \\ &= \frac{2^{n/2}}{\sqrt{\pi}}\Gamma\left(\frac{n}{2}+\frac{1}{2}\right) \\ &= \frac{2^{n/2}}{\sqrt{\pi}}\left(\frac{n}{2}-\frac{1}{2}\right)\left(\frac{n}{2}-\frac{3}{2}\right)\cdots\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma(1/2) \qquad \text{(Exercise 17.1)} \\ &= (n-1)(n-3)\cdots(5)(3)(1). \end{aligned}$$

Example 24.13. We can now compute the variance of a normal random variable with parameters μ and σ^2. Indeed,

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

Use the change of variable $z = (x-\mu)/\sigma$ to get $dx = \sigma dx$ and

$$\mathrm{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,dz = \sigma^2.$$

In the last step we used the previous exercise with $n = 2$ and recalled from Exercise 17.1 that $\Gamma(3/2) = \sqrt{\pi}/2$.

This is why one usually says that X is a normal random variable with mean μ and variance σ^2.

Example 24.14. Let $X \sim \mathrm{Gamma}(\alpha, \lambda)$. Then, we know $E[X] = \alpha/\lambda$. Now compute

$$\begin{aligned}
E[X^2] &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+1}e^{-\lambda x}\,dx \\
&= \frac{1}{\lambda^2\Gamma(\alpha)}\int_0^\infty z^{\alpha+1}e^{-z}\,dz \qquad (z = \lambda x) \\
&= \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} \\
&= \frac{(\alpha+1)\alpha}{\lambda^2}.
\end{aligned}$$

Thus,

$$\mathrm{Var}(X) = \frac{\alpha^2+\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

In particular, when $\alpha = 1$, we see that the variance of an exponential random variable with parameter λ equals $1/\lambda^2$.

Example 24.15. If X is a Cauchy random variable, then we have seen that its mean is not well defined (Exercise 22.4). Thus, it does not make sense to talk about its variance. Furthermore, the first moment is infinite: $E[|X|] < \infty$. (Otherwise, the mean would be defined and finite.) In fact, a direct computation shows that all moments are infinite: $E[|X|^n] = \infty$ for all $n \geqslant 1$. Indeed,

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{|x|^n}{1+x^2}\,dx = \frac{2}{\pi}\int_0^\infty \frac{x^n}{1+x^2}\,dx \geqslant \frac{2}{\pi}\int_1^\infty\frac{x^n}{1+x^2}\,dx \geqslant \frac{1}{\pi}\int_1^\infty x^{n-2}\,dx = \infty.$$

(In the second inequality we used the fact that if $x \geqslant 1$ then $1 + x^2 \leqslant 2x^2$.)

Homework Problems

Exercise 24.1. Let X and Y be two independent random variables, each exponentially distributed with parameter $\lambda = 1$.

(a) Compute $E[XY]$.
(b) Compute $E[X - Y]$.
(c) Compute $E[|X - Y|]$.

Exercise 24.2. Let X be a Binomial random variable with parameters n and p. Compute $E[X^2]$ and $E(X^2) - E(X)$.

Exercise 24.3. Let X be a Geometric random variable with parameter p. Compute $E[X^2]$ and $E(X^2) - E(X)$.

Exercise 24.4. Let X be uniformly distributed on $[0, 2\pi]$. Let $Y = \cos(X)$ and $Z = \sin(X)$. Prove that $E[YZ] = E[Y]E[Z]$ and $\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + \mathrm{Var}(Z)$. Then prove that Y and Z are not independent. This shows that the two equalities above *do not* imply independence.

Exercise 24.5. If X has the Poisson distribution with parameter λ, show that for any integer $k \geqslant 1$

$$E[X(X-1)(X-2)\cdots(X-k+1)] = \lambda^k.$$

Conclude that $E[X] = \mathrm{Var}(X) = \lambda$.

Exercise 24.6. If $E[X]$ exists, show that $|E[X]| \leqslant E[|X|]$.

Lecture 25

1. Joint distributions

If X and Y are two discrete random variables, then their *joint mass function* is

$$f(x\,,y) = P\{X = x\,,Y = y\}.$$

We might write $f_{X,Y}$ in place of f in order to emphasize the dependence on the two random variables X and Y.

Here are some properties of $f_{X,Y}$:

- $f(x\,,y) \geqslant 0$ for all x, y;
- $\sum_x \sum_y f(x\,,y) = 1$;
- $\sum_{(x,y)\in C} f(x\,,y) = P\{(X\,,Y) \in C\}$.

Example 25.1. You roll two fair dice. Let X be the number of 2s shown, and Y the number of 4s. Then X and Y are discrete random variables, and

$$\begin{aligned} f(x\,,y) &= P\{X = x\,,Y = y\} \\ &= \begin{cases} \frac{1}{36} & \text{if } x = 2 \text{ and } y = 0, \\ \frac{1}{36} & \text{if } x = 0 \text{ and } y = 2, \\ \frac{2}{36} & \text{if } x = y = 1, \\ \frac{8}{36} & \text{if } x = 0 \text{ and } y = 1, \\ \frac{8}{36} & \text{if } x = 1 \text{ and } y = 0, \\ \frac{16}{36} & \text{if } x = y = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Some times it helps to draw up a table of "joint probabilities":

$x \setminus y$	**0**	**1**	**2**
0	16/36	8/36	1/36
1	8/36	2/36	0
2	1/36	0	0

From this we can also calculate f_X and f_Y. For instance,

$$f_X(1) = P\{X = 1\} = f(1\,,0) + f(1\,,1) = \frac{10}{36}.$$

In general, you compute the row sums (f_X) and put them in the margin; you do the same with the column sums (f_Y) and put them in the bottom row. In this way, you obtain:

$x \setminus y$	**0**	**1**	**2**	f_X
0	16/36	8/36	1/36	25/36
1	8/36	2/36	0	10/36
2	1/36	0	0	1/36
f_Y	25/36	10/36	1/36	**1**

The "**1**" designates the right-most column sum (which should be one), and/or the bottom-row sum (which should also be one). This is also the sum of the elements of the table (which should also be one).

En route we have discovered the next result, as well.

Theorem 25.2. *For all* x, y*:*

(1) $f_X(x) = \sum_b f(x\,,b)$.
(2) $f_Y(y) = \sum_a f(a\,,y)$.

2. Independence

Definition 25.3. Let X and Y be discrete with joint mass function f. We say that X and Y are *independent* if for all x, y,

$$f_{X,Y}(x\,,y) = f_X(x) f_Y(y).$$

- Suppose A and B are two sets, and X and Y are independent. Then,

$$\begin{aligned} P\{X \in A, Y \in B\} &= \sum_{x \in A} \sum_{y \in B} f(x, y) \\ &= \sum_{x \in A} f_X(x) \sum_{y \in B} f_Y(y) \\ &= P\{X \in A\} P\{Y \in B\}. \end{aligned}$$

- Similarly, if h and g are functions, then $h(X)$ and $g(Y)$ are independent as well.
- All of this makes sense for more than 2 random variables as well.

Example 25.4 (Example 25.1, continued)**.** Note that in this example, X and Y are not independent. For instance,

$$f(1, 2) = 0 \neq f_X(1) f_Y(2) = \frac{10}{36} \times \frac{1}{36}.$$

Example 25.5. *Let* $X \sim \text{Geometric}(p_1)$ *and* $Y \sim \text{Geometric}(p_2)$ *be independent. What is the mass function of* $Z = \min(X, Y)$*?*

Let $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$ be the probabilities of failure. Recall from Lecture 10 that $P\{X \geqslant n\} = q_1^{n-1}$ and $P\{Y \geqslant n\} = q_2^{n-1}$ for all integers $n \geqslant 1$. Therefore,

$$\begin{aligned} P\{Z \geqslant n\} &= P\{X \geqslant n, Y \geqslant n\} = P\{X \geqslant n\} P\{Y \geqslant n\} \\ &= (q_1 q_2)^{n-1}, \end{aligned}$$

as long as $n \geqslant 1$ is an integer. Because $P\{Z \geqslant n\} = P\{Z = n\} + P\{Z \geqslant n+1\}$, for all integers $n \geqslant 1$,

$$\begin{aligned} P\{Z = n\} &= P\{Z \geqslant n\} - P\{Z \geqslant n+1\} = (q_1 q_2)^{n-1} - (q_1 q_2)^n \\ &= (q_1 q_2)^{n-1} (1 - q_1 q_2). \end{aligned}$$

Else, $P\{Z = n\} = 0$. Thus, $Z \sim \text{Geometric}(p)$, where $p = 1 - q_1 q_2$.

This makes sense: at each step we flip two coins and wait until the first time one of them comes up heads. In other words, we keep flipping as long as both coins land tails. Thus, this is the same as flipping one coin and waiting for the first time it comes up heads, as long as the probability of tails in this third coin is equal to the probability of both of the original coins coming up tails.

Homework Problems

Exercise 25.1. Let X and Y be two discrete random variables with joint mass function $f(x, y)$ given by

$x \mid y$	1	2
1	0.4	0.3
2	0.2	0.1

and $f(x, y) = 0$ otherwise.

(a) Determine if X and Y are independent.

(b) Compute $P(XY \leqslant 2)$.

Exercise 25.2. We roll two fair dice. Let X_1 (resp. X_2) be the smallest (resp. largest) of the two outcomes.

(a) What is the joint mass function of (X_1, X_2)?

(b) What are the probability mass functions of X_1 and X_2?

(c) Are X_1 and X_2 independent?

Exercise 25.3. We draw two balls with replacement out of an urn in which there are three balls numbered 2,3,4. Let X_1 be the sum of the outcomes and X_2 be the product of the outcomes.

(a) What is the joint mass function of (X_1, X_2)?

(b) What are the probability mass functions of X_1 and X_2?

(c) Are X_1 and X_2 independent?

Lecture 26

1. Jointly distributed continuous random variables

Definition 26.1. We say that (X, Y) is jointly distributed with *joint density function* f if f is piecewise continuous, and for all "nice" two-dimensional sets A,

$$P\{(X, Y) \in A\} = \iint_A f(x\,,y)\,dx\,dy.$$

If (X, Y) has a joint density function f, then:

(1) $f(x\,,y) \geqslant 0$ for all x and y;
(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x\,,y)\,dx\,dy = 1$.

For any function f of two variables that satisfies these properties, one can reverse engineer two random variables that will have f as their joint density function.

Example 26.2 (Uniform joint density)**.** Suppose E is a subset of the plane that has a well-defined finite area $|E| > 0$. Define

$$f(x\,,y) = \begin{cases} \dfrac{1}{|E|} & \text{if } (x\,,y) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Then, f is a joint density function, and the corresponding random vector (X, Y) is said to be distributed *uniformly* on E. Moreover, for all planar sets

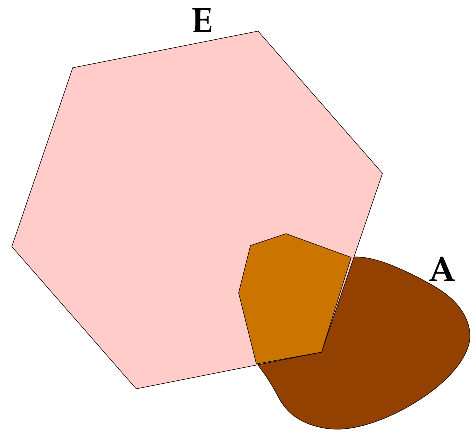


Figure 26.1. Region of integration in Example 26.2.

E with well-defined areas,

$$P\{(X,Y) \in A\} = \iint_{E\cap A} \frac{1}{|E|}\,dx\,dy = \frac{|E\cap A|}{|E|}.$$

See Figure 26.1. Thus, if the areas can be computed geometrically, then, in the case of a uniform distribution, there is no need to compute $\iint_A f(x\,,y)\,dx\,dy$.

Example 26.3. Let (X,Y) be uniformly distributed on $[-1,1]^2$. That is,

$$f_{X,Y}(x\,,y) = \begin{cases} \frac{1}{4} & \text{if } -1 \leqslant x \leqslant 1 \text{ and } -1 \leqslant y \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

We want to find $P\{|X+Y| \leqslant 1/2\}$. In this case, the areas are easy to compute geometrically; see Figure 26.2. The area of the square is $2^2 = 4$. The shaded area is the sum of the areas of two identical trapezoids and a parallelogram. It is thus equal to $2 \times \frac{1}{2} \times (1+\frac{1}{2})/2 + 1 \times 1 = 7/4$. Or, alternatively, the non-shaded area is that of two triangles. The shaded area is thus equal to $4 - 2 \times \frac{1}{2} \times \frac{3}{2} \times \frac{3}{2} = \frac{7}{4}$. Then, $P\{|X+Y| \leqslant 1/2\} = 7/16$. We could have used the definition of joint density functions and written

$$\begin{aligned} P\{|X+Y| \leqslant 1/2\} &= \iint_{|x+y|\leqslant 1/2} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-1}^{-1/2}\int_{-x-1/2}^{1} \frac{1}{4}\,dy\,dx + \int_{-1/2}^{1/2}\int_{-x-1/2}^{-x+1/2} \frac{1}{4}\,dy\,dx + \int_{1/2}^{1}\int_{-1}^{-x+1/2} \frac{1}{4}\,dy\,dx \\ &= \frac{7}{16}. \end{aligned}$$

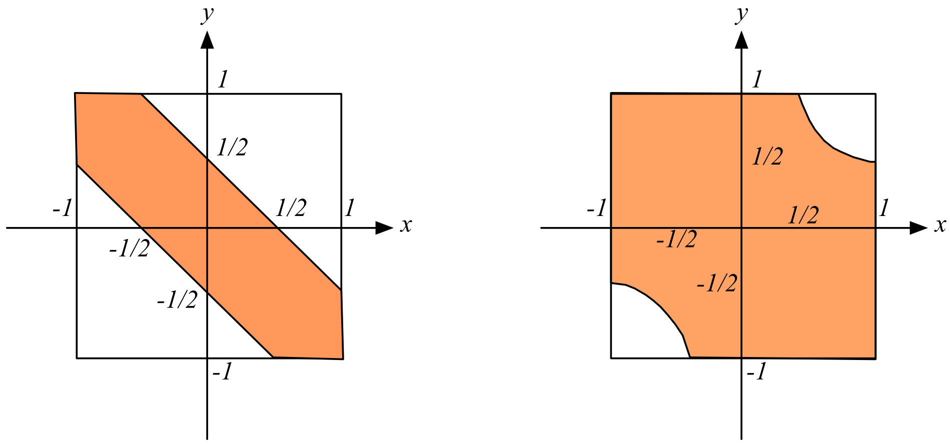


Figure 26.2. Regions of integration for Example 26.3. Left: $|x+y| \leqslant 1/2$. Right: $xy \leqslant 1/2$.

Next, we want to compute $P\{XY \leqslant 1/2\}$. This area is not easy to compute geometrically, in contrast to $|x+y| \leqslant 1/2$; see Figure 26.2. Thus, we need to compute it using the definition of joint density functions.

$$
\begin{aligned}
P\{XY \leqslant 1/2\} &= \iint\limits_{xy \leqslant 1/2} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-1}^{-1/2} \underbrace{\int_{1/2x}^{1} \frac{1}{4}\,dy}_{(1/4-1/8x)}\,dx + \int_{-1/2}^{1/2} \underbrace{\int_{-1}^{1} \frac{1}{4}\,dy}_{2/4}\,dx + \int_{1/2}^{1} \underbrace{\int_{-1}^{1/2x} \frac{1}{4}\,dy}_{(1/8x+1/4)}\,dx \\
&= \Big(\frac{x}{4} - \frac{\ln|x|}{8}\Big)\Big|_{-1}^{-1/2} + \frac{1}{2} + \Big(\frac{\ln|x|}{8} + \frac{x}{4}\Big)\Big|_{1/2}^{1} = \frac{3}{4} + \frac{\ln 2}{4}.
\end{aligned}
$$

Note that we could have computed the middle term geometrically: the area of the rectangle is $2 \times 1 = 2$ and thus the probability corresponding to it is $2/4 = 1/2$. An alternative way to compute the above probability is by computing one minus the integral over the non-shaded region in the right Figure 26.2. If, on top of that, one observes that both the pdf and the two non-shaded parts are symmetric relative to exchanging x and y, one can quickly compute

$$
P\{XY \leqslant 1/2\} = 1 - 2\int_{1/2}^{1} \Big(\int_{1/2x}^{1} \frac{1}{4}\,dy\Big)\,dx = 1 - 2\int_{1/2}^{1} \big(\frac{1}{4} - \frac{1}{8x}\big)\,dx = \frac{3}{4} + \frac{\ln 2}{4}.
$$

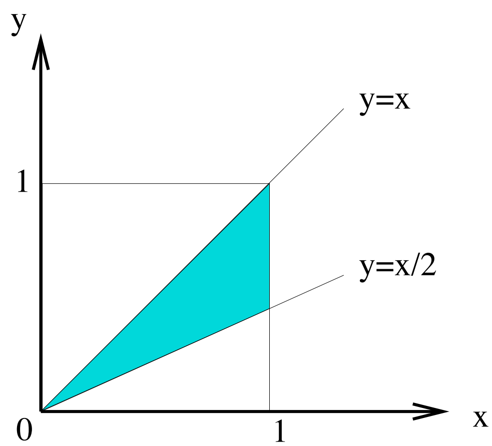


Figure 26.3. Region of integration in Example 26.4.

Example 26.4. Suppose (X, Y) has joint density

$$f(x, y) = \begin{cases} Cxy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let us first find C, and then $P\{X \leqslant 2Y\}$. To find C:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = \int_0^1 \int_0^x Cxy\, dy\, dx \\ &= C \int_0^1 x \underbrace{\left(\int_0^x y\, dy \right)}_{\frac{1}{2}x^2} dx = \frac{C}{2} \int_0^1 x^3\, dx = \frac{C}{8}. \end{aligned}$$

Therefore, $C = 8$, and hence

$$f(x, y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now

$$P\{X \leqslant 2Y\} = P\{(X, Y) \in A\} = \iint_A f(x, y)\, dx\, dy,$$

where A denotes the collection of all points (x, y) in the plane such that $x \leqslant 2y$. Therefore,

$$P\{X \leqslant 2Y\} = \int_0^1 \int_{x/2}^x 8xy\, dy\, dx = \frac{3}{4}.$$

See Figure 26.3. (Graphing a figure always helps!)

Homework Problems

Exercise 26.1. Let X and Y be two continuous random variables with joint density given by

$$f(x,y) = \begin{cases} \frac{1}{4} & \text{if } -1 \leqslant x \leqslant 1 \text{ and } -1 \leqslant y \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the following probabilities:

(a) $P\{X + Y \leqslant \frac{1}{2}\}$,
(b) $P\{X - Y \leqslant \frac{1}{2}\}$,
(c) $P\{XY \leqslant \frac{1}{4}\}$,
(d) $P\left\{\frac{Y}{X} \leqslant \frac{1}{2}\right\}$,
(e) $P\left\{\left|\frac{Y}{X}\right| \leqslant \frac{1}{2}\right\}$,
(f) $P\{|X| + |Y| \leqslant 1\}$,
(g) $P\{|Y| \leqslant e^X\}$.

Lecture 27

1. Marginals, distribution functions, etc.

If (X, Y) has joint density f, then

$$F_X(a) = P\{X \leqslant a\} = P\{(X, Y) \in A\},$$

where $A = \{(x, y) : x \leqslant a\}$. Thus,

$$F_X(a) = \int_{-\infty}^{a} \left(\int_{-\infty}^{\infty} f(x, y)\, dy \right) dx.$$

Differentiate, and apply the fundamental theorem of calculus, to find that

$$f_X(a) = \int_{-\infty}^{\infty} f(a, y)\, dy.$$

Similarly,

$$f_Y(b) = \int_{-\infty}^{\infty} f(x, b)\, dx.$$

Example 27.1 (Example 26.4, continued)**.** Let

$$f(x, y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$f_X(x) = \begin{cases} \int_0^x 8xy\, dy & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$
$$= \begin{cases} 4x^3 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

[Note the typo in Stirzaker's text, page 341.] Similarly,

$$f_Y(y) = \begin{cases} \int_y^1 8xy\,dx & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$
$$= \begin{cases} 4y(1-y^2) & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 27.2. Suppose (X, Y) is distributed uniformly in the circle of radius one about $(0,0)$. That is,

$$f(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } x^2+y^2 \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$f_X(x) = \begin{cases} \displaystyle\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,dy & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$
$$= \begin{cases} \dfrac{2}{\pi}\sqrt{1-x^2} & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

By symmetry, f_Y is the same function.

2. Independence

Just as in the discrete case, two continuous random variables are said to be independent if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, for all x and y. As a consequence, one has

$$P\{X \in A, Y \in B\} = \int_{A\times B} f_{X,Y}(x,y)\,dx\,dy = \int_{A\times B} f_X(x)f_Y(y)\,dx\,dy$$
$$= \int_A f_X(x)\,dx \int_B f_Y(y)\,dy = P\{X \in A\}P\{Y \in B\}.$$

This actually implies that if X and Y are independent, then $f(X)$ and $g(Y)$ are also independent, for any functions f and g. We omit the short proof.

Example 27.3. *Let $X \sim$ Exponential(λ_1) and $Y \sim$ Exponential(λ_2). What is $Z = \min(X, Y)$?*

Let us compute

$$F_Z(z) = P\{\min(X,Y) \leqslant z\} = 1 - P\{X > z, Y > z\}$$
$$= 1 - P\{X > z\}P\{Y > z\} = 1 - (1 - F_X(z))(1 - F_Y(z))$$
$$= 1 - e^{-\lambda_1 z}e^{-\lambda_2 z} = 1 - e^{-(\lambda_1+\lambda_2)z}.$$

Thus, $Z \sim \text{Exponential}(\lambda_1 + \lambda_2)$. This makes sense: say you have two stations, with the first serving about λ_1 people per unit time and the second serving about λ_2 people per unit time. Then, being served by these stations in a row is equivalent to being served by one station that serves about $\lambda_1 + \lambda_2$ people per unit time.

It is noteworthy that X and Y are independent as soon as one can write $f_{X,Y}(x,y)$ as the product of a function of x and a function of y. That is, if and only if $f_{X,Y}(x,y) = h(x)g(y)$, for some functions h and g. This is because we then have

$$f_X(x) = h(x)\Big(\int_{-\infty}^{\infty} g(y)\,dy\Big) \quad \text{and} \quad f_Y(y) = g(y)\Big(\int_{-\infty}^{\infty} h(x)\,dx\Big)$$

and

$$\Big(\int_{-\infty}^{\infty} h(x)\,dx\Big)\Big(\int_{-\infty}^{\infty} g(y)\,dy\Big) = 1$$

so that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. In other words, the functions h and g are really the same as the marginal density functions f_X and f_Y, up to the multiplicative constants that would make them integrate to one.

Example 27.4. Suppose (X,Y) is distributed uniformly on the square that joins the origin to the points $(1,0)$, $(1,1)$, and $(0,1)$. Then,

$$f_{X,Y}(x\,,y) = \begin{cases} 1 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, we see that $f_{X,Y}(x,y)$ does split into a product of a function of x and a function of y. Indeed, both $1 = 1 \times 1$ and $0 = 0 \times 0$. Furthermore, the set $0 < x < 1$ and $0 < y < 1$ is a set that involves two independent conditions on x and y. In fact, the marginals are equal to

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

and

$$f_Y(y) = \begin{cases} 1 & \text{if } 0 < y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and thus we see clearly that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Note that we have just shown that X and Y are both uniformly distributed on $(0,1)$.

Example 27.5. Let X and Y have joint density $f_{X,Y}(x,y) = \frac{1}{4}(1 + xy)$, for $-1 \leqslant x \leqslant 1$ and $-1 \leqslant y \leqslant 1$. Then, the marginals are

$$f_X(x) = \int_{-1}^{1} \frac{1}{4}\,dy + \frac{x}{4}\int_{-1}^{1} y\,dy = \frac{1}{2}\,,$$

for $-1 \leqslant x \leqslant 1$, and similarly $f_Y(y) = \frac{1}{2}$, for $-1 \leqslant y \leqslant 1$. However, clearly $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$. This shows that X and Y are not independent.

To confirm this we compute $P\{X \geqslant 0 \text{ and } Y \geqslant 0\}$ and $P\{X \geqslant 0\}P\{Y \geqslant 0\}$. First,

$$P\{X \geqslant 0, Y \geqslant 0\} = \int_0^1 \int_0^1 \frac{1}{4}(1+xy)\,dx\,dy = \frac{1}{4} + \frac{1}{4} \times \frac{1}{2} \times \frac{1}{2} = \frac{5}{16}.$$

On the other hand, $P\{X \geqslant 0\} = P\{Y \geqslant 0\} = 1/2$, since both X and Y are Uniform$(-1,1)$. (Alternatively, compute $\int_{-1}^1 \int_0^1 \frac{1}{4}(1+xy)\,dx\,dy = 1/2$.) Thus

$$P\{X \geqslant 0\}P\{Y \geqslant 0\} = \frac{1}{4} \neq \frac{5}{16} = P\{X \geqslant 0 \text{ and } Y \geqslant 0\}.$$

Example 27.6 (Example 27.1, continued)**.** Let

$$f(x,y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is tempting to say that X and Y are then independent, since $f(x,y)$ seems to be a product of two functions, one of x and one of y. However, one has to be careful with the set: $0 < y < x < 1$. This is where the dependence occurs. Indeed, if we know that $x = 1/2$, then we know that y cannot be larger than $1/2$. This is made clear once we compute the marginals, in Example 27.1, and observe that indeed $f_{X,Y}(x,y)$ is not equal to $f_X(x)f_Y(y)$.

The same caution needs to be applied to Example 27.2.

Example 27.7 (Order Statistic)**.** Let $X_1, \ldots, X_n$ be independent random variables with the same CDF $F(x)$. We want to compute the CDF of $S = \min(X_1, \ldots, X_n)$, the smallest of them. Then,

$$\begin{aligned} F_S(s) &= P\{S \leqslant s\} = 1 - P\{S > s\} = 1 - P\{X_1 > s, \ldots, X_n > s\} \\ &= 1 - P\{X_1 > s\}P\{X_2 > s\} \cdots P\{X_n > s\} \end{aligned}$$

by independence. Because the variables have the same CDF, $P\{X_1 > s\} = \cdots = P\{X_n > s\} = 1 - F(s)$. Thus,

$$F_S(s) = 1 - (1 - F(s))^n.$$

Similarly, if $T = \max(X_1, \ldots, X_n)$ is the largest of the variables, then

$$\begin{aligned} F_T(t) &= P\{T \leqslant t\} = P\{X_1 \leqslant t, \ldots, X_n \leqslant t\} \\ &= P\{X_1 \leqslant t\}P\{X_2 \leqslant t\} \cdots P\{X_n \leqslant t\} \\ &= P\{X_1 \leqslant t\}^n, \end{aligned}$$

and

$$F_T(t) = (F(t))^n.$$

Homework Problems

Exercise 27.1. Let X and Y be two continuous random variables with joint density given by

$$f(x,y) = \begin{cases} c(x+y) & \text{if } 0 \leqslant x \leqslant 1 \text{ and } 0 \leqslant y \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find c.
(b) Compute $P\{X < Y\}$.
(c) Find the marginal densities of X and Y.
(d) Compute $P\{X = Y\}$.

Exercise 27.2. Let X and Y be two continuous random variables with joint density given by

$$f(x,y) = \begin{cases} 4xy & \text{if } 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1 \text{ and } x \geqslant y \\ 6x^2 & \text{if } 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1 \text{ and } x < y \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the marginal densities of X and Y.
(b) Let $A = \{X \leqslant \frac{1}{2}\}$ and $B = \{Y \leqslant \frac{1}{2}\}$. Find $P(A \cup B)$.

Exercise 27.3. Let X and Y be two continuous random variables with joint density given by

$$f(x,y) = \begin{cases} 2e^{-(x+y)} & \text{if } 0 \leqslant y \leqslant x, \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal densities of X and Y.

Exercise 27.4. Let (X,Y) be uniformly distributed over the parallelogram with vertices $(-1,0)$, $(1,0)$, $(2,1)$, and $(0,1)$.

(a) Find and sketch the density functions of X and Y.
(b) A new random variable Z is defined by $Z = X + Y$. Show that Z is a continuous random variable and find and sketch its probability density function.

Exercise 27.5. Let (X,Y) be continuous random variables with joint density $f(x,y) = (x+y)/8$, $0 \leqslant x \leqslant 2$, $0 \leqslant y \leqslant 2$; $f(x,y) = 0$ elsewhere.

(a) Find the probability that $X^2 + Y \leqslant 1$.
(b) Find the conditional probability that exactly one of the random variables X and Y is $\leqslant 1$, given that at least one of the random variables is $\leqslant 1$.
(c) Determine whether or not X and Y are independent.

Lecture 28

1. Functions of a random vector

Basic problem: If (X, Y) has joint density f, then what, if any, is the joint density of (U, V), where $U = u(X, Y)$ and $V = v(X, Y)$? Or equivalently, $(U, V) = T(X, Y)$, where

$$T(x\,,y) = \begin{pmatrix} u(x\,,y) \\ v(x\,,y) \end{pmatrix}.$$

Example 28.1. Let (X, Y) be distributed uniformly in the disk of radius $\rho > 0$ about the origin in the plane. Thus,

$$f_{X,Y}(x\,,y) = \begin{cases} \dfrac{1}{\pi\rho^2} & \text{if } x^2 + y^2 \leqslant \rho^2, \\ 0 & \text{otherwise.} \end{cases}$$

We wish to write (X, Y), in polar coordinates, as (R, Θ), where $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arctan(Y/X)$. Then, we compute first the *joint distribution function* $F_{R,\Theta}$ of (R, Θ):

$$\begin{aligned} F_{R,\Theta}(r\,,\theta) &= P\{R \leqslant r\,, \Theta \leqslant \theta\} \\ &= P\{(X, Y) \in A\}, \end{aligned}$$

where A is the "partial cone" $\{(x\,,y) : x^2 + y^2 \leqslant r^2\,, \arctan(y/x) \leqslant \theta\}$. If $0 < r < \rho$ and $-\pi < \theta < \pi$, then

$$\begin{aligned} F_{R,\Theta}(r\,,\theta) &= \iint_A f_{X,Y}(x\,,y)\, dx\, dy \\ &= \int_0^r \int_0^\theta \frac{1}{\pi\rho^2} s\, ds\, d\varphi, \end{aligned}$$

after the change of variables $s = \sqrt{x^2+y^2}$ and $\varphi = \arctan(y/x)$. Therefore, for all $r \in (0\,,\rho)$ and $\theta \in (-\pi\,,\pi)$,

$$F_{R,\Theta}(r\,,\theta) = \frac{r^2\theta}{2\pi\rho^2}.$$

Since, by definition, $F_{R,\Theta}(r,\theta) = \int_{-\infty}^{r}\int_{-\infty}^{\theta} f_{R,\Theta}(s,\varphi)\,ds\,d\varphi$, we see that

$$f_{R,\Theta}(r\,,\theta) = \frac{\partial^2 F_{R,\Theta}}{\partial r \partial \theta}(r\,,\theta).$$

It is also clear that $f_{R,\Theta}(r,\theta) = 0$ if $r \notin (0,\rho)$ or $\theta \notin (-\pi,\pi)$. Therefore,

$$f_{R,\Theta}(r\,,\theta) = \begin{cases} \dfrac{r}{\pi\rho^2} & \text{if } 0 < r < \rho \text{ and } -\pi < \theta < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that the above yields $f_\Theta(\theta) = \frac{1}{2\pi}$, if $-\pi < \theta < \pi$, which implies that Θ is Uniform$(-\pi,\pi)$. On the other hand, $f_R(r) = \frac{2r}{\rho^2}$, if $0 < r < \rho$, which implies that R is not Uniform$(0,\rho)$. Indeed, it is more likely to pick a point with a larger radius (since there are more of them!).

The previous example can be generalized. Suppose T is invertible with inverse function

$$T^{-1}(u\,,v) = \begin{pmatrix} x(u\,,v) \\ y(u\,,v) \end{pmatrix}.$$

The *Jacobian* of this transformation is

$$\frac{\partial(x\,,y)}{\partial(u\,,v)} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Theorem 28.2. *If T is "nice," then*

$$f_{U,V}(u\,,v) = f_{X,Y}(x(u\,,v)\,,y(u\,,v))\left|\frac{\partial(x\,,y)}{\partial(u\,,v)}\right|.$$

Compare the above to Theorem 18.2. The Jacobian is really what comes up when doing change of variables in calculus, which is what the above theorem is all about. We thus omit the proof.

Example 28.3. In the polar coordinates example ($r = u$, $\theta = v$),

$$r(x\,,y) = \sqrt{x^2+y^2}, \quad \theta(x\,,y) = \arctan(y/x) = \theta,$$
$$x(r\,,\theta) = r\cos\theta, \quad y(r\,,\theta) = r\sin\theta.$$

Therefore, for all $r > 0$ and $\theta \in (-\pi\,,\pi)$, the Jacobian equals

$$(\cos(\theta) \times r\cos(\theta)) - (-r\sin(\theta) \times \sin(\theta)) = r.$$

Hence,

$$f_{R,\Theta}(r\,,\theta) = r\,f_{X,Y}(r\cos\theta\,,r\sin\theta).$$

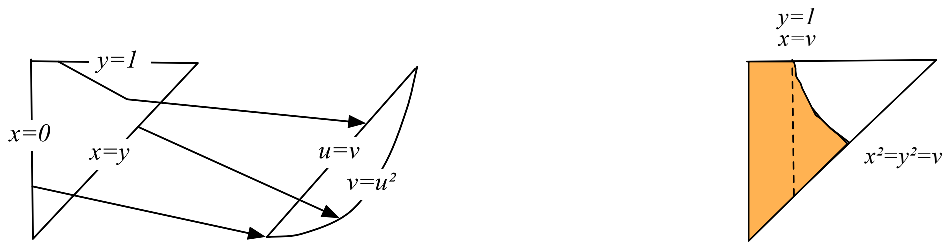


Figure 28.1. Domains for pdfs in Example 28.4. Left: domain transformation. Right: integration area for CDF calculation.

You should check that this yields Example 28.1, for instance.

Example 28.4. Let $f_{X,Y}(x,y) = 2(x+y)$, if $0 < x < y < 1$, and 0 otherwise. We want to find f_{XY}. In this case, we will first find $f_{X,XY}$, and then integrate the first coordinate out. This means we will use the transformation $(u,v) = (x, xy)$. Solving for x and y we get $(x,y) = (u, v/u)$, with $0 < v < u < \sqrt{v} < 1$; see Figure 28.1. The Jacobian is then equal to

$$1 \times \frac{1}{u} - 0 \times \frac{-v}{u^2} = \frac{1}{u}.$$

As a result, $f_{U,V}(u,v) = 2(u + v/u)/u$, with $0 < v < u < \sqrt{v} < 1$, and

$$f_V(v) = 2\int_v^{\sqrt{v}} (1 + v/u^2)\,du = 2(1-v), \quad \text{for } 0 < v < 1.$$

Alternatively, we could have computed the CDF of XY and then took its derivative to find the pdf. Clearly, $0 < XY < 1$ and thus $F_{XY}(v) = 0$ for $v \leqslant 0$ and $F_{XY}(v) = 1$ for $v \geqslant 1$. For $0 < v < 1$ we have (see Figure 28.1)

$$\begin{aligned}
F_{XY}(v) = P\{XY \leqslant v\} &= \iint_{xy \leqslant v} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_0^v \int_x^1 2(x+y)\,dy\,dx + \int_v^{\sqrt{v}} \int_x^{v/x} 2(x+y)\,dy\,dx \\
&= \int_0^v (2x - 2x^2 + 1 - x^2)\,dx + \int_v^{\sqrt{v}} (2v - 2x^2 + v^2/x^2 - x^2)\,dx \\
&= (v^2 - 2v^3/3 + v - v^3/3) \\
&\quad + (2v\sqrt{v} - 2v^2 - 2v^{3/2}/3 + 2v^3/3 - v^2/\sqrt{v} + v^2/v - v^{3/2}/3 + v^3/3) \\
&= 2v - v^2.
\end{aligned}$$

And $f_{XY}(v) = F'_{XY}(v) = 2 - 2v$, for $0 < v < 1$ (and 0 otherwise).

Example 28.5. Let us compute the joint density of $U = X$ and $V = X + Y$. Here,

$$u(x, y) = x, \quad v(x, y) = x + y$$
$$x(u, v) = u, \quad y(u, v) = v - u.$$

Therefore, the Jacobian equals

$$(1 \times 1) - (0 \times -1) = 1.$$

Consequently,

$$f_{U,V}(u, v) = f_{X,Y}(u, v - u).$$

This has an interesting by-product: The density function of $V = X + Y$ is

$$f_V(v) = \int_{-\infty}^{\infty} f_{U,V}(u, v)\, du = \int_{-\infty}^{\infty} f_{X,Y}(u, v - u)\, du.$$

Compare with Theorem **??**.

Example 28.6. Let X and Y be two independent standard normal random variables. We want to find the joint density of $U = X + Y$ and $V = X - Y$. Solving for x and y we get $x = (u + v)/2$ and $y = (u - v)/2$. The Jacobian is then equal to

$$\frac{1}{2} \times \frac{-1}{2} - \frac{1}{2} \times \frac{1}{2} = -\frac{1}{2}.$$

The joint pdf of X and Y is $f_{X,Y}(x, y) = f_X(x) f_Y(y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$. Thus,

$$f_{U,V}(u, v) = \frac{1}{2\pi} \exp\Big\{ -\big((u+v)^2 + (u-v)^2\big)/4\Big\} \times \frac{1}{2} = \frac{1}{2\pi} e^{-u^2/4} e^{-v^2/4}.$$

This in fact shows that U and V are independent, even though they are both mixtures of both X and Y. It also shows that they are both normal random variables with parameters (mean) 0 and (variance) 2; i.e. $N(0, 2)$.

Theorem 28.7. *If* (X, Y) *have joint mass function* $f(x, y)$ *and* $g(x, y)$ *is some function, then*

$$E[g(X, Y)] = \sum_{x,y} g(x, y)\, f(x, y),$$

provided the sum is well defined.

The same holds in the continuous case.

Theorem 28.8. *If* (X, Y) *have joint density function* $f(x, y)$ *and* $g(x, y)$ *is some function, then*

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f(x, y)\, dx,$$

provided the integral is well defined. In particular, if X *has density* $f(x)$ *and* $g(x)$ *is some function, then*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx,$$

provided the integral is well defined.

Homework Problems

Exercise 28.1. Let X and Y be independent and uniformly distributed between 0 and 1. Find and sketch the distribution and density functions of the random variable $Z = Y/X^2$.

Exercise 28.2. Let X and Y be two continuous independent random variables with densities given by

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

(a) Find the probability density function of $Z = X + Y$.

(b) Find the probability density function of $W = \frac{Y}{X}$.

Exercise 28.3. Let X and Y be two independent random variables both with distribution $N(0, 1)$. Find the probability density function of $Z = \frac{Y}{X}$.

Exercise 28.4. A point-size worm is inside an apple in the form of the sphere $x^2 + y^2 + z^2 = 4a^2$. (Its position is uniformly distributed.) If the apple is eaten down to a core determined by the intersection of the sphere and the cylinder $x^2 + y^2 = a^2$, find the probability that the worm will be eaten.

Exercise 28.5. A point (X, Y, Z) is uniformly distributed over the region described by $x^2 + y^2 \leqslant 4$, $0 \leqslant z \leqslant 3x$. Find the probability that $Z \leqslant 2X$.

Exercise 28.6. Let $T_1, \ldots, T_n$ be the order statistics of $X_1, \ldots, X_n$. That is, T_1 is the smallest of the X's, T_2 is the second smallest, and so on. T_n is the largest of the X's. Assume $X_1, \ldots, X_n$ are independent, each with density f. Show that the joint density of $T_1, \ldots, T_n$ is given by $g(t_1, \ldots, t_n) = n!\, f(t_1, \ldots, t_n)$ if $t_1 < t_2 < \cdots < t_n$ and 0 otherwise.

Hint: First find $P\{T_1 \leqslant t_1, \ldots, T_n \leqslant t_n, X_1 < X_2 < \ldots < X_n\}$.

Exercise 28.7. Let X, Y and Z be three continuous independent random variables with densities given by

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

Compute $P\{X \geqslant 2Y \geqslant 3Z\}$.

Exercise 28.8. A man and a woman agree to meet at a certain place some time between 10 am and 11 am. They agree that the first one to arrive will wait 10 minutes for the other to arrive and then leave. If the arrival times are independent and uniformly distributed, what is the probability that they will meet?

Exercise 28.9. When commuting to work, John can take public transportation (first a bus and then a train) or walk. Buses ride every 20 minutes and trains ride every 10 minutes. John arrives at the bus stop at 8 am precisely, but he doesn't know the exact schedule of buses, nor the exact schedule of trains. The total travel time on foot (resp. by public transportation) is 27 minutes (resp. 12 minutes).

(a) What is the probability that taking public transportation will take more time than walking?

(b) If buses are systematically 2 minutes late, how does it change the probability in (a)?

Exercise 28.10. Let X and Y be two independent random variable, both with distribution $N(0, \sigma^2)$ for some $\sigma > 0$. Let R and Θ be two random variables defined by

$$\begin{aligned} X &= R\cos(\Theta), \\ Y &= R\sin(\Theta), \end{aligned}$$

where $R \geqslant 0$. Prove that R and Θ are independent and find their density functions.

Exercise 28.11. A chamber consists of the inside of the cylinder $x^2 + y^2 = 1$. A particle at the origin is given initial velocity components $v_x = U$ and $v_y = V$, where (U, V) are independent random variables, each with standard normal density. There is no motion in the z-direction and no force acting on the particle after the initial push at time $t = 0$. If T is the time at which the particle strikes the wall of the chamber, find the distribution and density functions of T.

Lecture 29

1. Covariance

Theorem 29.1. *If* $E[X^2] < \infty$ *and* $E[Y^2] < \infty$ *then* $E[X]$, $E[Y]$, *and* $E[XY]$ *are all well-defined and finite.*

Proof. First observe that if $|X| \geqslant 1$ then $|X| \leqslant X^2$ and thus also $|X| \leqslant X^2 + 1$. If, on the other hand, $|X| \leqslant 1$ then also $|X| \leqslant X^2 + 1$. So in any case, $|X| \leqslant X^2 + 1$. This implies that $E[|X|] < \infty$ and by the triangle inequality $E[X]$ is well-defined and finite. The same reasoning goes for $E[Y]$. Lastly, observe that $(X + Y)^2 \geqslant 0$ and $(X - Y)^2 \geqslant 0$ imply

$$-X^2 - Y^2 \leqslant 2XY \leqslant X^2 + Y^2$$

and thus $|XY| \leqslant (X^2 + Y^2)/2$ and $E[XY]$ is well-defined and finite. $\square$

From the above theorem we see that if $E[X^2]$ and $E[Y^2]$ are finite then we can define the *covariance* between X and Y to be

$$\mathrm{Cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]. \tag{29.1}$$

Because $(X-E[X])(Y-E[Y]) = XY-XE[Y]-YE[X]+E[X]E[Y]$, we obtain the following, which is the computationally useful formula for covariance:

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]. \tag{29.2}$$

Here are some properties of the covariance.

Theorem 29.2. *Suppose* $E[X^2]$, $E[Y^2]$, *and* $E[Z^2]$ *are finite and let* a *be a nonrandom number.*

(1) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$;
(2) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$;

(3) $\mathrm{Cov}(X, a) = 0$ (*and thus also* $\mathrm{Cov}(a, Y) = 0$);

(4) $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$ (*and thus also* $\mathrm{Cov}(X, aY) = a\,\mathrm{Cov}(X, Y)$);

(5) $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$
(*and thus also* $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$);

(6) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.

The proofs go by directly applying the definition of covariance. Try them as an exercise! The above shows that covariance is bilinear. So if a, b, c, and d are nonrandom numbers and $E[X^2] < \infty$, $E[Y^2] < \infty$, $E[Z^2] < \infty$, and $E[U^2] < \infty$, then

$$\mathrm{Cov}(aX+bY, cZ+dU) = ac\,\mathrm{Cov}(X, Z)+ad\,\mathrm{Cov}(X, U)+bc\,\mathrm{Cov}(Y, Z)+bd\,\mathrm{Cov}(Y, U).$$

Example 29.3 (Example 25.1, continued)**.** Observe that the only nonzero value XY takes with positive probability is 1×1. (For example, 2×1 and 2×2 have 0 probability.) Thus,

$$E[XY] = 1 \times 1 \times \frac{2}{36} = \frac{2}{36}.$$

Also,

$$E[X] = E[Y] = 0 \times \frac{25}{36} + 1 \times \frac{10}{36} + 2 \times \frac{1}{36} = \frac{12}{36}.$$

Therefore,

$$\mathrm{Cov}(X, Y) = \frac{2}{36} - \frac{12}{36} \times \frac{12}{36} = -\frac{72}{1296} = -\frac{1}{18}.$$

2. Correlation

The *correlation* between X and Y is the quantity,

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\ \mathrm{Var}(Y)}}. \tag{29.3}$$

Example 29.4 (Example 25.1, continued)**.** Note that

$$E[X^2] = E[Y^2] = 0^2 \times \frac{25}{36} + 1^2 \times \frac{10}{36} + 2^2 \times \frac{1}{36} = \frac{14}{36}.$$

Therefore, the correlation between X and Y is

$$\rho(X, Y) = -\frac{1/18}{\sqrt{\left(\frac{5}{18}\right)\left(\frac{5}{18}\right)}} = -\frac{1}{5}.$$

We say that X and Y are negatively correlated. But what does this mean? The following few sections will help explain this.

Correlation is always a number between -1 and 1.

Theorem 29.5. *If* $E[X^2]$ *and* $E[Y^2]$ *are positive and finite, then* $-1 \leqslant \rho(X, Y) \leqslant 1$.


Figure 29.1. Left: Karl Hermann Amandus Schwarz (Jan 25, 1843 – Nov 30, 1921, Hermsdorf, Silesia [now Jerzmanowa, Poland]). Right: Victor Yakovlevich Bunyakovsky (Dec 16, 1804 – Dec 12, 1889, Bar, Ukraine, Russian Empire)

This is a straightforward variant of the following inequality. [How?]

Theorem 29.6 (Cauchy–Bunyakovsky–Schwarz inequality)**.** *If* $E[X^2]$ *and* $E[Y^2]$ *are finite, then*

$$|E[XY]| \leqslant \sqrt{E[X^2]\ E[Y^2]}.$$

Proof. Note that

$$X^2\left(E[Y^2]\right)^2 + Y^2\left(E[XY]\right)^2 - 2XY\,E[Y^2]E[XY] = \left(XE[Y^2] - YE[XY]\right)^2 \geqslant 0.$$

Therefore, taking expectation, we find

$$E[X^2]\left(E[Y^2]\right)^2 + E[Y^2]\left(E[XY]\right)^2 - 2E[Y^2]\left(E[XY]\right)^2 \geqslant 0$$

which leads to

$$E[Y^2]\Big(E[X^2]E[Y^2] - \left(E[XY]\right)^2\Big) \geqslant 0.$$

If $E[Y^2] > 0$, then we get

$$E[X^2]E[Y^2] \geqslant \left(E[XY]\right)^2,$$

which is the claim. Else, if $E[Y^2] = 0$, then $P\{Y = 0\} = 1$ and $P\{XY = 0\} = 1$. In this case the the result is true because it says $0 \leqslant 0$. □

Applying the above inequality to two special cases one deduces two very useful inequalities in mathematical analysis.

Example 29.7. Fix numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$. Let $P\{(X,Y) = (a_i, b_i)\} = 1/n$, for all $i = 1, \ldots, n$. Then, $P\{X = a_i\} = 1/n$ and $P\{Y = b_i\} = 1/n$. Applying the above theorem we get that

$$\Big(\sum_{i=1}^n a_i b_i\Big)^2 \leqslant \Big(\sum_{i=1}^n a_i^2\Big)\Big(\sum_{i=1}^n b_i^2\Big).$$

Example 29.8. Let $U \sim \mathrm{Uniform}(a,b)$ and let $X = g(U)$ and $Y = h(U)$ for some functions g and h. Applying the above theorem we deduce that

$$\Big(\int_a^b g(u)h(u)\,du\Big)^2 \leqslant \Big(\int_a^b |g(u)|^2\,du\Big)\Big(\int_a^b |h(u)|^2\,du\Big).$$

3. Correlation and independence

We say that X and Y are *uncorrelated* if $\rho(X,Y) = 0$; equivalently, if $\mathrm{Cov}(X,Y) = 0$. A significant property of uncorrelated random variables is that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$; see Theorem 29.2(2). We saw that this also happens when the variables are independent; see Theorem 24.7. This is not a coincidence.

Theorem 29.9. *If X and Y are independent and $E[X^2]$ and $E[Y^2]$ are finite, then X and Y are uncorrelated.*

Proof. It suffices to prove that $E[XY] = E[X]E[Y]$. But this is a consequence of Theorem 23.2(6). □

Example 29.10 (A counter example)**.** Sadly, it is only too common that people some times think that the converse to Theorem 29.9 is also true. So let us dispel this with a counterexample: Let Y and Z be two independent random variables such that $Z = \pm 1$ with probability $1/2$ each; and $Y = 1$ or 2 with probability $1/2$ each. Define $X = YZ$. Then, I claim that X and Y are uncorrelated but not independent.

First, note that $X = \pm 1$ and ± 2, with probability $1/4$ each. Therefore, $E[X] = 0$. Also, $XY = Y^2 Z = \pm 1$ and ± 4 with probability $1/4$ each. Therefore, again, $E[XY] = 0$. It follows that

$$\mathrm{Cov}(X,Y) = \underbrace{E[XY]}_{0} - \underbrace{E[X]}_{0}\,E[Y] = 0.$$

Thus, X and Y are uncorrelated. But they are not independent. Intuitively speaking, this is clear because $|X| = Y$. Here is one way to logically justify our claim:

$$P\{X = 1\,, Y = 2\} = 0 \neq \frac{1}{8} = P\{X = 1\}P\{Y = 2\}.$$

4. Correlation and linear dependence

Observe that if $Y = aX + b$ for some nonrandom constants $a \neq 0$ and b, then $\mathrm{Cov}(X, Y) = a\mathrm{Cov}(X, X) = a\mathrm{Var}(X)$. Furthermore, $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$. Therefore, $\rho(X, Y) = a/|a|$, which equals 1 if $a > 0$ and -1 if a is negative.

In other words, if Y follows X linearly and goes up when X does, then its correlation to X is $+1$. If it follows X linearly but goes down when X goes up, then its correlation is -1. The converse is in fact true.

Theorem 29.11. *Assume none of* X *and* Y *is constant (i.e. Var*$(X) > 0$ *and Var*$(Y) > 0$*). If* $\rho(X, Y) = 1$*, then there exist constants* b *and* $a > 0$ *such that* $P\{Y = aX + b\} = 1$*. Similarly, if* $\rho(X, Y) = -1$*, then there exist constants* b *and* $a < 0$ *such that* $P\{Y = aX + b\} = 1$.

Proof. Let $a = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. Note that a has the same sign as $\rho(X, Y)$. Recalling that $\rho(X, Y) = 1$ means $(\mathrm{Cov}(X, Y))^2 = \mathrm{Var}(X)\mathrm{Var}(Y)$, we have

$$\begin{aligned}
\mathrm{Var}(Y - aX) &= \mathrm{Var}(Y) + \mathrm{Var}(-aX) + 2\mathrm{Cov}(-aX, Y)\\
&= \mathrm{Var}(Y) + a^2\mathrm{Var}(X) - 2a\mathrm{Cov}(X, Y)\\
&= \mathrm{Var}(Y) - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{Var}(X)}\\
&= 0.
\end{aligned}$$

By Theorem 24.2 this implies the existence of a constant b such that

$$P\{Y - aX = b\} = 1. \qquad \square$$

Consider now the function

$$\begin{aligned}
f(a, b) &= E[(Y - aX - b)^2]\\
&= E[X^2]\, a^2 + b^2 + 2E[X]\, ab - 2E[XY]\, a - 2E[Y]\, b + E[Y^2].
\end{aligned}$$

This represents "how far" Y is from the line $aX+b$. Using some elementary calculus one can find that the optimal a and b that minimize f (and thus make Y as close as possible to $aX + b$) are the solutions to

$$E[X^2]\, a + E[X]\, b - E[XY] = 0 \quad \text{and} \quad b + E[X]\, a - E[Y] = 0.$$

$$Var(X)\, a + E[X]E[Y] - E[X]^2 a - E[XY] = 0 \quad \text{and} \quad b + E[X]\, a - E[Y] = 0.$$

Finding b in terms of a from the second equation and plugging back in the first one one gets

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \quad \text{and} \quad b = E[Y] - E[X]\, a.$$

Plugging back into f one has

$$\begin{aligned} f(a,b) &= E\Big[\Big((Y-aX)-E[Y-aX]\Big)^2\Big] = \operatorname{Var}(Y-aX) \\ &= \operatorname{Var}(Y) + a^2 \operatorname{Var}(X) - 2a \operatorname{Cov}(X,Y) \\ &= \operatorname{Var}(Y)(1-\rho(X,Y)^2). \end{aligned}$$

So the closer $|\rho(X,Y)|$ is to 1, the closer Y is to being a linear function of X.

Homework Problems

Exercise 29.1. If X and Y are affinely dependent (i.e. there exist numbers a and b such that $Y = aX + b$), show that $|\rho(X, Y)| = 1$.

Exercise 29.2. Show that equality occurs in the Cauchy-Bunyakovsky-Schwarz inequality (i.e. $E[XY]^2 = E[X^2]E[Y^2]$) if and only if X and Y are linearly dependent (i.e. there exists a number a such that $Y = aX$).

Exercise 29.3. Prove the following.

(a) For any real numbers $a_1, \dots, a_n$ and $b_1, \dots, b_n$,

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \leqslant \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2.$$

(b) If $\int_a^b g^2(x)\,dx$ and $\int_a^b h^2(x)\,dx$ are finite, then so is $\int_a^b g(x)h(x)\,dx$ and furthermore

$$\Big(\int_a^b g(x)h(x)\,dx\Big)^2 \leqslant \int_a^b g^2(x)\,dx \int_a^b h^2(x)\,dx.$$

Exercise 29.4. Let $X_1, \dots, X_n$ be a sequence of random variables with $E[X_i^2] < \infty$ for all $i = 1, \dots, n$. Prove that

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathrm{Cov}(X_i, X_j).$$

Lecture 30

1. Conditioning

Say X and Y are two random variables. If they are independent, then knowing something about Y does not say anything about X. So, for example, if $f_X(x)$ were the pdf of X, then knowing that $Y = 2$ the pdf of X is still $f_X(x)$. If, on the other hand, the two are dependent, then knowing $Y = 2$ must change the pdf of X. For example, consider the case $Y = |X|$ and $X \sim N(0,1)$. If we do not know anything about Y, then the pdf of X is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. However, if we know $Y = 2$, then X can only take the values 2 and -2 (with equal probability in this case). So knowing $Y = 2$ makes X a discrete random variable with mass function $f(2) = f(-2) = 1/2$.

1.1. Conditional mass functions. We are given two discrete random variables X and Y with mass functions f_X and f_Y, respectively. For all y, define the conditional mass function of X given that $Y = y$ as

$$f_{X|Y}(x\,|\,y) = P\left\{X = x \,\middle|\, Y = y\right\} = \frac{P\{X = x\,, Y = y\}}{P\{Y = y\}} = \frac{f_{X,Y}(x\,,y)}{f_Y(y)},$$

provided that $f_Y(y) > 0$ (i.e. y is a possible value for Y).

As a function in x, $f_{X|Y}(x\,|\,y)$ is a probability mass function. That is:

(1) $0 \leqslant f_{X|Y}(x\,|\,y) \leqslant 1$;

(2) $\sum_x f_{X|Y}(x\,|\,y) = 1$.

Example 30.1 (Example 25.1, continued)**.** In this example, the joint mass function of (X, Y), and the resulting marginal mass functions, were given by the following:

$x \setminus y$	**0**	**1**	**2**	f_X
0	16/36	8/36	1/36	25/36
1	8/36	2/36	0	10/36
2	1/36	0	0	1/36
f_Y	25/36	10/36	1/36	**1**

Let us calculate the conditional mass function of X, given that Y = 0:

$$f_{X|Y}(0\,|\,0) = \frac{f_{X,Y}(0\,,0)}{f_Y(0)} = \frac{16}{25}, \; f_{X|Y}(1\,|\,0) = \frac{f_{X,Y}(1\,,0)}{f_Y(0)} = \frac{8}{25},$$
$$f_{X|Y}(2\,|\,0) = \frac{f_{X,Y}(2\,,0)}{f_Y(0)} = \frac{1}{25}, \; f_{X|Y}(x\,|\,0) = 0 \text{ for other values of } x.$$

Similarly,

$$f_{X|Y}(0\,|\,1) = \frac{8}{10}, \; f_{X|Y}(1\,|\,1) = \frac{2}{10}, \; f_{X|Y}(x\,|\,1) = 0 \text{ for other values of } x.$$

and

$$f_{X|Y}(0\,|\,2) = 1, \; f_{X|Y}(x\,|\,2) = 0 \text{ for other values of } x.$$

These conditional mass functions are really just the relative frequencies in each column of the above table. Similarly, $f_{Y|X}(y\,|\,x)$ would be the relative frequencies in each row.

Observe that if we know $f_{X|Y}$ and f_Y, then $f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)f_Y(y)$. This is really Bayes' formula. The upshot is that one way to describe how two random variables interact is by giving their joint mass function, and another way is by giving the mass function of one and then the conditional mass function of the other (i.e. describing how the second random variable behaves, when the value of the first variable is known).

Example 30.2. Let $X \sim \text{Poisson}(\lambda)$ and if $X = x$ then let $Y \sim \text{Binomial}(x, p)$. By the above observation, the mass function for Y is

$$\begin{aligned} f_Y(y) &= \sum_x f_{X,Y}(x,y) = \sum_x f_X(x) f_{Y|X}(y\,|\,x) \\ &= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\binom{x}{y}p^y(1-p)^{x-y} \\ &= \frac{p^y\lambda^y}{y!}e^{-\lambda}\sum_{x=y}^{\infty}\frac{(\lambda(1-p))^{x-y}}{(x-y)!} = e^{-\lambda p}\frac{(\lambda p)^y}{y!}. \end{aligned}$$

In other words, $Y \sim \text{Poisson}(\lambda p)$. This in fact makes sense: after having finished shopping you stand in line to pay. The length of the line (X) is a Poisson random variable with average λ. But you decide now to use the fact that you own the store and you give each person ahead of you a coin to flip. The coin gives heads with probability p. If it comes up heads, the person stays in line. But if it comes up tails, the person leaves the store! Now, you still have a line of length Y in front of you. This is thus again a Poisson random variable. Its average, though, is λp (since you had on average λ people originally and then only a fraction of p of them stayed).

1.2. Conditional expectations. Define conditional expectations, as we did ordinary expectations. But use conditional probabilities in place of ordinary probabilities, viz.,

$$E[X \mid Y = y] = \sum_x x\, f_{X|Y}(x \mid y). \tag{30.1}$$

Example 30.3 (Example 30.1, continued)**.** Here,

$$E[X \mid Y = 1] = \left(0 \times \frac{8}{10}\right) + \left(1 \times \frac{2}{10}\right) = \frac{2}{10} = \frac{1}{5}.$$

Similarly,

$$E[X \mid Y = 0] = \left(0 \times \frac{16}{25}\right) + \left(1 \times \frac{8}{25}\right) + \left(2 \times \frac{1}{25}\right) = \frac{10}{25} = \frac{2}{5},$$

and

$$E[X \mid Y = 2] = 0.$$

Note that $E[X] = 10/36 + 2 \times 1/36 = 12/36 = 1/3$, which is none of the preceding. If you know, for example, that $Y = 0$, then your best bet for X is $2/5$. But if you have no extra knowledge, then your best bet for X is $1/3$.

However, let us note Bayes' formula in action:

$$\begin{aligned} E[X] &= E[X \mid Y = 0]P\{Y = 0\} + E[X \mid Y = 1]P\{Y = 1\} + E[X \mid Y = 2]P\{Y = 2\} \\ &= \left(\frac{2}{5} \times \frac{25}{36}\right) + \left(\frac{1}{5} \times \frac{10}{36}\right) + \left(0 \times \frac{1}{36}\right) = \frac{12}{36}, \end{aligned}$$

as it should be.

The proof of Bayes' formula is elementary:

$$\begin{aligned} E[X] &= \sum_x x\, P\{X = x\} = \sum_x x \sum_y P\{X = x, Y = y\} \\ &= \sum_{x,y} x\, P\{X = x \mid Y = y\}P\{Y = y\} = \sum_y \Big(\sum_x x\, f_{X|Y}(x \mid y)\Big)P\{Y = y\} \\ &= \sum_y E[X \mid Y = y]P\{Y = y\}, \end{aligned}$$

provided $E[X]$ exists, of course, so that we can interchange summations over x and y at will.

Example 30.4. Roll a fair die fairly n times. Let X be the number of 3's and Y the number of 6's. We want to compute the conditional mass function $f_{X|Y}(x\,|\,y)$. The possible values for Y are the integers from 0 to n. If we know $Y = y$, for $y = 0, \dots, n$, then we know that the possible values for X are the integers from 0 to $n - y$. If we know we got y 6's, then the probability of getting x 3's is

$$f_{X|Y}(x\,|\,y) = \binom{n-y}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{n-y-x},$$

for $y = 0, \dots, n$ and $x = 0, \dots, n-y$ (and it is 0 otherwise). In other words, given that $Y = y$, X is a Binomial$(n-y, 1/5)$. This makes sense, doesn't it? (You can also compute $f_{X|Y}$ using the definition.) Now, the expected value of X, given $Y = y$, is clear: $E[X\,|\,Y = y] = (n-y)/5$, for $y = 0, \dots, n$.

Example 30.5 (Example 20.3, continued)**.** Last time we computed the average amount one wins by considering a long table of all possible outcomes and their corresponding probabilities. Now, we can do things much cleaner. If we know the outcome of the die was x (an integer between 1 and 6), we lose $-x$ dollars right away. Then, we toss a fair coin x times and the expected amount we win at each toss is $2 \times \frac{1}{2} - 1 \times \frac{1}{2} = \frac{1}{2}$ dollars. So after x tosses the expected amount we win is $x/2$. Subtracting the amount we already lost we have that, given the die rolls an x, the expected amount we win is $x/2 - x = -x/2$. The probability of the die rolling x is $1/6$. Hence, Baye's formula gives that the expected amount we win is

$$E[W] = \sum_{x=1}^{6} E[W\,|\,X = x]P\{X = x\} = \sum_{x=1}^{6}\Big(-\frac{x}{2}\Big)\Big(\frac{1}{6}\Big) = -\frac{7}{4},$$

as we found in the longer computation. Here, we wrote W for the amount we win in this game and X for the outcome of the die.

Homework Problems

Exercise 30.1. If a single die is tossed independently n times, find the average number of 2's, given that the number of 1's is k.

Exercise 30.2. Of the 100 people in a certain village, 50 always tell the truth, 30 always lie, and 20 always refuse to answer. A single unbiased die is tossed. If the result is 1, 2, 3, or 4, a sample of size 30 is taken with replacement. If the result is 5 or 6, a samle of size 30 is taken without replacement. A random variable X is defined then as follows:

$X = 1$ if the resulting sample contains 10 people of each category.

$X = 2$ if the sample is taken with replacement and contains 12 liars.

$X = 3$ otherwise.

Find $E[X]$.

Lecture 31

1. Conditioning, continued

1.1. Conditional density functions. We are now given two continuous random variables X and Y with density functions f_X and f_Y, respectively. For all y, define the conditional density function of X given that $Y = y$ as

$$f_{X|Y}(x\,|\,y) = \frac{f(x\,,y)}{f_Y(y)}, \tag{31.1}$$

provided that $f_Y(y) > 0$.

As a function in x, $f_{X|Y}(x\,|\,y)$ is a probability density function. That is:

(1) $f_{X|Y}(x\,|\,y) \geqslant 0$;
(2) $\int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)\,dx = 1$.

Example 31.1. Let X and Y have joint density $f_{X,Y}(x,y) = e^{-y}$, $0 < x < y$. If we do not have any information about Y, the pdf of X is

$$f_X(x) = \int_x^{\infty} e^{-y}\,dy = e^{-x},\; x > 0$$

which means that $X \sim$ Exponential(1). But say we know that $Y = y > 0$. We would like to find $f_{X|Y}(x\,|\,y)$. To this end, we first compute

$$f_Y(y) = \int_0^y e^{-y}\,dx = ye^{-y},\; y > 0.$$

Then,

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{y},\; 0 < x < y.$$

This means that given $Y = y > 0$, $X \sim \text{Uniform}(0,y)$.

Example 31.2. Now, say X is a random variable with pdf $f_X(x) = xe^{-x}$, $x > 0$. Given $X = x > 0$, Y is a uniform random variable on $(0, x)$. This means that Y has the conditional pdf $f_{Y|X}(y\,|\,x) = \frac{1}{x}$, $0 < y < x$. Then, $f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y\,|\,x) = e^{-x}$, $0 < y < x$. This allows us to compute, for example, the pdf of Y:

$$f_Y(y) = \int_y^\infty e^{-x}\,dx = e^{-y},\ y > 0.$$

So $Y \sim$ Exponential(1). We can also compute things like $P\{X + Y \leqslant 2\}$. First, we need to figure out the boundary of integration. We know that $0 < y < x$ and now we also have $x + y \leqslant 2$. So x can go from 0 to 2 and then y can go from 0 to x or $2 - x$, whichever is smaller. The switch happens at $x = 2 - x$, and so at $x = 1$. Now we compute:

$$\begin{aligned} P\{X + Y \leqslant 2\} &= \int_0^1 \left(\int_0^x e^{-x}\,dy\right) dx + \int_1^2 \left(\int_0^{2-x} e^{-x}\,dy\right) dx \\ &= \int_0^1 xe^{-x}\,dx + \int_1^2 (2 - x)e^{-x}\,dx \\ &= -(x+1)e^{-x}\Big|_0^1 - 2e^{-x}\Big|_1^2 + (x+1)e^{-x}\Big|_1^2 = 1 + e^{-2} - 2e^{-1}. \end{aligned}$$

1.2. Conditional expectations. Define conditional expectations, as we did ordinary expectations. But use conditional probabilities in place of ordinary probabilities, viz.,

$$E[X\,|\,Y = y] = \int_{-\infty}^\infty x\, f_{X|Y}(x\,|\,y)\,dx.$$

Similarly, if g is a function of x, then

$$E[g(X)\,|\,Y = y] = \int_{-\infty}^\infty g(x)\, f_{X|Y}(x\,|\,y)\,dx.$$

Example 31.3 (Example 31.2, continued)**.** If we are given that $X = x > 0$, then $Y \sim \text{Uniform}(0, x)$. This implies that $E[Y\,|\,X = x] = x/2$. Now, say we are given that $Y = y > 0$. Then, to compute $E[X\,|\,Y = y]$ we need to find

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{e^{-x}}{e^{-y}} = e^{-(x-y)},\ 0 \leqslant y < x.$$

As a consequence, given $Y = y > 0$,

$$E[X\,|\,Y = y] = \int_y^\infty xe^{-(x-y)}\,dx = \int_0^\infty (z + y)e^{-z}\,dz = 1 + y.$$

We can also compute, for $y > 0$,

$$E[e^{X/2}\,|\,Y = y] = \int_y^\infty e^{x/2} e^{-(x-y)}\,dx = e^y \int_y^\infty e^{-x/2}\,dx = 2e^y e^{-y/2} = 2e^{y/2}.$$

Let us note Bayes' formula in action. On the one hand,

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^{\infty} x^2 e^{-x}\, dx = 2.$$

(To see the last equality, either use integration by parts, or the fact that this is the second moment of an Exponential(1), which is equal to its variance plus the square of its mean: $1 + 1^2 = 2$.) On the other hand,

$$E[X] = \int_{-\infty}^{\infty} E[X \mid Y = y]\, f_Y(y)\, dy = \int_0^{\infty} (1+y)e^{-y}\, dy = 2,$$

as it should be.

The proof of Bayes' formula is similar to the discrete case. (Do it!)

2. Conditioning on events

So far, we have learned how to compute the conditional pdf and expectation of X given $Y = y$. But what about the same quantities, conditional on knowing that $Y \leqslant 2$, instead of a specific value for Y? This is quite simple to answer in the discrete case. The mass function of X, given $Y \in B$, is:

$$f_{X\mid Y\in B}(x) = P\{X = x \mid Y \in B\} = \frac{P\{X = x, Y \in B\}}{P\{Y \in B\}} = \frac{\sum_{y\in B} f_{X,Y}(x,y)}{\sum_{y\in B} f_Y(y)}.$$

The analogous formula in the continuous case is for the pdf of X, given $Y \in B$:

$$f_{X\mid Y\in B}(x) = \frac{\int_B f_{X,Y}(x,y)\, dy}{\int_B f_Y(y)\, dy}. \tag{31.2}$$

Once we know the pdf (or mass function), formulas for expected values become clear:

$$E[X \mid Y \in B] = \frac{\sum_x \sum_{y\in B} x\, f_{X,Y}(x,y)}{\sum_{y\in B} f_Y(y)},$$

in the discrete case, and

$$E[X \mid Y \in B] = \frac{\int_{-\infty}^{\infty} x \left(\int_B f_{X,Y}(x,y)\, dy \right) dx}{\int_B f_Y(y)\, dy}, \tag{31.3}$$

in the continuous case. Observe that this can also be written as:

$$\begin{aligned} E[X \mid Y \in B] &= \frac{\int_B \left(\int_{-\infty}^{\infty} x\, f_{X\mid Y}(x\mid y)\, dx \right) f_Y(y)\, dy}{P\{Y \in B\}} \\ &= \frac{\int_B E[X \mid Y = y]\, f_Y(y)\, dy}{P\{Y \in B\}}. \end{aligned} \tag{31.4}$$

Example 31.4. Let (X, Y) have joint density function $f_{X,Y}(x, y) = e^{-x}$, $0 < y < x$. We want to find the expected value of Y, conditioned on $X \leqslant 5$. First, we find the conditional pdf. One part we need to compute is $P\{X \leqslant 5\}$. The pdf of X is

$$f_X(x) = \int_0^x e^{-x}\,dy = xe^{-x},\; x > 0,$$

and, using integration by parts, we have

$$P\{X \leqslant 5\} = \int_0^5 xe^{-x}\,dx = 1 - 6e^{-5}.$$

Now, we can go ahead with computing the conditional pdf using (31.3). If $X \leqslant 5$, then also $Y < 5$ (since $Y < X$) and

$$f_{Y|X\leqslant 5}(y) = \frac{\int_0^5 f_{X,Y}(x, y)\,dx}{1 - 6e^{-5}} = \frac{\int_y^5 e^{-x}\,dx}{1 - 6e^{-5}} = \frac{e^{-y} - e^{-5}}{1 - 6e^{-5}},\; 0 < y < 5.$$

(Check that this pdf integrates to 1!) Finally, using integration by parts, we can compute:

$$E[Y\,|\,X \leqslant 5] = \int_{-\infty}^{\infty} y\, f_{Y|X\leqslant 5}(y)\,dy = \frac{\int_0^5 y(e^{-y} - e^{-5})\,dy}{1 - 6e^{-5}} = \frac{1 - 18.5e^{-5}}{1 - 6e^{-5}} \approx 0.912.$$

Remark 31.5. We have $f_Y(y) = \int_y^\infty e^{-x}dx = e^{-y}$, $y > 0$. Thus, $Y \sim$ Exponential(1) and $E[Y] = 1$. Note now that the probability that $X \leqslant 5$ is $1 - 6e^{-5} \approx 0.96$, which is very close to 1. So knowing that $X \leqslant 5$ gives very little information. This explains why $E[Y\,|\,X \leqslant 5]$ is very close to $E[Y]$. Try to compute $E[Y\,|\,X \leqslant 1]$ and see how it is not that close to $E[Y]$ anymore. Try also to compute $E[Y\,|\,X \leqslant 10]$ and see how it is even closer to $E[Y]$ than $E[Y\,|\,X \leqslant 5]$.

We could have done things in a different order to find $E[Y\,|\,X \leqslant 5]$. First, we find the conditional expectation $E[Y\,|\,X = x]$. To do so, we need to find $f_{Y|X}$ and thus to find first $f_X(x) = \int_0^x e^{-x}\,dy = xe^{-x}$, $x > 0$. Hence, $f_{Y|X}(y\,|\,x) = 1/x$, for $0 < y < x$. Now, we see that $E[Y\,|\,X = x] = \int_0^x y\frac{1}{x}\,dy = \frac{x}{2}$. (This, of course, is not surprising since given $X = x$ we found that $Y \sim$ Uniform$(0, x)$.) Finally, we can apply (31.4) and use integration by parts to compute:

$$E[Y\,|\,X \leqslant 5] = \frac{\int_0^5 \frac{x}{2} xe^{-x}\,dx}{P\{X \leqslant 5\}} = \frac{\frac{1}{2}\int_0^5 x^2 e^{-x}\,dx}{\int_0^5 xe^{-x}\,dx} = \frac{1 - 18.5e^{-5}}{1 - 6e^{-5}}.$$

Example 31.6. Let X and Y be independent uniformly distributed on $(0, 1)$. Then, $P\{X + Y \leqslant 1\} = 1/2$. (This is the area of the triangle that is half the square $[0, 1]^2$.) Conditioned on knowing that $X + Y \leqslant 1$, the pair (X, Y) is still uniformly distributed but on the triangle $\{(x,y) \in [0,1]^2 : x+y \leqslant 1\}$. Consequently,

$$\mathrm{E}[X \,|\, X+Y \leqslant 1] = \frac{\int_0^1 x\left(\int_0^{1-x} dy\right) dx}{1/2} = 2\int_0^1 x(1-x)\,dx = \frac{1}{3}.$$

Alternatively, let $U = X+Y$. Using the transformation method we find that $f_{X,U}(x,u) = 1$, $0 < x < 1$ and $x < u < x+1$. (Do it!) This implies that $f_U(u) = \int_0^u dx = u$, for $0 < u < 1$, and $f_U(u) = \int_{u-1}^1 dx = 2-u$, for $1 < u < 2$. (This clearly integrates to 1. Use geometry to see that, rather than doing the (easy) computation!) We can readily see that $\mathrm{E}[U] = 1/2$. (Again, use geometry rather than the (easy) computation.) Furthermore, $f_{X|U}(x \,|\, u) = 1/u$, for $0 < x < u < 1$, and $f_{X|U}(x \,|\, u) = 1/(2-u)$, for $0 < u-1 < x < 1$. Thus,

$$\begin{aligned}\mathrm{E}[X \,|\, U=u] &= \begin{cases} \int_0^u x\frac{1}{u}\,dx = \frac{u}{2} & \text{if } 0 < u < 1,\\[2ex] \int_{u-1}^1 x\frac{1}{2-u}\,dx = \frac{1-(2-u)^2}{2(2-u)} & \text{if } 1 < u < 2.\end{cases}\\ &= \frac{u}{2}\,.\end{aligned}$$

Finally, using (31.4),

$$\mathrm{E}[X \,|\, X+Y \leqslant 1] = \mathrm{E}[X \,|\, U \leqslant 1] = \frac{\int_0^1 \mathrm{E}[X \,|\, U=u] f_U(u)\,du}{\mathrm{P}\{U \leqslant 1\}} = \frac{\int_0^1 \frac{u}{2} u\,du}{1/2} = \frac{1}{3}.$$

If instead we wanted to use (31.3), then we first write

$$f_{X|U\leqslant 1}(x) = \frac{\int_{-\infty}^{\infty} f_{X,U}(x,u)\,du}{\mathrm{P}\{U \leqslant 1\}} = \frac{\int_x^1 du}{1/2} = 2(1-x).$$

Then, applying (31.3),

$$\mathrm{E}[X \,|\, X+Y \leqslant 1] = \mathrm{E}[X \,|\, U \leqslant 1] = \int_{-\infty}^{\infty} x f_{X|U\leqslant 1}(x)\,dx = \int_0^1 2x(1-x)\,dx = \frac{1}{3}.$$

Homework Problems

Exercise 31.1. A number X is chosen with density $f_X(x) = 1/x^2$, $x \geqslant 1$; $f_X(x) = 0$, $x < 1$. If $X = x$, let Y be uniformly distributed between 0 and x. Find the distribution and density functions of Y.

Exercise 31.2. Let (X, Y) have density $f(x, y) = e^{-y}$, $0 \leqslant x \leqslant y$, $f(x, y) = 0$ elsewhere. Find the conditional density of Y given X, and $P\{Y \leqslant y \,|\, X = x\}$, the conditional distribution function of Y given $X = x$.

Exercise 31.3. Let (X, Y) have density $f(x, y) = k|x|$, $-1 \leqslant y \leqslant x \leqslant 1$; $f(x, y) = 0$ elsewhere. Find k; also find the individual densities of X and Y, the conditional density of Y given X, and the conditional density of X given Y.

Exercise 31.4. Let (X, Y) have density $f(x, y) = e^{-y}$, $0 \leqslant x \leqslant y$, $f(x, y) = 0$ elsewhere. Let $Z = Y - X$. Find the conditional density of Z given $X = x$. Also find $P\{1 \leqslant Z \leqslant 2 \,|\, X = x\}$.

Exercise 31.5. Let (X, Y) have density $f(x, y) = 8xy$, $0 \leqslant y \leqslant x \leqslant 1$; $f(x, y) = 0$ elsewhere.

(a) Find the conditional expectation of Y given $X = x$, and the conditional expectation of X given $Y = y$.

(b) Find the conditional expectation of Y^2 given $X = x$.

(c) Find the conditional expectation of Y given $A = \{X \leqslant 1/2\}$.

Exercise 31.6. Let (X, Y) be uniformly distributed over the parallelogram with vertices $(0, 0)$, $(2, 0)$, $(3, 1)$, $(1, 1)$. Find $E[Y|X = x]$.

Exercise 31.7. Let X and Y be independent random variables, each uniformly distributed between 0 and 2.

(a) Find the conditional probability that $X \geqslant 1$, given that $X + Y \leqslant 3$.

(b) Find the conditional expectation of X, given that $X + Y \leqslant 3$.

Exercise 31.8. The density for the time T required for the failure of a light bulb is $f(t) = \lambda e^{-\lambda t}$, $t \geqslant 0$. Find the conditional density function for $T - t_0$, given that $T > t_0$, and interpret the result intuitively.

Exercise 31.9. Let X and Y be independent random variables, each uniformly distributed between 0 and 1. Find the conditional expectation of $(X + Y)^2$ given $X - Y$. Hint: first find the joint density of $(X + Y, X - Y)$.

Exercise 31.10. Let X and Y be independent random variables, each with density $f(x) = (1/2)e^{-x}$, $x \geqslant 0$; $f(x) = 1/2$, $-1 \leqslant x \leqslant 0$; $f(x) = 0$, $x < -1$. Let $Z = X^2 + Y^2$. Find $E[Z|X = x]$.

Exercise 31.11. Let X be the number of successes in n Bernoulli trials, with probability p of success on a given trial. Find the conditional expectation of X, given that $X \geqslant 2$.

Exercise 31.12. Let X be uniformly distributed between 0 and 10, and define Y by

$$Y = \begin{cases} X^2 & \text{if } 0 \leqslant X \leqslant 6, \\ 3 & \text{if } 6 < X \leqslant 10. \end{cases}$$

Find the conditional expectation of Y given that $2 \leqslant Y \leqslant 4$.

Lecture 32

1. The binomial distribution, the golden theorem, and a normal approximation

Consider n independent coin tosses, each giving heads with probability p and tails with probability $1-p$. As was mentioned at the very beginning of the course, one expects that as n becomes very large the proportion of heads approaches p. While this cannot be used as the definition of the probability of heads being equal to p, it certainly is a consequence of the probability models we have been learning about. We will later see the following fact.

Theorem 32.1 (Bernoulli's golden theorem a.k.a. the law of large numbers, 1713)**.** *Suppose $0 \leqslant p \leqslant 1$ is fixed. Then, with probability 1, as $n \to \infty$,*

$$\frac{\textit{Number of heads}}{n} \approx p.$$

In other words: *in a large sample (n large), the probability is nearly one that the percentage in the sample is quite close to the percentage the population (p); i.e. with high probability, random sampling works well for large sample sizes.*

Next, a natural question comes to mind: for a given $a \leqslant b$ with $0 \leqslant a, b \leqslant n$, we know that

$$P\{\text{Number of heads is somewhere between } a \text{ and } b\} = \sum_{j=a}^{b} \binom{n}{j} p^j (1-p)^{n-j}.$$

Can we estimate this sum, if n is large? The answer is "yes." Another remarkable fact we will see later on is the following:

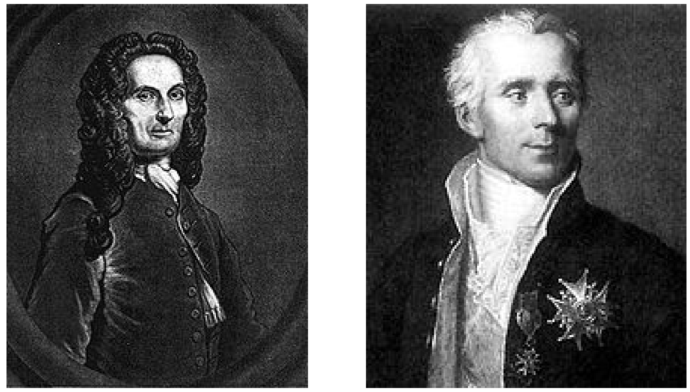


Figure 32.1. Left: Abraham de Moivre (May 26, 1667 – Nov 27, 1754, France). Right: Pierre-Simon, marquis de Laplace (Mar 23, 1749 – Mar 5, 1827, France).

Theorem 32.2 (The De Moivre–Laplace central limit theorem, 1733)**.** *Suppose $0 < p < 1$ is fixed. Then, as $n \to \infty$,*

$$\mathrm{P}\left\{\textit{Between } a \textit{ and } b \textit{ successes}\right\} \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right), \tag{32.1}$$

where Φ is the standard normal CDF (cumulative distribution function),

$$\Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \qquad \textit{for all } -\infty < z < \infty.$$

Remark 32.3. Taking $a \to -\infty$ it follows that

$$\mathrm{P}\{\text{Less than } b \text{ successes}\} \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right).$$

Similarly, taking $b \to \infty$ we have

$$\mathrm{P}\{\text{More than } a \text{ successes}\} \approx 1 - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right).$$

Next we learn to use this theorem; we will learn to understand its actual meaning later on.

Recall that $\Phi(z)$ is equal to the area to the left of z and under the standard normal curve and that, according to (16.1), the difference on the right-hand side in (32.1) is the area between a and b under the normal curve $N(np, np(1-p))$. The De Moivre–Laplace central limit theorem tells us then that *if n is large, then the binomial probability of having between a and b successes is approximately equal to the area between a and b under the normal curve with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.*

Example 32.4. The evening of a presidential election the ballots were opened and it was revealed that the race was a tie between the democratic and the republican candidates. In a random sample of 1963 voters what is the chance that more than 1021 voted for the republican candidate?

The exact answer to this question is computed from a binomial distribution with $n = 1963$ and $p = 0.5$. We are asked to compute

$$P\{\text{more than 1021 republican voters}\} = \sum_{j=1021}^{1963} \binom{1963}{j} \left(\frac{1}{2}\right)^j \left(1 - \frac{1}{2}\right)^{1963-j}.$$

Because $np = 981.5$ and $\sqrt{np(1-p)} = 22.15$, the normal approximation (Theorem 32.2) yields the following which turns out to be a quite good approximation:

$$\begin{aligned} P\{\text{more than 1021 republican voters}\} &\approx 1 - \Phi\left(\frac{1021 - 981.5}{22.15}\right) \\ &\approx 1 - \Phi(1.78) \approx 1 - 0.9625 = 0.0375. \end{aligned}$$

In other words, the chances are approximately 3.75% that the number of republican voters in the sample is more than 1021.

Example 32.5. A certain population is comprised of half men and half women. In a random sample of 10,000 what is the chance that the percentage of the men in the sample is somewhere between 49% and 51%?

The exact answer to this question is computed from a binomial distribution with $n = 10,000$ and $p = 0.5$. We are asked to compute

$$P\{\text{between 4900 and } 5,100 \text{ men}\} = \sum_{j=4900}^{5100} \binom{10000}{j} \left(\frac{1}{2}\right)^j \left(1 - \frac{1}{2}\right)^{10000-j}.$$

Because $np = 5000$ and $\sqrt{np(1-p)} = 50$, the normal approximation (Theorem 32.2) yields the following which turns out to be a quite good approximation:

$$\begin{aligned} P\{\text{between 4900 and 5100 men}\} &\approx \Phi\left(\frac{5100 - 5000}{50}\right) - \Phi\left(\frac{4900 - 5000}{50}\right) \\ &= \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) \\ &= 2\Phi(2) - 1 \\ &\approx (2 \times 0.9772) - 1 = 0.9544. \end{aligned}$$

In other words, the chances are approximately 95.44% that the percentage of men in the sample is somewhere between 49% and 51%. This is consistent with law of large numbers: in a large sample, the probability is nearly one that the percentage of the men in the sample is quite close to the percentage of men in the population.

Homework Problems

Exercise 32.1. Let X be the number of successes in a sequence of $10,000$ Bernoulli trials with probability of success 0.8. Estimate $P\{7940 \leqslant X \leqslant 8080\}$.

Lecture 33

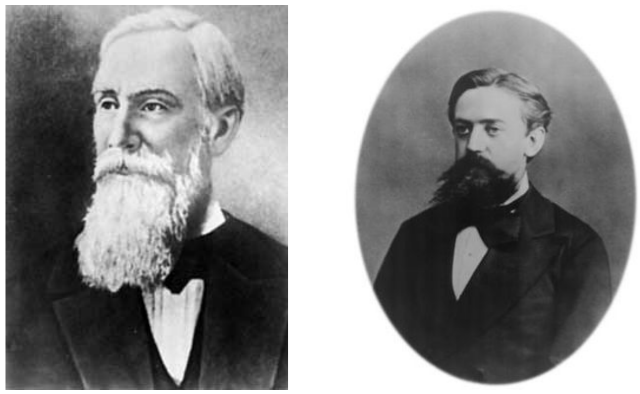


Figure 33.1. Left: Pafnuty Lvovich Chebyshev (May 16, 1821 – Dec 8, 1894, Kaluga, Russia). Right: Andrei Andreyevich Markov (Jun 14, 1856 – Jul 20, 1922, Ryazan, Russia)

1. Indicator functions

Let A be an event. The indicator function of A is the random variable defined by

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

It *indicates* whether x is in A or not!

Example 33.1. If A and B are two events, then $\mathbb{1}_{A\cap B} = \mathbb{1}_A \mathbb{1}_B$. This is because $\mathbb{1}_A(x)\mathbb{1}_B(x)$ equals 1 when both indicators are 1, and equals 0 otherwise. But both indicators equal 1 only when x is in both A and B, i.e. when $x \in A \cap B$.

The following is a useful "trick."

Lemma 33.2. *If A is an event, then* $P(A) = E[\mathbb{1}_A]$.

Proof. The proof is simple. $\mathbb{1}_A$ takes only two values: 0 and 1. Thus,

$$E[\mathbb{1}_A] = P(A^c) \times 0 + P(A) \times 1 = P(A). \qquad \square$$

Next, we prove a very useful inequality.

Lemma 33.3 (Chebyshev's inequality)**.** *If* h *is a nonnegative function, then for all* $\lambda > 0$,

$$P\{h(X) \geqslant \lambda\} \leqslant \frac{E[h(X)]}{\lambda}.$$

Proof. Let A be the event $\{x : h(x) \geqslant \lambda\}$. Then, because h is nonnegative,

$$h(x) \geqslant h(x)\mathbb{1}_A(x) \geqslant \lambda \mathbb{1}_A(x).$$

Thus,

$$E[h(X)] \geqslant \lambda E[\mathbb{1}_A] = \lambda P(A) = \lambda P\{h(X) \geqslant \lambda\}.$$

Divide by λ to finish. □

Thus, for example,

$$P\{|X| \geqslant \lambda\} \leqslant \frac{E[|X|]}{\lambda} \qquad \text{(Markov's inequality)}$$

$$P\{|X - E[X]| \geqslant \lambda\} \leqslant \frac{\mathrm{Var}(X)}{\lambda^2}, \tag{33.1}$$

$$P\{|X - E[X]| \geqslant \lambda\} \leqslant \frac{E[|X - E[X]|^4]}{\lambda^4}. \tag{33.2}$$

To get Markov's inequality, apply Lemma 33.3 with $h(x) = |x|$. To get the second inequality, first note that $|X-E[X]| \geqslant \lambda$ if and only if $|X-E[X]|^2 \geqslant \lambda^2$. Then, apply Lemma 33.3 with $h(x) = |x - E[X]|^2$ and with λ^2 in place of λ. The third inequality is similar: use $h(x) = |x - E[X]|^4$ and λ^4 in place of λ.

In words:

- If $E[|X|] < \infty$, then the probability that $|X|$ is large is small.
- If $\mathrm{Var}(X) = E[|X - E[X]|^2]$ is small, then with high probability $X \approx E[X]$.
- If $E[|X - E[X]|^4]$ is small, then with even higher probability $X \approx E[X]$.

We are now ready for the link between the intuitive understanding of probability (relative frequency) and the mathematical one (state space, probability of an event, random variable, expectation, etc).

2. The law of large numbers

Theorem 33.4 (Weak Law of Large Numbers)**.** *Suppose* $X_1, X_2, \dots, X_n$ *are independent, all with the same (well defined) mean* μ *and (finite) variance* $\sigma^2 < \infty$*. Then for all* $\varepsilon > 0$*, however small,*

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geqslant \varepsilon\right\} = 0. \tag{33.3}$$

To see why this theorem is a step towards the connection with the intuitive understanding of probability, think of the X_i's as being the results of independent coin tosses: $X_i = 1$ if the i-th toss results in heads and $X_i = 0$ otherwise. Then $(X_1 + \cdots + X_n)/n$ is precisely the relative frequency of heads: the fraction of time we got heads, up to the n-th toss. On the other hand, $\mu = E[X_1]$ equals the probability of getting heads (since X_1 is really a Bernoulli random variable). Thus, the theorem says that if we toss a coin a lot of times, the relative frequency of heads will, with very high chance, be close to the probability the coin lands heads. If the coin is fair, the relative frequency of heads will, with high probability, be close to 0.5.

The reason the theorem is called the *weak* law of large numbers is that it does not say that the relative frequency will *always* converge, as $n \to \infty$, to the probability the coin lands heads. It only says that the odds the relative frequency is far from the probability of getting heads (even by a tiny, but fixed, amount) get smaller as n grows. We will later prove the stronger version of this theorem, which then completes the link with the intuitive understanding of a probability. But let us, for now, prove the weak version.

Proof of Theorem 33.4. Let $\overline{X} = (X_1 + \cdots + X_n)/n$. (This is simply the sample mean.) Observe that

$$\begin{aligned}
E[\overline{X}] &= \frac{1}{n}E[X_1 + \cdots + X_n] \\
&= \frac{1}{n}\Big(E[X_1 + \cdots + X_{n-1}] + E[X_n]\Big) \\
&= \frac{1}{n}\Big(E[X_1 + \cdots + X_{n-2}] + E[X_{n-1}] + E[X_n]\Big) \\
&= \cdots = \frac{1}{n}\Big(E[X_1] + \cdots + E[X_n]\Big) \\
&= \frac{1}{n}(n\mu) = \mu.
\end{aligned}$$

Similarly, since X_i's are independent,

$$\begin{aligned}\mathrm{Var}(\overline{X}) &= \frac{1}{n^2}\mathrm{Var}(X_1+\cdots+X_n)\\ &= \frac{1}{n^2}\Big(\mathrm{Var}(X_1)+\cdots+\mathrm{Var}(X_n)\Big)\\ &= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.\end{aligned}$$

Applying Chebyshev's inequality, we find

$$P\left\{\left|\frac{X_1+\cdots+X_n}{n}-\mu\right| \geqslant \varepsilon\right\} \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

Let $n \nearrow \infty$ to finish. □

Now, we will state and prove the stronger version of the law of large numbers.

Theorem 33.5 (Strong Law of Large Numbers)**.** *Suppose* $X_1, X_2, \ldots, X_n$ *are independent, all with the same (well defined) mean* μ *and finite fourth moment* $\beta^4 = E[X_1^4] < \infty$. *Then,*

$$P\Big\{\lim_{n\to\infty}\frac{X_1+\cdots+X_n}{n} = \mu\Big\} = 1. \tag{33.4}$$

This theorem implies that if we flip a fair coin a lot of times and keep track of the relative frequency of heads, then it will converge, as the number of tosses grows, to 0.5, the probability of the coin landing heads.

There is a subtle difference between the statements of the two versions of the law of large numbers. This has to do with the different definitions of convergence for a sequence of random variables.

Definition 33.6. A sequence Y_n of random variables is said to converge *in probability* to a random variable Y if for any $\varepsilon > 0$ (however small) the quantity $P\{|Y_n - Y| > \varepsilon\}$ converges to 0 as $n \to \infty$.

Convergence in probability means that the probability that Y_n is far from Y by more than the fixed amount ε gets small as n gets large. In other words, Y_n is very likely to be close to Y for large n.

Definition 33.7. A sequence Y_n of random variables is said to converge *almost surely* to a random variable Y, if

$$P\Big\{Y_n \underset{n\to\infty}{\longrightarrow} Y\Big\} = 1.$$

Almost sure convergence means that the odds that Y_n does not converge to Y are nill. It is a fact that almost sure convergence implies convergence in probability. We omit the proof. However, the converse is not true, as the following example shows.

Example 33.8. Let Y_n be a sequence of independent random variables such that $P\{Y_n = 3\} = 1/n$ and $P\{Y_n = 2\} = 1 - 1/n$. Then, for any $\varepsilon \in (0,1)$, $P\{|Y_n - 2| > \varepsilon\} = P\{Y_n = 3\} = 1/n$ converges to 0 as $n \to \infty$. This proves that Y_n converges to the constant random variable $Y = 2$, in probability.

However, to say that Y_n converges to 2 almost surely would mean to say that Y_n becomes equal to 2, for large n (since Y_n takes only the values 2 and 3). If we fix two integers $M \geqslant N$, then the probability that $Y_n = 2$ for all n between N and M is equal, by independence, to $(1 - \frac{1}{N})(1 - \frac{1}{N+1}) \cdots (1 - \frac{1}{M})$. Observe now that if $x \in (0,1)$, then $(1-x) \leqslant e^{-x}$. Thus, the above probability is smaller than

$$\exp\Big\{ -\sum_{n=N}^{M} \frac{1}{n} \Big\},$$

which goes to 0 as $M \to \infty$, since the series with general term $1/n$ is divergent. This proves that there is 0 probability that $Y_n = 2$ for all $n \geqslant N$, no matter what N is. In other words, Y_n cannot converge to 2, almost surely.

In words: as n grows, the odds that Y_n is far from 2 decrease to 0. Thus, Y_n is converging to 2 in probability. However, with probability 1, Y_n will take the value 3 infinitely often and thus cannot be converging to 2 almost surely.

Back to the strong law of large numbers. It is noteworthy that the theorem actually holds without the assumption of finiteness of the fourth moment. In fact, one only needs that the mean of the X_i's is well defined and finite. The proof, however, becomes quite harder.

Proof of Theorem 33.5. For $n \geqslant 1$ define the event

$$A_n = \Big\{ \Big| \frac{X_1 + \cdots + X_n}{n} - \mu \Big| \geqslant \frac{1}{n^{1/8}} \Big\}.$$

We start similarly to the proof of the weak version. By Chebyshev's inequality (33.2), we have

$$\begin{aligned} P(A_n) &= P\Big\{ \Big| \frac{X_1 + \cdots + X_n}{n} - \mu \Big|^4 \geqslant \frac{1}{\sqrt{n}} \Big\} \\ &= P\{|(X_1 - \mu) + \cdots + (X_n - \mu)|^4 \geqslant n^{7/2}\} \\ &\leqslant \frac{E[|(X_1 - \mu) + \cdots + (X_n - \mu)|^4]}{n^{7/2}}. \end{aligned}$$

To compute $E[|(X_1 - \mu) + \cdots + (X_n - \mu)|^4]$ we expand the expression and notice that if $i \neq j$ then by independence we have $E[(X_i - \mu)(X_j - \mu)^3] =$

$E[X_i - \mu]E[(X_j - \mu)^3]$, and this equals 0 because $E[X_i] = \mu$. Also, there are n terms of the form

$$E[(X_i - \mu)^4] = E[X_i^4] - 3E[X_i^3]\mu + 3E[X_i]\mu^2 + \mu^4.$$

Observe that by the Cauchy-Schwarz inequality (Theorem 29.6),

$$E[X_i^2] = E[1 \times X_i^2] \leqslant \sqrt{1 \times E[X_i^4]} = \beta^2 < \infty$$

and

$$E[|X_i|^3] = E[|X_i| \times X_i^2] \leqslant \sqrt{E[X_i^2]E[X_i^4]} \leqslant \beta^3 < \infty.$$

Thus, $E[(X_i - \mu)^4] \leqslant \beta^4 + 3|\mu|\beta^3 + 3|\mu|^3 + \mu^4 = \gamma < \infty$. Similarly, we have $n(n-1)$ terms of the form $E[(X_i - \mu)^2(X_j - \mu)^2]$, with $i \neq j$. Here too the Cauchy-Schwarz inequality gives

$$E[(X_i - \mu)^2(X_j - \mu)^2] \leqslant \sqrt{E[(X_i - \mu)^4]E[(X_j - \mu)^4]} \leqslant \gamma < \infty.$$

In conclusion, $E[|(X_1 - \mu) + \cdots + (X_n - \mu)|^4] \leqslant (n + n(n-1))\gamma = n^2\gamma$. This gives us that

$$P(A_n) \leqslant \frac{\gamma}{n^{3/2}}.$$

But then

$$P\{\cup_{n \geqslant N} A_n\} \leqslant \sum_{n \geqslant N} P(A_n) \leqslant \gamma \sum_{n \geqslant N} \frac{1}{n^{3/2}}.$$

Next, observe that the sets $B_N = \cup_{n \geqslant N} A_n$ are decreasing; i.e. $B_{N+1} \subset B_N$. Thus, by Lemma 13.1,

$$P\{\cap_{N \geqslant 1} B_N\} = \lim_{N \to \infty} P(B_N) \leqslant \gamma \lim_{N \to \infty} \sum_{n \geqslant N} \frac{1}{n^{3/2}}.$$

Because the series with general term $1/n^{3/2}$ is summable, the right-most term above converges to 0 as $N \to \infty$. Thus,

$$P\Big\{\bigcap_{N \geqslant 1} \bigcup_{n \geqslant N} \{|(X_1 + \cdots + X_n)/n - \mu| \geqslant 1/n^{1/8}\}\Big\} = 0.$$

Taking complements we have

$$P\Big\{\bigcup_{N \geqslant 1} \bigcap_{n \geqslant N} \{|(X_1 + \cdots + X_n)/n - \mu| < 1/n^{1/8}\}\Big\} = 1.$$

The event in question reads:

"There is an $N \geqslant 1$ such that for all $n \geqslant N$, $\Big|\dfrac{X_1 + \cdots + X_n}{n} - \mu\Big| < \dfrac{1}{n^{1/8}}$."

This has probability one, as we have shown, and implies that $(X_1 + \cdots + X_n)/n$ converges to μ. Thus the latter has probability one as well. $\square$

Homework Problems

Exercise 33.1. Establish the following properties of indicator functions:

(a) $\mathbb{1}_\Omega = 1$, $\mathbb{1}_\varnothing = 0$

(b) $\mathbb{1}_{A\cap B} = \mathbb{1}_A \mathbb{1}_B$, $\mathbb{1}_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A\cup B}$

(c) $\mathbb{1}_{\cup_{i=1}^\infty A_i} = \sum_{i=1}^\infty \mathbb{1}_{A_i}$ if the A_i are disjoint

(d) If $A_1, A_2, \ldots$ is an increasing sequence of events ($A_n \subset A_{n+1}$ for all n) and $\cup_{n=1}^\infty A_n = A$, or if $A_1, A_2, \ldots$ is a decreasing sequence of events ($A_{n+1} \subset A_n$ for all n) and $\cap_{n=1}^\infty A_n = A$, then $\mathbb{1}_{A_n}(\omega) \to \mathbb{1}_A(\omega)$ for all ω.

Exercise 33.2. (a) Prove that

$$\mathbb{1}_{A_1\cup\cdots\cup A_n} = \sum_{i=1}^n (-1)^{i-1} \sum_{\substack{1\leqslant j_1,\ldots,j_i\leqslant n \\ j_1,\ldots,j_i \text{ all different}}} \mathbb{1}_{A_{j_1}\cap\cdots\cap A_{j_i}}.$$

(b) Deduce the inclusion-exclusion formula (3.5).

Exercise 33.3. 100 balls are tossed independently and at random into 50 boxes. Let X be the number of empty boxes. Find $E[X]$.

Exercise 33.4. Let X have the exponential density $f(x) = e^{-x}$, $x \geqslant 0$; $f(x) = 0$, $x < 0$. Let $\mu = E[X]$ and $\sigma^2 = \mathrm{Var}(X)$. Evaluate $P\{|X-\mu| \geqslant k\sigma\}$ and compare with Chebyshev's inequality.

Exercise 33.5. Suppose that X_n is the amount you win on trial n in a game of chance. Assume that the X_i are independent random variables, each with finite mean μ and finite variance σ^2. Make the realistic assumption that $\mu < 0$. Show that

$$P\left\{\frac{R_1 + \ldots + R_n}{n} < \mu/2\right\} \underset{n\to\infty}{\longrightarrow} 1.$$

Lecture 34

1. The moment generating function

The *moment generating function* (mgf) of a random variable X is the function of t given by

$$M(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} f(x), & \text{in the discrete setting,} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx, & \text{in the continuous setting.} \end{cases}$$

provided that the sum (or integral) exists. This is precisely the Laplace transform of the mass function (or pdf).

Note that $M(0)$ always equals 1 and $M(t)$ is always nonnegative.

A related transformation is the *characteristic function* of a random variable, given by

$$\Phi(t) = E[e^{itX}] = \begin{cases} \sum_x e^{itx} f(x), & \text{in the discrete setting,} \\ \int_{-\infty}^{\infty} e^{itx} f(x)\,dx, & \text{in the continuous setting.} \end{cases}$$

While the moment generating function may be infinite at some (or even all) nonzero values of t, the characteristic function is always defined and finite. It is precisely the Fourier transform of the mass function (or the pdf). In this course we will restrict attention to the moment generating function. However, one can equally work with the characteristic function instead, with the added advantage that it is always defined.

Example 34.1 (Bernoulli)**.** If $X \sim \text{Bernoulli}(p)$, then its mgf is

$$M(t) = 1 - p + pe^t.$$

Example 34.2 (Uniform)**.** If $X \sim \mathrm{Uniform}(a\,,b)$, then

$$M(t) = E[e^{tX}] = \frac{1}{b-a}\int_a^b e^{tx}\,dx = \frac{e^{bt}-e^{at}}{(b-a)t}.$$

Example 34.3 (Exponential)**.** If $X \sim \mathrm{Exponential}(\lambda)$, then

$$M(t) = E[e^{tX}] = \lambda\int_0^\infty e^{tx}e^{-\lambda x}\,dx.$$

This is infinite if $t \geqslant \lambda$ and otherwise equals

$$M(t) = \frac{\lambda}{\lambda - t}\,, \text{ if } t < \lambda.$$

This is indeed a useful transformation, viz.,

Theorem 34.4 (Uniqueness)**.** *If X and Y are two random variables—discrete* or *continuous—with moment generating functions M_X and M_Y, and if there exists $\delta > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\delta\,,\delta)$, then $M_X = M_Y$ and X and Y have the same distribution. More precisely:*

(1) *X is discrete if and only if Y is, in which case their mass functions are the same;*

(2) *X is continuous if and only if Y is, in which case their density functions are the same.*

We omit the proof. The theorem says that if we compute the mgf of some random variable and recognize it to be the mgf of a distribution we already knew, then that is precisely what the distribution of the random variable is. In other words, there is only one distribution that corresponds to any given mgf.

Example 34.5. If

$$M(t) = \frac{1}{2}e^{t} + \frac{1}{4}e^{-\pi t} + \frac{1}{4}e^{\sqrt{2}t},$$

then M is the mgf of a random variable with mass function

$$f(x) = \begin{cases} 1/2 & \text{if } x = 1,\\ 1/4 & \text{if } x = -\pi \text{ or } x = \sqrt{2},\\ 0 & \text{otherwise.}\end{cases}$$

2. Sums of independent random variables

Here is another reason why moment generating functions are a powerful tool.

Theorem 34.6. *If $X_1, \ldots, X_n$ are independent, with respective moment generating functions $M_{X_1}, \ldots, M_{X_n}$, then $\sum_{i=1}^n X_i$ has the mgf,*

$$M(t) = M_{X_1}(t) \times \cdots \times M_{X_n}(t).$$

Proof. By induction, it suffices to do this for $n = 2$ (why?). But then

$$M_{X_1+X_2}(t) = E\left[e^{t(X_1+X_2)}\right] = E\left[e^{tX_1} \times e^{tX_2}\right].$$

By independence, this is equal to the product of $E[e^{tX_1}]$ and $E[e^{tX_2}]$, which is the desired result. $\square$

Example 34.7 (Binomial)**.** Suppose $X \sim \text{Binomial}(n\,, p)$. Then we can write $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent Bernoulli(p). We can apply Theorem 34.6 then to find that

$$M_X(t) = (1 - p + pe^t)^n.$$

Example 34.8. If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m\,, p)$ are independent, then by the previous example and Theorem 34.6,

$$M_{X+Y}(t) = (1 - p + pe^t)^n (1 - p + pe^t)^m = (1 - p + pe^t)^{n+m}.$$

By the uniqueness theorem, $X + Y \sim \text{Binomial}(n + m\,, p)$. We found this out earlier by applying much harder methods. See Example **??**.

Example 34.9 (Poisson)**.** If $X \sim \text{Poisson}(\lambda)$, then

$$\begin{aligned} M(t) = E\left[e^{tX}\right] &= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!}. \end{aligned}$$

The sum gives the Taylor expansion of $e^{\lambda e^t}$. Therefore,

$$M(t) = e^{\lambda(e^t - 1)}.$$

Example 34.10. Now suppose $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\gamma)$ are independent. We apply the previous example and Theorem 34.6, in conjunction, to find that

$$M_{X+Y}(t) = e^{\lambda(e^t-1)} e^{\gamma(e^t-1)} = e^{(\lambda+\gamma)(e^t-1)}.$$

Thus, $X + Y \sim \text{Poisson}(\gamma + \lambda)$, thanks to the uniqueness theorem and Example 34.9. For a harder derivation of the same fact see Example **??**.

Example 34.11 (Geometric)**.** Let $X \sim \text{Geometric}(p)$. Then,

$$M(t) = \sum_{k=1}^{\infty} e^{kt} (1-p)^{k-1} p = \frac{p}{1-p} \sum_{k=1}^{\infty} ((1-p)e^t)^k.$$

The sum converges only when $(1-p)e^t < 1$, and thus when $t < -\ln(1-p)$. For example, the mgf of a Geometric(1/2) is only defined on the interval $(-\infty, \ln 2)$. So for a Geometric(p), the mgf is

$$M(t) = \frac{pe^t}{1-(1-p)e^t}\,, \text{ for } t < -\ln(1-p).$$

Example 34.12 (Negative Binomial)**.** Since a negative binomial with parameters r and p is the sum of r independent geometrics with parameter p, it has the mgf

$$M(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r, \text{ for } t < -\ln(1-p).$$

Example 34.13 (Gamma)**.** If $X \sim \mathrm{Gamma}(\alpha, \lambda)$, then

$$M(t) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\,dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x}\,dx.$$

If $t \geqslant \lambda$, then the integral is infinite. On the other hand, if $t < \lambda$, then

$$\begin{aligned} M(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{z^{\alpha-1}}{(\lambda-t)^{\alpha-1}} e^{-z} \frac{dz}{\lambda-t} \qquad (z = (\lambda-t)x) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha) \times (\lambda-t)^\alpha} \underbrace{\int_0^\infty z^{\alpha-1} e^{-z}\,dz}_{\Gamma(\alpha)} \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}. \end{aligned}$$

Thus,

$$M(t) = \left(\frac{\lambda}{\lambda-t}\right)^\alpha, \text{ if } t < \lambda.$$

In particular, if $\alpha = 1$ then we see (again) that the mgf of an Exponential(λ) is

$$M(t) = \frac{\lambda}{\lambda-t}\,, \text{ if } t < \lambda.$$

Example 34.14 (Normal)**.** If $X = N(\mu\,,\sigma^2)$, then

$$\begin{aligned}
M(t) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{tx}e^{-(x-\mu)^2/(2\sigma^2)}\,dx\\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2-2(\sigma^2t+\mu)x+(\sigma^2t+\mu)^2}{2\sigma^2}\right)\,dx\\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\sigma^2t-\mu)^2}{2\sigma^2}\right)\,dx\\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2}\,du \qquad (u=(x-\sigma^2t-\mu)/\sigma)\\
&= e^{\mu t+\sigma^2t^2/2}.
\end{aligned}$$

In particular, the mgf of a standard normal N(0,1) is

$$M(t) = e^{t^2/2}.$$

Homework Problems

Exercise 34.1. Let $X \sim \mathrm{Gamma}(\alpha, \lambda)$ and $Y \sim \mathrm{Gamma}(\beta, \lambda)$. Assume X and Y are independent. What is the distribution of $X + Y$?

Exercise 34.2. Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$. That is, each of them is normally distributed with its own mean an variance. Show that $X_1 + \cdots + X_n$ is again normally distributed, with mean $\mu_1 + \cdots + \mu_n$ and variance $\sigma_1^2 + \cdots + \sigma_n^2$.

Exercise 34.3. In each of the following, indicate whether or not the given function can be a moment generating function. If it can, then find the mass function or pdf of the corresponding random variable.

(a) $M(t) = 1 - t$.

(b) $M(t) = 2e^{-t}$.

(c) $M(t) = 1/(1 - t)$, for $t < 1$.

(d) $M(t) = \frac{1}{3} + \frac{1}{2}e^{2t} + \frac{1}{12}e^{-2t} + \frac{1}{12}e^{13t}$.

Exercise 34.4. Show that if $Y = aX + b$, with nonrandom constants a and b, then

$$M_Y(t) = e^{bt} M_X(at).$$

Exercise 34.5. Let X and Y take only the values 0, 1, or 2. Prove that if $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the same mass function. Do not quote the Uniqueness Theorem 34.4.

Lecture 35

1. Relation of MGF to moments

Suppose we know the function $M(t) = E[e^{tX}]$. Then, we can compute the moments of X from the function M by successive differentiation. For instance, suppose X is a continuous random variable with moment generating function M and density function f, and note that

$$M'(t) = \frac{d}{dt}\left(E[e^{tX}]\right) = \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

Now, loosely speaking, if the integral of the derivative converges absolutely, then a general fact states that we can take the derivative under the integral sign. That is,

$$M'(t) = \int_{-\infty}^{\infty} x e^{tx} f(x)\, dx = E\left[Xe^{tX}\right].$$

The same end-result holds if X is discrete with mass function f, but this time,

$$M'(t) = \sum_x x e^{tx} f(x) = E\left[Xe^{tX}\right].$$

Therefore, in any event:

$$M'(0) = E[X].$$

In general, this procedure yields,

$$M^{(n)}(t) = E\left[X^n e^{tX}\right].$$

Therefore,

$$M^{(n)}(0) = E\left[X^n\right].$$

Example 35.1 (Uniform)**.** We saw earlier that if X is distributed uniformly on $(0,1)$, then for all real numbers t,

$$M(t) = \frac{e^t - 1}{t}.$$

Therefore,

$$M'(t) = \frac{te^t - e^t + 1}{t^2}, \quad M''(t) = \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3},$$

whence

$$E[X] = M'(0) = \lim_{t \searrow 0} \frac{te^t - e^t + 1}{t^2} = \lim_{t \searrow 0} \frac{te^t}{2t} = \frac{1}{2},$$

by l'Hopital's rule. Similarly,

$$E\left[X^2\right] = \lim_{t \searrow 0} \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3} = \lim_{t \searrow 0} \frac{t^2e^t}{3t^2} = \frac{1}{3}.$$

Alternatively, these can be checked by direct computation, using the fact that $E[X^n] = \int_0^1 x^n \, dx = 1/(n+1)$.

Example 35.2 (Standard Normal)**.** If $X \sim N(0,1)$, then we have seen that $M(t) = e^{t^2/2}$. Thus, $M'(t) = te^{t^2/2}$ and $E[X] = M'(0) = 0$. Also, $M''(t) = (t^2+1)e^{t^2/2}$ and $E[X^2] = M''(0) = 1$.

Homework Problems

Exercise 35.1. Let $X_1, \dots, X_n$ be independent Poisson(λ) random variables. What is the distribution of $X_1 + \cdots + X_n$? What is the moment generating function of $(X_1 + \cdots + X_n - n\lambda)/\sqrt{n\lambda}$? Find the limit of this function as $n \to \infty$. Can you recognize the outcome as a moment generating function?

Exercise 35.2. Let X have pdf $f(x) = e^{-(x+2)}$ for $x > -2$, and $f(x) = 0$ otherwise. Find its mgf and use it to find E[X] and E[X^2].

Exercise 35.3. Let $X_n \sim \text{Geometric}(\lambda/n)$. Show that X_n/n converges in distribution to an Exponential(λ).

Hint: show that $M_{X_n/n}(t) = M_{X_n}(t/n)$. Then, when taking $n \to \infty$, write $h = 1/n$ and use de l'Hôpital's rule.

Exercise 35.4. Let $X_n \sim \text{Negative Binomial}(r, \lambda/n)$. Show that X_n/n converges in distribution to a Gamma(r, λ).

Remark 35.3. The above exercise shows that Gamma(α, λ) is a continuous version of a Negative Binomial with a fractional r!

Lecture 36

1. The Central Limit Theorem

In this section we will address one of the most important theorems in probability and statistics. In particular, we will answer the question "why is the normal distribution so important?"

To start, let us consider a sequence $X_1, \ldots, X_n$ of independent identically distributed (iid) random variables. Assume $\mu = E[X_1]$, the common average value of these random variables, exists and is finite. Then, $X_1 - \mu, \ldots, X_n - \mu$ represent the successive "measurement errors". The law of large numbers tells us that $(X_1 + \cdots + X_n - n\mu)/n$ converges to 0, with probability 1. So the cumulative measurement error $X_1 + \cdots + X_n - n\mu$ is not growing as fast as n. The question is then: how fast is it growing, if it is growing at all? To get an idea of the answer, let us compute the variance of this error. Let us assume $\sigma^2 = \mathrm{Var}(X_1)$, the common variance of the random variables X_i, to be finite. Then,

$$\mathrm{Var}(X_1+\cdots+X_n-n\mu) = \mathrm{Var}(X_1+\cdots+X_n) = \mathrm{Var}(X_1)+\cdots+\mathrm{Var}(X_n) = n\sigma^2.$$

So to have a quantity that has finite and positive variation we need to consider $(X_1+\cdots+X_n-n\mu)/\sqrt{n}$. In fact, we will consider $(X_1+\cdots+X_n-n\mu)/(\sigma\sqrt{n})$, just to normalize things to have a variance of 1.

What we are doing here is similar to the following simple idea: we know $a_n = 2n^2 + 5n - 10$ is growing to infinity as n grows, but how fast is it growing? Note that $a_n/n \to \infty$ and so a_n is growing faster than n is. On the other hand, $a_n/n^3 \to 0$ means a_n is growing slower than n^2. To find how fast a_n is growing we need to find just the right power α for which a_n/n^α will not grow to infinity nor will it go to 0. The correct answer in this particular example is $\alpha = 2$. In the case of $X_1 + \cdots + X_n - n\mu$, the correct answer seems to be $\alpha = 1/2$.

Now that we have an idea of how fast the error grows (like $\sqrt{n}$), we start wondering what the distribution of our cumulative error would look like. In other words, if we draw a histogram of $(X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$, then will it have some recognizable shape as n grows large? Note that another way to write this cumulative error is as

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

where $\overline{X}$ is the *sample mean* $(X_1 + \cdots + X_n)/n$. So the question we are really asking is: what does the distribution of the sample mean $\overline{X}$ look like, for large samples (n large)?

Let us first find the answer to our question in the case when the random variables are normally distributed.

Example 36.1. Let $X_1, \ldots, X_n$ be a sequence of independent random variables that are all $N(\mu, \sigma^2)$. Then, $E[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Let us compute the moment generating function of $Z_n = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$:

$$\begin{aligned}
M_{Z_n}(t) &= E[e^{tZ_n}] = E\Big[\exp\Big\{\frac{t}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu)\Big\}\Big] \\
&= e^{-\sqrt{n}\mu t/\sigma}\big(M_{X_1}(t/(\sigma\sqrt{n}))\big)^n \\
&= e^{-\sqrt{n}\mu t/\sigma}\big(e^{\mu t/(\sigma\sqrt{n}))+\sigma^2 t^2/(2\sigma^2 n)}\big)^n \\
&= e^{t^2/2}.
\end{aligned}$$

Thus, for any $n \geqslant 1$, Z_n is a standard normal.

Motivated by the above use of the moment generating function, the following theorem will be helpful in our quest.

Theorem 36.2 (Lévy's continuity theorem)**.** *Let X_n be a random variables—discrete* or *continuous—with moment generating functions M_n. Also, let X be a random variable with moment generating function M. Suppose there exists $\delta > 0$ such that:*

(1) *If $-\delta < t < \delta$, then $M_n(t), M(t) < \infty$ for all $n \geqslant 1$; and*

(2) $\lim_{n\to\infty} M_n(t) = M(t)$ *for all $t \in (-\delta, \delta)$, then*

$$\lim_{n\to\infty} F_{X_n}(a) = \lim_{n\to\infty} P\{X_n \leqslant a\} = P\{X \leqslant a\} = F_X(a),$$

for all numbers a at which F_X is continuous.

The convergence of CDFs in (2) above roughly means that if we fix a very large n and graph the cumulative histogram of the X_n, then it will look like that of X. It will in fact look closer and closer to the CDF of X, as we take n larger and larger. This is called convergence in distribution.

Example 36.3. Let $X_n \sim \text{Uniform}(0, 1/n)$. Since $0 \leqslant X_n \leqslant 1/n$, it is clear that $P\{X_n \to 0\} = 1$; i.e. $X_n \to 0$ almost-surely. It also converges to 0 in distribution: $M_{X_n}(t) = \frac{e^{t/n}-1}{t/n} \to 1$ as $n \to \infty$. To see this, write $h = t/n$ and observe that we are looking for the limit of $(e^h - 1)/h$ as $h \to 0$. Now, either use the definition of derivative to see that the answer is the derivative of e^x at $x = 0$, or use de l'Hôpital's rule. Now, since $M(t) = 1$ is the moment generating function of the random variable $X = 0$, Lévy's continuity theorem implies that the CDF of X_n must converge to that of X at points of continuity of the latter. But $F_X(x) = 0$ if $x < 0$ and 1 if $x \geqslant 0$. (Recall that CDFs are right-continuous.) Note that $F_{X_n}(x) = 0$ if $x < 0$ and thus converges in that case to $F_X(x)$. Similarly, $F_{X_n}(x) = 1$ if $x > 1/n$ and thus as $n \to \infty$, $F_{X_n}(x) \to 1$ for $x > 0$. However, $F_{X_n}(0) = 0$ which does not converge to $F_X(0) = 1$. This is not a problem, though, because $x = 0$ is a point of discontinuity for F_X.

Example 36.4 (Law of rare events)**.** Suppose $X_n \sim \text{Binomial}(n\,, \lambda/n)$, where $\lambda > 0$ is fixed, and $n \geqslant \lambda$. Then,

$$M_{X_n}(t) = \left(1 - p + pe^{-t}\right)^n = \left(1 - \frac{\lambda}{n} + \frac{\lambda e^{-t}}{n}\right)^n \to \exp\left(-\lambda + \lambda e^{-t}\right).$$

Note that the right-most term is $M_X(t)$, where $X = \text{Poisson}(\lambda)$. Therefore, by Lévy's continuity theorem,

$$\lim_{n\to\infty} P\{X_n \leqslant a\} = P\{X \leqslant a\}, \tag{36.1}$$

at all a where F_X is continuous. But X is discrete and integer-valued. Therefore, F_X is continuous at a if and only if a is not a nonnegative integer. If a *is* a nonnegative integer, then we can choose a non-integer $b \in (a\,, a + 1)$ to find that

$$\lim_{n\to\infty} P\{X_n \leqslant b\} = P\{X \leqslant b\}.$$

Because X_n and X are both non-negative integers, $X_n \leqslant b$ if and only if $X_n \leqslant a$, and $X \leqslant b$ if and only if $X \leqslant a$. Therefore, this time (36.1) holds for all a, i.e. even at points where F_X is discontinuous.

Let us now find the answer to our question about the distribution of the sample mean in a few cases.

Example 36.5. Let $X_1, \ldots, X_n$ be a sequence of independent random variables that are all Poisson(λ). Then, $E[X_1] = \lambda$, $\text{Var}(X_1) = \lambda$, and $M_{X_1}(t) =$

$e^{\lambda(e^t-1)}$. Let us compute the mgf of $Z_n = (X_1 + \cdots + X_n - n\lambda)/\sqrt{n\lambda}$:

$$\begin{aligned}M_{Z_n}(t) &= E[e^{tZ_n}] = E\Big[\exp\Big\{\frac{t}{\sqrt{n\lambda}}(X_1 + \cdots + X_n - n\lambda)\Big\}\Big]\\ &= e^{-t\sqrt{\lambda n}}M_{X_1}(t/\sqrt{\lambda n})^n = \exp\Big\{-t\sqrt{\lambda n} + n\lambda(e^{t/\sqrt{\lambda n}} - 1)\Big\}.\end{aligned}$$

According to the Taylor–MacLaurin expansion,

$$e^{t/\sqrt{\lambda n}} = 1 + \frac{t}{\sqrt{\lambda n}} + \frac{t^2}{2\lambda n} + \text{smaller terms}.$$

Thus,

$$M_{Z_n}(t) = \exp\Big\{-t\sqrt{\lambda n} + t\sqrt{\lambda n} + \frac{t^2}{2} + \text{smaller terms}\Big\} \underset{n\to\infty}{\longrightarrow} e^{t^2/2}.$$

Since $e^{t^2/2}$ is the mgf of a standard normal, Lévy's continuity theorem and the fact that the CDF of a standard normal is continuous imply that

$$P\{Z_n \leqslant a\} \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}\,dx, \text{ for all } a.$$

Example 36.6 (The de Moivre–Laplace central limit theorem)**.** Suppose $S_n \sim \text{Binomial}(n\,,p)$, where $p \in (0\,,1)$ is fixed, and define Z_n to be its *standardization*. That is, $Z_n = (S_n - E[S_n])/\sqrt{\text{Var}(S_n)}$. Alternatively,

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

Recall that S_n is really the sum of n independent Bernoulli(p) random variables and that the mean of a Bernoulli(p) is p and its variance is $p(1-p)$. Thus, the question of what the asymptotic distribution of Z_n looks like is precisely what we have been asking about in this section.

We know that for all real numbers t, $M_{S_n}(t) = \left(1 - p + pe^t\right)^n$. We can use this to compute M_{Z_n} as follows:

$$\begin{aligned}M_{Z_n}(t) &= E\left[\exp\left(t \cdot \frac{S_n - np}{\sqrt{np(1-p)}}\right)\right]\\ &= e^{-npt/\sqrt{np(1-p)}}M_{S_n}\Big(\frac{t}{\sqrt{np(1-p)}}\Big)\\ &= e^{-t\sqrt{np/(1-p)}}\Big(1 - p + pe^{t/\sqrt{np(1-p)}}\Big)^n\\ &= \Big((1-p)e^{-t\sqrt{\frac{p}{n(1-p)}}} + pe^{t\sqrt{\frac{1-p}{np}}}\Big)^n.\end{aligned}$$

According to the Taylor–MacLaurin expansion,

$$\exp\left\{t\sqrt{\frac{1-p}{np}}\right\} = 1 + t\sqrt{\frac{1-p}{np}} + \frac{t^2(1-p)}{2np} + \text{smaller terms},$$
$$\exp\left\{-t\sqrt{\frac{p}{n(1-p)}}\right\} = 1 - t\sqrt{\frac{p}{n(1-p)}} + \frac{t^2 p}{2n(1-p)} + \text{smaller terms}.$$

Therefore,

$$\begin{aligned}
&p\exp\left\{t\sqrt{\frac{1-p}{np}}\right\} + (1-p)\exp\left\{-t\sqrt{\frac{p}{n(1-p)}}\right\}\\
&= p\left(1 + t\sqrt{\frac{1-p}{np}} + \frac{t^2(1-p)}{2np} + \cdots\right) + (1-p)\left(1 - t\sqrt{\frac{p}{n(1-p)}} + \frac{t^2 p}{2n(1-p)} + \cdots\right)\\
&= p + t\sqrt{\frac{p(1-p)}{n}} + \frac{t^2(1-p)}{2n} + \cdots + (1-p) - t\sqrt{\frac{p(1-p)}{n}} + \frac{t^2 p}{2n} + \cdots\\
&= 1 + \frac{t^2}{2n} + \text{smaller terms}.
\end{aligned}$$

Consequently,

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \text{smaller terms}\right)^n \to e^{t^2/2}.$$

We recognize the right-hand side as $M_Z(t)$, where $Z \sim N(0,1)$. Because F_Z is continuous, this prove the *central limit theorem* of de Moivre: For all real numbers a,

$$\lim_{n\to\infty} P\{Z_n \leqslant a\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}\,dx.$$

The Central Limit Theroem (i.e. the limit theorem that is central in probability theory), states that the above three results are not a coincidence.

Theorem 36.7 (Central Limit Theorem)**.** *Let $X_1,\ldots,X_n,\ldots$ be independent random variables identically distributed. Assume $\sigma^2 = Var(X_1)$ is finite. Then, if we let $\mu = E[X_1]$ (which exists because the variance is finite),*

$$Z = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

converges in distribution, as $n \to \infty$, to a standard normal random variable.

If $E[|X_1|^3] < \infty$, then the proof of the above theorem goes exactly the same way as in the two examples above, i.e. through a Taylor–MacLaurin expansion. We leave the details to the student.

A nice visualization of the Central Limit Theorem in action is done using a Galton board. Look it up on Google and on YouTube.

One way to use this theorem is to approximately compute percentiles of the sample mean $\overline{X}$.

Example 36.8. The waiting time at a certain toll station is exponentially distributed with an average waiting time of 30 seconds. If we use minutes to compute things, then this average waiting time is $\mu = 0.5$ a minute and thus $\lambda = 1/\mu = 2$. Consequently, the variance is $\sigma^2 = 1/\lambda^2 = 1/4$. If 100 cars are in line, we know the average waiting time is 50 minutes. This is only an estimate, however. So, for example, we want to estimate of the probabilities they wait between 45 minutes and an hour. If X_i is the waiting time of car number i, then we want to compute $P\{45 < X_1 + \cdots + X_{100} < 60\}$. We can use the central limit theorem for this. The average waiting time for the 100 cars is 50 minutes. The theorem tells us that the distribution of

$$Z = \frac{X_1 + \cdots + X_{100} - 50}{0.5\sqrt{100}}$$

is approximately standard normal. Thus,

$$P\{45 < X_1 + \cdots + X_{100} < 60\} = P\{-5/5 < Z < 10/5\} \approx \frac{1}{\sqrt{2\pi}} \int_{-1}^{2} e^{-z^2/2}\,dz,$$

which we can find using the tables for the so-called *error function*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\,dz$$

(which is simply the CDF of a standard normal). The tables give that $\varphi(2) \approx 0.9772$. Most tables do not give $\varphi(x)$ for negative numbers x. This is because symmetry implies that $\varphi(-x) = 1 - \varphi(x)$. Thus, $\varphi(-1) = 1 - \varphi(1) \approx 1 - 0.8413 = 0.1587$. Hence, the probability we are looking for is approximately equal to $0.9772 - 0.1587 = 0.8185$, i.e. about 82%.

Example 36.9. In the 2004 presidential elections the National Election Pool ran an exit poll. At 7:32 PM it was reported that 1963 voters from Ohio responded to the poll, of which 941 said they voted for President Bush and 1022 for Senator Kerry. It is safe to assume the sampling procedure was done correctly without any biases (e.g. nonresponse, etc). We are wondering if this data has significant evidence that President Bush had lost the race in Ohio.

To answer this question, we assume the race resulted in a tie and compute the odds that only 941 of the 1963 voters would vote for President

Bush. The de Moivre–Laplace central limit theorem tells us that

$$Z = \frac{S_{1963} - 0.5 \times 1963}{\sqrt{1963 \times 0.5(1-0.5)}}$$

is approximately standard normal. Hence,

$$P\{S_{1963} \leqslant 941\} = P\left\{Z \leqslant \frac{941 - 981.5}{\sqrt{490.75}}\right\} \approx \varphi(-1.8282) = 1 - \varphi(1.8282) \approx 0.03376.$$

In other words, had the result been a tie, there is chance of at most 3.4% no more than 941 of the 1963 voters would have voted for President Bush.

Homework Problems

Exercise 36.1. A carton contains 144 baseballs, each of which has a mean weight of 5 ounces and a standard deviation of 2/5 ounces. (Standard deviation is the square root of the variance.) Find an approximate value for the probability that the total weight of the baseballs in the carton is no more than 725 ounces.

Exercise 36.2. Let $X_i \sim \text{Uniform}(0, 1)$, where $X_1, \ldots, X_{20}$ are independent. Find normal approximations for each of the following:

(a) $P\left\{\sum_{i=1}^{20} X_i \leqslant 12\right\}$.

(b) The 90-th percentile of $\sum_{i=1}^{20} X_i$; i.e. the number a for which

$$P\left\{\sum_{i=1}^{20} X_i \leqslant a\right\} = 0.9.$$

Exercise 36.3. Let X_i be the weight of the i-th passenger's luggage. Assume that the weights are independent, each with pdf

$$f(x) = 3x^2/80^3, \text{ for } 0 < x < 80,$$

and 0 otherwise. Approximate $P\left\{\sum_{i=1}^{100} X_i > 6025\right\}$.

Exercise 36.4. Let X be the number of baskets scored in a sequence of $10,000$ free throws attempts by some NBA player. This player's rate of scoring is 80%. Estimate the probability that he scores between 7940 and 8080 baskets.

Appendix A

1. What next?

This course has covered some basics in probability theory. There are several (not necessarily exclusive) directions to follow from here.

One direction is learning some statistics. For example, if you would like to compute the average height of students at the university, one way would be to run a census asking each student for their height. Thanks to the law of the large numbers, a more efficient way would be to collect a sample and compute the average height in that sample. Natural questions arise: how many students should be in the sample? How to take the sample? Is the average height in the sample a good estimate of the average height of all university students? If it is, then how large an error are we making? These are very important practical issues. Example 36.8 touched on this matter. The same kind of questions arise, for example, when designing exit polls. The main question is really about estimating parameters of the distribution of the data; e.g. the mean, the variance, etc. This is the main topic of Statistical Inference I (Math 5080).

Another situation where statistics is helpful is, for example, when someone claims the average student at the university is more than 6 feet tall. How would you collect data and check this claim? Clearly, the first step is to use a sample to estimate the average height. But that would be just an estimate and includes an error that is due to the randomness of the sample. So if you find in your sample an average height of 6.2 feet, is this large enough to conclude the average height of all university students is indeed larger then 6 feet? What about if you find an average of 6.01 feet? Or 7.5 feet? Can one estimate the error due to random sampling and thus guarantee that an average of 7.5 feet is not larger than 6 feet only due to randomness but because the average height of all students is really more than 6 feet? Inline with this, example 36.9 shows how one can use probability theory to check the validity of certain claims. These issues are addressed in Statistical Inference II (Math 5090).

Another direction is learning more probability theory. Here is a selected subset of topics you would learn in Math 6040. The notion of algebra of events is established more seriously, a very important topic that we touched upon very lightly and then brushed under the rug for the rest of this course. The two main theorems are proved properly: the strong law of large numbers (with just one finite moment, instead of four as we did in this course) and the central limit theorem (with just two moments instead of three). The revolutionary object "Brownian motion" is introduced and explored. Markov chains may also be covered, depending on the instructor and time. And more... We will talk a little bit about Brownian motion in the next two sections. Brownian motion (and simulations) is also explored in Stochastic Processes and Simulation I & II (Math 5040 and 5050).

1.1. History of Brownian motion, as quoted from the Wiki. The Roman Lucretius's scientific poem On the Nature of Things (c. 60 BC) has a remarkable description of Brownian motion of dust particles. He uses this as a proof of the existence of atoms:

"Observe what happens when sunbeams are admitted into a building and shed light on its shadowy places. You will see a multitude of tiny particles mingling in a multitude of ways... their dancing is an actual indication of underlying movements of matter that are hidden from our sight... It originates with the atoms which move of themselves [i.e. spontaneously]. Then those small compound bodies that are least removed from the impetus of the atoms are set in motion by the impact of their invisible blows and in turn cannon against slightly larger bodies. So the movement mounts up from the atoms and gradually emerges to the level of our senses, so that those bodies are in motion that we see in sunbeams, moved by blows that remain invisible."

Although the mingling motion of dust particles is caused largely by air currents, the glittering, tumbling motion of small dust particles is, indeed, caused chiefly by true Brownian dynamics.

Jan Ingenhousz had described the irregular motion of coal dust particles on the surface of alcohol in 1785. Nevertheless Brownian motion is traditionally regarded as discovered by the botanist Robert Brown in 1827. It is believed that Brown was studying pollen particles floating in water under the microscope. He then observed minute particles within the vacuoles of the pollen grains executing a jittery motion. By repeating the experiment with particles of dust, he was able to rule out that the motion was due to pollen particles being 'alive', although the origin of the motion was yet to be explained.

The first person to describe the mathematics behind Brownian motion was Thorvald N. Thiele in 1880 in a paper on the method of least squares. This was followed independently by Louis Bachelier in 1900 in his PhD thesis "The theory of speculation", in which he presented a stochastic analysis of the stock and option markets. However, it was Albert Einstein's (in his 1905 paper) and Marian Smoluchowski's (1906) independent research of the problem that brought the solution to the attention of physicists, and presented it as a way to indirectly confirm the existence of atoms and molecules.

However, at first the predictions of Einstein's formula were refuted by a series of experiments, by Svedberg in 1906 and 1907, which gave displacements of the particles as 4 to 6 times the predicted value, and by Henri in 1908 who found displacements 3 times greater than Einstein's formula predicted. But Einstein's predictions were finally confirmed in a series of experiments carried out by Chaidesaigues in 1908 and Perrin in 1909. The confirmation of Einstein's theory constituted empirical progress for the kinetic theory of heat. In essence, Einstein showed that the motion can be predicted directly from the kinetic model of thermal equilibrium. The importance of the theory lay in the fact that it confirmed the kinetic theory's account of the second law of thermodynamics as being an essentially statistical law.

1.2. More history. Einstein predicted that the one-dimensional Brownian motion is a random function of time, written as $W(t)$ for "time" $t \geqslant 0$, such that:

(a) At time 0, the random movement starts at the origin; i.e. $W(0) = 0$.

(b) At any given time $t > 0$, the position $W(t)$ of the particle has the normal distribution with mean 0 and variance t.

(c) If $t > s > 0$, then the displacement from time s to time t is independent of the past until time s; i.e., $W(t) - W(s)$ is independent of all the values $W(r)$; $r \leqslant s$.

(d) The displacement is time-homogeneous; i.e., the distribution of $W(t) - W(s)$ is the same as the distribution of $W(t - s)$ which is in turn normal with mean 0 and variance $t - s$.

(e) The random function W is continuous.

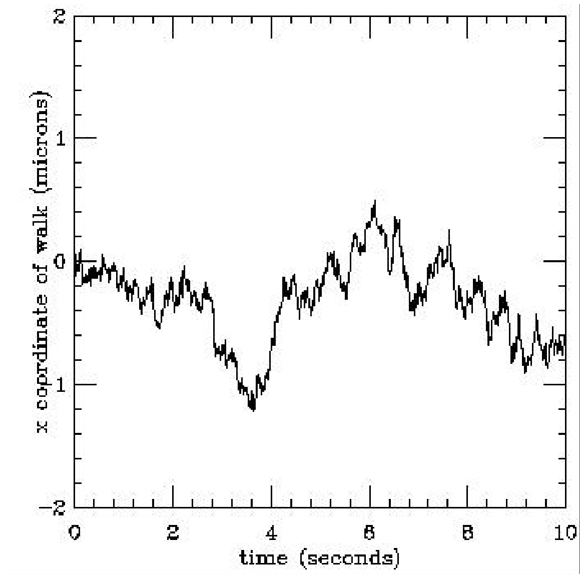
In 1923, Norbert Wiener (a professor at MIT and a child prodigy) proved the existence of Brownian motion and set down a rm mathematical foundation for its further development and analysis. Wiener used the recently-developed mathematics of É. Borel and H. Steinhaus (the subject is called measure theory), and cleverly combined it with a nice idea from a different mathematical discpline (harmonic analysis).

Finally, the classical development of Brownian motion was complete in a 1939 work of Paul Lévy who proved the following remarkable fact: If you replace the normal distribution by any other distribution in Einsteins predicates, then either there is no stochastic process that satises the properties (a)–(d), or (e) fails to hold! Lévys work was closely related to the concurrent and independent work of A. I. Khintchine in Russia, and is nowadays called the Lévy-Khintchine Formula.

The work of Paul Lévy started the modern age of random processes, and at its center, the theory of Brownian motion. The modern literature on this is truly vast. But all probabilists would (or should) agree that a centerpiece of the classical literature is the 1942/1946 work of K. Itô who derived a calculus – and thereby a theory of stochastic differential equations – that is completely different from the ordinary nonstochastic theory. This theory is nowadays at the very heart of the applications of probability theory to mathematical nance, mathematical biology, turbulence, oceanography, etc.

For us, the final important step in the analysis of Brownian motion was the 1951 work of Donsker who was a Professor of mathematics at The New York University. Amongst other things, Donsker verified a 1949 conjecture of the great American mathematician J. L. Doob by showing that once you run them for a long time, all mean-zero variance-one random walks look like Brownian motion! We will say more on this in the next section.

1.3. Random Walk and Brownian Motion. We describe the one-dimensional case, but adding more dimensions is not too hard. In one dimension, imagine the lattice of integer numbers. Say a particle starts at position 0. If the particle is at position x, then it flips a fair coin and moves to $x + 1$ if the coin lands Heads and to $x - 1$ if it lands Tails. The motion of the particle is a random process called *simple symmetric r andom walk*. The increments of the random walker are simply a sequence of independent random variables taking the value 1 with probability 1/2 and -1 with probability 1/2. If we call the k-th increment X_k, then the position of the walker at time n is $S_n = X_1 + \cdots + X_n$. This is a simplistic model of a dust particle moving in discrete time and discrete space. As we have done a few times in this course, we can try and derive a continuum model. The idea is to plot the positions S_n against the times n and connect the dots. If we do this for n very large and look from afar, so that the fine details are washed out,



but not too far, so that there is still a process going on and we do not just see a straight line!, then a continuous curve emerges. This is the so-called *Brownian motion*. This is not a trivial fact to work out mathematically and is expressed in the following theorem. A picture explains it nicely, though; see Figure 1.3.

DONSKER'S THEOREM. *Let* $X_1, X_2, \dots$ *denote independent, identically distributed random variables with mean zero and variance one. The random walk is then the random sequence* $S_n = X_1 + \cdots + X_n$ *, and for all* n *large, the random graph* $(0,0), (1, S_1/\sqrt{n}), (2, S_2/\sqrt{n}), \cdots, (n, Sn/\sqrt{n})$ *(linearly interpolated in between the values), is close to the graph of Brownian motion run until time one.*

Once it is shown that the polygonal graph does have a limit, Einstein's predicates (a)–(d) are natural ((b) being the result of the central limit theorem). (e) is not trivial at all and is a big part of the hard work.

Appendix B

Solutions

Exercise 1.1

(a) $\{4\}$

(b) $\{0, 1, 2, 3, 4, 5, 7\}$

(c) $\{0, 1, 3, 5, 7\}$

(d) $\emptyset$

Exercise 1.2

(a) Let $x \in (A \cup B) \cup C$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cup B) \cup C \quad &\Leftrightarrow \quad x \in A \cup B \text{ or } x \in C \\ &\Leftrightarrow \quad x \in A \text{ or } x \in B \text{ or } x \in C \\ &\Leftrightarrow \quad x \in A \text{ or } x \in (B \cup C) \\ &\Leftrightarrow \quad x \in A \cup (B \cup C) \end{aligned}$$

This proves the assertion.

(b) Let $x \in A \cap (B \cup C)$. Then we have the following equivalences:

$$\begin{aligned} x \in A \cap (B \cup C) \quad &\Leftrightarrow \quad x \in A \text{ and } x \in B \cup C \\ &\Leftrightarrow \quad (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C) \\ &\Leftrightarrow \quad (x \in A \cap B) \text{ or } (x \in A \cap C) \\ &\Leftrightarrow \quad x \in (A \cap B) \cup (A \cap C) \end{aligned}$$

This proves the assertion.

(c) Let $x \in (A \cup B)^c$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cup B)^c \quad &\Leftrightarrow \quad x \notin A \cup B \\ &\Leftrightarrow \quad (x \notin A \text{ and } x \notin B) \\ &\Leftrightarrow \quad x \in A^c \text{ and } x \in B^c \\ &\Leftrightarrow \quad x \in A^c \cap B^c \end{aligned}$$

This proves the assertion.

(d) Let $x \in (A \cap B)^c$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cap B)^c \quad &\Leftrightarrow \quad x \notin A \cap B \\ &\Leftrightarrow \quad (x \notin A \text{ or } x \notin B) \\ &\Leftrightarrow \quad x \in A^c \text{ or } x \in B^c \\ &\Leftrightarrow \quad x \in A^c \cup B^c \end{aligned}$$

This proves the assertion.

Exercise 1.3

(a) $A \cap B \cap C^c$

(b) $A \cap B^c \cap C^c$

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

(d) $(A^c \cap B^c \cap C^c)^c$

(e) $(A \cap B \cap C^c) \cup (A \cap B^c \cap C) \cup (A^c \cap B \cap C)$

(f) $(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$

(g) $(A^c \cap B^c \cap C^c) \cup (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$

Exercise 1.4 First of all, we can see that A^c and B^c are not disjoint: any element that is not in A, nor in B will be in $A^c \cap B^c$. Then, $A \cap C$ and $B \cap C$ are disjoint as $(A \cap C) \cap (B \cap C) = A \cap B \cap C = \emptyset \cap C = \emptyset$. Finally, $A \cup C$ and $B \cup C$ are not disjoint as they both contain the elements of C (if this one is not empty).

Exercise 1.5 The standard sample space for this experiment is to consider $\Omega = \{1, 2, 3, 4, 5, 6\}^3$, i.e. the set of all sequences of 3 elements chosen from the set $\{1, 2, 3, 4, 5, 6\}$. In other words,

$$\Omega = \{(1, 1, 1), (1, 1, 2), \ldots, (6, 6, 6)\}.$$

There are $6^3 = 216$ elements in Ω. As Ω is finite and the information is complete, we can choose $\mathcal{F}$ to be the set of all possible subsets of Ω. As for the probability we can consider all outcomes equally likely and assign $P(A) = \frac{\#A}{216}$ for any subset A of Ω. ($\#A$ is the number of elements in A.) For example, $P(\{(1, 1, 1)\}) = 1/216$ and $P(\{(1, 1, 1), (1, 1, 2)\}) = 2/216$.

Exercise 1.6 We can choose $\Omega = \{B, G, R\}$ where B denotes the black chip, G the green one and R the red one. As the set is finite and the information complete, we can choose $\mathcal{F}$ to be the set of all possible subsets of Ω. Namely,

$$\mathcal{F} = \{\emptyset, \{B\}, \{G\}, \{R\}, \{B, G\}, \{B, R\}, \{R, G\}, \Omega\}.$$

We can consider the outcomes are equally likely and then $P(A) = \frac{\#A}{3}$ as in the previous exercise.

Exercise 1.7 See Ash's exercise 1.2.7.

Exercise 2.1 There are 150 people who favor the Health Care Bill, do not approve of Obama's performance and are not registered Democrats.

Exercise 2.2

(a) We have

$$\begin{aligned}(A \cap B) \setminus (A \cap C) &= (A \cap B) \cap (A \cap C)^c = (A \cap B) \cap (A^c \cup C^c)\\ &= (A \cap B \cap A^c) \cup (A \cap B \cap C^c) = A \cap (B \cap C^c) = A \cap (B \setminus C).\end{aligned}$$

(b) We have

$$\begin{aligned}A \setminus (B \cup C) &= A \cap (B \cup C)^c = A \cap (B^c \cap C^c) = (A \cap B^c) \cap C^c\\ &= (A \setminus B) \cap C^c = (A \setminus B) \setminus C.\end{aligned}$$

(c) Let $A = \{1, 2, 3\}$, $B = \{2, 3, 4\}$, $C = \{2, 4, 6\}$. We have $(A \setminus B) \cup C = \{1, 2, 4, 6\}$ and $(A \cup C) \setminus B = \{1, 6\}$. Hence, the proposition is wrong.

Exercise 2.3 See Ash's exercise 1.2.5.

Exercise 3.1

(a) We can choose $\Omega = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}, \mathsf{TT}\}$, where H stands for heads and T for tails. The first letter is the outcome of the first toss and the second letter, the outcome of the second toss.

(b) The sample space Ω is finite and, at step 3, all information is available, so we can choose $\mathcal{F}_3 = \mathcal{P}(\Omega)$, the set of possible subsets of Ω.

(c) At step 2, we do not know the result of the second toss. Hence, if an observable subset contains $x\mathsf{H}$, it has to contain $x\mathsf{T}$, because we would have no way to distinguish both. Hence, we have to choose

$$\mathcal{F}_2 = \{\emptyset, \{\mathsf{HH}, \mathsf{HT}\}, \{\mathsf{TH}, \mathsf{TT}\}, \Omega\}.$$

Other subsets, such as $\{\mathsf{HT}\}$ cannot be observed at this step. Indeed, one would need to know the outcome of the second toss to decide if $\{\mathsf{HT}\}$ happens or not. As for the sets in $\mathcal{F}_2$ above, you do not need to know the second outcome to decide if they happen or not.

(d) At step 1, we know neither of the outcomes, so we can just decide about the probability of the trivial events and we have to pick

$$\mathcal{F}_1 = \{\emptyset, \Omega\}.$$

Exercise 3.2

(a) We can choose Ω to be the set of all sequences of outcomes that are made of only tails and one head at the end. That is

$$\Omega = \{\mathsf{H}, \mathsf{TH}, \mathsf{TTH}, \mathsf{TTTH}, \ldots\}.$$

Notice that we can also consider $\Omega = \mathbb{N}$, where $\omega = n$ means that the game ended at toss number n. We also would like to point out that it is customary to add an outcome Δ called the *cemetery* outcome which corresponds to the case where the experiment never ends.

(b) With the different sample spaces above, we can describe the events below this way.

$$\{\text{Aaron wins}\} = \{\mathsf{H}, \mathsf{TTH}, \mathsf{TTTTH}, \ldots\}, \quad \{\text{Bill wins}\} = \{\mathsf{TH}, \mathsf{TTTH}, \mathsf{TTTTTH}, \ldots\}$$

and

$$\{\text{no one wins}\} = \emptyset.$$

If we consider the case where $\Omega = \mathbb{N}$, we obtain

$$\{\text{Aaron wins}\} = \{\text{odd integers}\}, \quad \{\text{Bill wins}\} = \{\text{even integers}\}$$

and

$$\{\text{no one wins}\} = \emptyset.$$

We notice that if you consider the cemetery outcome Δ, then {no one wins} becomes $\{\Delta\}$ rather than $\emptyset$. We will see later that with a fair coin, the probability that no one wins is 0, hence modelling this by $\emptyset$ is ok. To illustrate why Δ could be useful, imagine that the players play with a coin with two tails faces. Then the only possible outcome is Δ and this one can't have probability 0. Hence, modelling by $\emptyset$ would not be appropriate in this case.

Exercise 3.3 Let's consider the case where we toss a coin twice and let $A = \{\text{we get H on the first toss}\}$ and $B = \{\text{we get T on the second toss}\}$. Hence,

$$A = \{HH, HT\}, \quad B = \{HT, TT\}, \quad \text{and } A \setminus B = \{HH\}.$$

Hence,

$$P(A \setminus B) = P\{HH\} = \frac{1}{4},$$

but

$$P(A) - P(B) = \frac{1}{2} - \frac{1}{2} = 0.$$

Exercise 4.1 (a) There are 6^5 possible outcomes. (b) There are only 6^3 possible outcomes with the first and last rolls being 6. So the probability in question is $6^3/6^5$.

Exercise 4.2 There are $10^3 = 1000$ ways to choose a 3-digit number at random. Now, there are 3 ways to choose the position of the single digit larger than 5, 4 ways to choose this digit (6 to 9) and $6 \cdot 6$ ways to choose the two other digits (0 to 5). Hence, there are $3 \cdot 4 \cdot 6 \cdot 6$ ways to choose a 3-digit number with only one digit larger than 5. The probability then becomes:

$$p = \frac{3 \cdot 4 \cdot 6 \cdot 6}{10^3} = \frac{432}{1000} = 43.2\%.$$

Exercise 4.3

(a) We can apply the principles of counting and choosing each symbol on the license plate in order. We obtain $26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17'576'000$ diffrent license plates.

(b) Similarly, we have $10^3 \times 1 \times 26 \times 26 = 676'000$ license plates starting with an A.

Exercise 4.4 Let A be the event that balls are of the same color, R, Y and G the event that they are both red, yellow and green, respectively. Then, as R, Y and G are disjoint,

$$P(A) = P(R \cup Y \cup G) = P(R) + P(Y) + P(G) = \frac{3}{24} \cdot \frac{5}{18} + \frac{8}{24} \cdot \frac{7}{18} + \frac{13}{24} \cdot \frac{6}{18} = \frac{149}{432} = 0.345.$$

Exercise 5.1 Each hunter has 10 choices: hunter 1 makes one of 10 choices, then hunter 2 makes 1 of 10 choices, etc. So over all there are 10^5 possible options. On the other hand, the number of ways to get 5 ducks shot is: 10 for hunter 1 then 9 for hunter 2, etc. So $10 \times 9 \times 8 \times 7 \times 6$ ways. The answer thus is the ratio of the two numbers: $\frac{10\times9\times8\times7\times6}{10^5}$.

Exercise 5.2 There are $\binom{64}{8}$ ways to place 8 rooks on a chessboard. If we want the rooks not to check each other, we place the first rook on column 1 (8 ways) then the second rook on column 2 but not on the same row as the first rook (7 ways) and so on. I total there are 8! ways to do this. So the probability the 8 rooks are not checking each other is $8!/\binom{64}{8}$ which equals the claimed number. (Check this last claim yourself!)

Exercise 5.3 We put the women together in order to form one entity. Hence, this problem comes back to seating $m+1$ entity (m men and one entity for the women). We have $(m+1)!$ ways to do it. Then, among the women together, we have $w!$ ways to seat them. As we have $(m+w)!$ ways to seat this people, the probability is

$$P(\text{women together}) = \frac{(m+1)!w!}{(m+w)!}.$$

Exercise 5.4 See Ash's exercise 1.4.7.

Exercise 5.5

(a) There are $\binom{54}{6}$ possible combinations of 6 numbers (the order doesn't matter). Only one of them will match the one you played. Hence, the probability to win the first prize is

$$p = \frac{1}{\binom{54}{6}} = \frac{1}{25,827,165}.$$

(b) We have $\binom{6}{5}\binom{48}{1}$ ways to choose a combination of 6 numbers that shares 5 numbers with the one played (5 numbers out of the 6 played and 1 out of the 48 not played). Hence, the probability to win the second prize is

$$p = \frac{\binom{6}{5}\cdot\binom{48}{1}}{\binom{54}{6}} = \frac{6\cdot 48}{\binom{54}{6}} = \frac{288}{25,827,165} = \frac{3}{269,033}.$$

Exercise 5.6

(a) There are $\binom{50}{5}$ possible combinations of 5 numbers in the first list and $\binom{9}{2}$ combinations of 2 numbers in the second list. That makes $\binom{50}{5}\cdot\binom{9}{2}$ possible results for this lottery. Only one will match the combination played. Hence, the probability to win the first prize is

$$p = \frac{1}{\binom{50}{5}\binom{9}{2}} = \frac{1}{76,275,360}$$

(b) Based only on the probability to win the first price, you would definitely choose the first one which has a larger probability of winning.

Exercise 5.7 The number of possible poker hands is $\binom{52}{5}$ as we have seen in class.

(a) We have 13 ways to choose the value for the four cards. The suits are all taken. Then, there are 48 ways left to choose the fifth card. Hence, the probability to get four of a kind is

$$\mathrm{P}(\text{four of a kind}) = \frac{13 \cdot 48}{\binom{52}{5}} = 0.00024.$$

(b) We have 13 ways to choose the value for the three cards. The, $\binom{4}{3}$ ways to choose the suits. Then, there are $\binom{12}{2} \cdot 4 \cdot 4$ ways left to choose the last two cards (both of different values). Hence, the probability to get three of a kind is

$$\mathrm{P}(\text{three of a kind}) = \frac{13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4^2}{\binom{52}{5}} = 0.0211.$$

(c) In order to make a straight flush, we have 10 ways to choose the highest card of the straight and 4 ways to choose the suit. Hence, the probability to get a straight flush is

$$\mathrm{P}(\text{straight flush}) = \frac{10 \cdot 4}{\binom{52}{5}} = 0.0000154.$$

(d) In order to make a flush, we have 4 ways to choose the suit and then $\binom{13}{5}$ ways to choose the five cards among the 13 of the suit selected. Nevertheless, among those flushes, some of them are straight flushes, so we need to subtract the number of straight flushes obtained above. Hence, the probability to get a flush is

$$\mathrm{P}(\text{flush}) = \frac{4 \cdot \binom{13}{5} - 40}{\binom{52}{5}} = 0.00197.$$

(e) In order to make a straight, we have 10 ways to choose the highest card of the straight and then 4^5 ways to choose the suits (4 for each card). Nevertheless, among those straights, some of them are straight flushes, so we need to subtract the number of straight flushes obtained above. Hence, the probability to get a straight is

$$\mathsf{P}(\text{straight}) = \frac{10 \cdot 4^5 - 40}{\binom{52}{5}} = 0.00392.$$

Exercise 6.1 Observe first that once people are seated moving everyone one seat to the right gives the same seating arrangement! So at a round table the first person can sit anywhere. Then, the next person has $n-1$ possible places to sit at, the next has $n-2$ places, and so on. In total, there are $(n-1)!$ ways to seat n people at a round table.

Exercise 6.2

(a) For the draw with replacement, there are 52^{10} possible hands. If we want no to cards to have the same face values, we have $13 \cdot 12 \cdots 4$ ways to pick diffrent values and then, 4^{10} ways to choose the suits (4 for each card drawn). Hence, the probability becomes

$$p = \frac{13 \cdot \cdots \cdot 4 \cdot 4^{10}}{52^{10}} = 0.00753.$$

(b) In the case of the draw without replacement, we have $\binom{52}{10}$ possible hands. The number of hands that have at least 9 card of the same suit can have 9 of them or 10 of them. The first case corresponds to $4 \cdot \binom{13}{9} \cdot 39$ possibilities (4 possible suits, 9 cards out of this suit and 1 additionnal card from the 39 remaining) and the second case corresponds to $4 \cdot \binom{13}{10}$ possibilities. Hence, the probability becomes

$$p = \frac{4 \cdot \binom{13}{9} \cdot 39 + 4 \cdot \binom{13}{10}}{\binom{52}{10}} = 0.00000712.$$

Exercise 6.3 As the order doesn't count, we have $\binom{10}{5}$ possible ways to draw the balls. If we want the second largest number to be 8, we need to pick the 8, then pick one larger number among the two possible and pick 3 numbers among the 7 lower numbers. Hence, the probability becomes

$$p = \frac{\binom{2}{1} \cdot \binom{7}{3}}{\binom{10}{5}} = 0.2778.$$

Exercise 6.4 See Ash's exercise 1.4.8.

Exercise 6.5

(a) This comes back to counting the number of permutations of 8 different people. Hence, there are $8!$ $(= 40320)$ possibles ways to seat those 8 people in a row.

(b) People A and B want to be seated together. Hence, we will consider them as one single entity that we will first treat as a single person. Hence, we will assign one spot to each person and one spot to the entity AB. There are 7 entities (6 people and the group AB). There are $7!$ ways to seat them. For each of these ways, A

and B can be seated in 2 different ways in the group. As a consequence, there are $2 \cdot 7!$ $(= 10080)$ possible ways to seat these 8 people with A and B seated together.

(c) First of all, notice that there are two possible ways to sit men and women in alternance, namely

$$wmwmwmwm \text{ or } mwmwmwmw,$$

where w stands for a woman and m for a man. Then, for each of the repartitions above, we have to choose the positions of the women among themselves. There are 4! permutations. For each repartition of the women, we need to choose the positions of the men. There are 4! permutations as well. Hence, there are $2{\cdot}4!{\cdot}4!$ $(= 1152)$ ways to seat 4 women and 4 men in alternance.

(d) Similarly as in (b), the 5 men form an entity that we will treat as a single person. Then, there are 4 entities (3 women and 1 group of men) to position. There are 4! ways to do it. For each of these ways, the 5 men can be seated differently on the 5 consecutive chairs they have. There are 5! to do it. Hence, there are $4! \cdot 5!$ $(= 2880)$ possible ways to seat those 8 people with the 5 men seated together.

(e) We consider that each married couple forms an entity that we will treat as a single person. There are then 4! ways to assign seats to the couples. For each of these repartitions, there are two ways to seat each person within the couple. Hence, there are $4! \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 4! \cdot 2^4$ $(= 384)$ possible ways to seat 4 couples.

Exercise 6.6

(a) There are 6 discs to store on the shelf. As they are all different, there are $6!$ $(= 720)$ ways to do it.

(b) Assume the classical discs, as well as the jazz discs form two entities, that we will consider as a single disc. Then, there are 3 entities to store and 3! ways to do it. For each of these repartitions, the classical discs have 3! ways to be stored within the group and the jazz discs hav 2 ways to be stored within the group. Globally, there are $3! \cdot 3! \cdot 2$ $(= 72)$ ways to store the 6 discs respecting the styles.

(c) If only the classical discs have to be stored together, we have 4 entities (the classical group, the three other discs). We have then 4! ways to assign their position. For each of their repartitions, we have 3! to store the classical discs within the group.

Hence, we have $4! \cdot 3!$ ways to store the discs with the classical together. Nevertheless, among those repartitions, some of them have the jazz discs together, which we don't want. Hence, we subtract to the number above, the number of ways to store the discs according to the styles found in (b). Hence, there are $(4! \cdot 3!) - (3! \cdot 3! \cdot 2)$ $(= 144 - 72 = 72)$ ways to store the discs with only the classical together.

Exercise 6.7

(a) The 5 letters of the word "bikes" being different, there are 5! (= 120) ways to form a word.

(b) Among the 5 letters of the word "paper", there are two p's. First choose their position, we have $\binom{5}{2}$ ways to do it. Then, there are 3! ways to position the other 3 different letters. Hence, we have $\binom{5}{2} \cdot 3! = \frac{5!}{2!} = 60$ possible words.

(c) First choose the positions of the e's, then of the t's and finally the ones of the other letters. Hence, we have $\binom{6}{2}\binom{4}{2} \cdot 2! = \frac{6!}{2!2!} = 180$ possible words.

(d) Choose the poisition of the three m, then the ones of the two i's and finally the ones of the other different letters. Hence, we have $\binom{7}{3}\binom{4}{2} \cdot 2! = \frac{7!}{3!2!} = 420$ possible words.

Exercise 7.1 In $(a+b)^8$ the coefficient of b^5 is $\binom{8}{5} \times a^3$. Hence if $a = 2$ and $b = 3x$, the coefficient of x^5 is $\binom{8}{5} \times 2^3 \times 3^5$.

Exercise 7.2 See Ash's exercise 1.4.9.

Exercise 7.3 In order to prove that

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0},$$

we will consider a group of people made of m women and n men. Let $0 \leqslant r \leqslant \min(m, n)$. We will count the number of teams of r people that we can form from the $m+n$ people available. On the one hand, this number is $\binom{m+n}{r}$. On another hand, we can count the number of teams of r people with k women (and $r-k$ men), with $k \leqslant r$. We have $\binom{m}{k}\binom{n}{r-k}$ of them. Summing the number of such teams over all possible values of k, we obtain the total number of teams of r people, namely

$$\sum_{k=0}^{r} \binom{n}{k}\binom{m}{r-k} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}.$$

The two results are solutions of a same combinatorial problem and have to be equal. The result is proved.

Exercise 7.4

(a) The number of players in the team is not fixed. If we have to form a k-player team, we have $\binom{n}{k}$ ways to pick the players. Then, there are k ways to choose the captain of this team. Summing from $k = 1$ to n, we obtain $\sum_{k=1}^{n} k\binom{n}{k}$ possible teams. On another hand, we can first choose a captain (n ways), and then choose for each of the $n-1$ remaining people if they are part of the team or not (2^{n-1} ways). That gives $n\,2^{n-1}$ possible teams.

(b) In that case, we proceed as in (a), but we choose a captain and an assistant-captain. On the left hand-side, $k(k-1)$ represents the number of ways to choose the captain and his assistant. On the right-hand side, we first choose the captain and the assistant ($n(n-1)$). There are $n-2$ people left to be part or not of the team. That gives the result.

(c) The binomial theorem gives

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

Taking derivatives on both sides with respect to x, we have

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

Taking $x = 1$ in the equation above gives (a). Differentiating a second time, we get

$$n(n-1)(1+x)^{n-2} = \sum_{k=2}^{\infty} k(k-1)\binom{n}{k}x^{k-2}.$$

Taking $x = 1$ gives (b).

Exercise 7.5

(a) For each digit, except the zero, we can build a 4-digit number. Hence, there are 9 possible numbers.

(b) Two cases can occur. First of all, the number of ways to build a 4-digit number made of two pairs of different digits, different from 0 is $\binom{9}{2}\binom{4}{2}$. Indeed, we choose two digits among 9 and two places among four two place the first-type digit. Secondly, if one of the pairs is a pair of 0's, we have $9 \cdot \binom{3}{2}$ possible numbers. Indeed, there are 9 ways to choose the second pair and we need to choose two spots among three for the 0's (we cannot put the 0 upfront). Finally, there are $\binom{9}{2}\binom{4}{2} + 9 \cdot \binom{3}{2} = 243$ 4-digit numbers made of two pairs of two different digits.

(c) We again distinguish the cases with or without 0. There are $\binom{9}{4} \cdot 4!$ numbers with 4 different digits without 0. Indeed, we choose 4 digits among 9 that we can place in any order. Moreover, there are $\binom{9}{3} \cdot 3 \cdot 3!$ 4-digit numbers with 0. We choose the three other numbers among 9, the position of 0 and finally, we can place the others in any order. Hence, there are $\binom{9}{4} \cdot 4! + \binom{9}{3} \cdot 3 \cdot 3! = 4536$ 4-digit numbers with different digits.

(d) In the case where the number have to be ordered in increasing order, there are $\binom{9}{4}$ ways to choose the 4 different digits (0 can't be chosen) and only one way to place them in order. Hence, there are $\binom{9}{4}$ 4-digit ordered numbers.

(e) In (a), there are 9 possible numbers, for any value of n. In (d), following the same argument as for $n = 4$, we notice that there are $\binom{9}{n}$ n-digit ordered numbers for $1 \leqslant n < 10$. There are none of them for $n \geqslant 10$. In (c), for $2 \leqslant n \leqslant 9$, we have $\binom{9}{n} \cdot n!$ n-digit numbers with different digits without 0 and $\binom{9}{n-1} \cdot (n-1) \cdot (n-1)!$ numbers with 0. Hence, we have $\binom{9}{n} \cdot n! + \binom{9}{n-1} \cdot (n-1) \cdot (n-1)! = \frac{9 \cdot 9!}{(10-n)!}$ n-digit numbers with different digits. There are $9 \cdot 9!$ for $n = 10$ and none for $n > 10$.

Exercise 8.1 See Ash's exercise 1.6.1.

Exercise 8.2 Let's assume $n \geqslant 3$, otherwise the answer is 0. We will denote by X the number of heads that we obtain. We want to find $P\{X \geqslant 3 | X \geqslant 1\}$. We have

$$\begin{aligned} P\{X \geqslant 3 | X \geqslant 1\} &= \frac{P(\{X \geqslant 3\} \cap \{X \geqslant 1\})}{P\{X \geqslant 1\}} = \frac{P\{X \geqslant 3\}}{P\{X \geqslant 1\}} \\ &= \frac{1 - \frac{1}{2^n}(1 + n + \binom{n}{2})}{1 - \frac{1}{2^n}} = \frac{2^n - 1 - n - \binom{n}{2}}{2^n - 1}, \end{aligned}$$

where we used that the probability to get k heads out of n tosses is given by $\binom{n}{k}\frac{1}{2^n}$.

Exercise 8.3 Let F denote the event the a fair coin is used and H the event that the first n outcome of the coin are heads. We want to find $P(F|H)$. We know that

$$P(F) = P\{\text{outcome of the die is odd}\} = \frac{1}{2}$$

and that

$$P(H|F) = 2^{-n} \qquad P(H|F^c) = p^n.$$

We can use Bayes' theorem to obtain

$$P(F|H) = \frac{P(H|F)P(F)}{P(H|F)P(F) + P(H|F^c)P(F^c)} = \frac{2^{-n} \cdot \frac{1}{2}}{2^{-n} \cdot \frac{1}{2} + p^n \cdot \frac{1}{2}} = \frac{2^{-n}}{2^{-n} + p^n}.$$

Exercise 8.4 We will use the Law of Total Probability with an infinite number of events. Indeed, for every $n \geqslant 1$, the events $\{I = n\}$ are disjoint (we can't choose two different integers) and their union is Ω (one integer is necessarily chosen). Hence, letting H denote the event that the outcome is heads, we have

$$P(H|I = n) = e^{-n}.$$

Then, by the Law of Total Probability, we have

$$P(H) = \sum_{n=1}^{\infty} P(H|I = n)P\{I = n\} = \sum_{n=1}^{\infty} e^{-n}\, 2^{-n} = \sum_{n=1}^{\infty} (2e)^{-n} = \frac{1}{1 - \frac{1}{2e}} - 1 = \frac{1}{2e - 1},$$

because $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ for $|x| < 1$.

Exercise 8.5 See Ash's exercise 1.6.5.

Exercise 8.6 See Ash's exercise 1.6.6.

Exercise 8.7 Let D denote the event that a random person has the disease, P the event that the test is positive and R the event that the person has the rash. We want to find $P(D|R)$. We know that

$$P(D) = 0.2 \quad P(P|D) = 0.9 \quad P(P|D^c) = 0.3 \text{ and } P(R|P) = 0.25.$$

First of all, let's notice that we have $P(R|D) = P(R|P\cap D)P(P|D) = 0.25\cdot 0.9 = 0.225$ and $P(R|D^c) = P(R|P \cap D^c)P(P|D^c) = 0.25 \cdot 0.3 = 0.075$. Now, by Bayes' theorem, we have

$$P(D|R) = \frac{P(R|D)P(D)}{P(R|D)P(D) + P(R|D^c)P(D^c)} = \frac{0.225 \cdot 0.2}{0.225 \cdot 0.2 + 0.075 \cdot 0.8} = \frac{0.045}{0.105} = \frac{3}{7}.$$

Exercise 8.8 Let A denote the event "the customer has an accident within one year" and let R denote the event "the customer is likely to have accidents".

(a) We want to find $P(A)$. By the Law of Total Probability, we have

$$P(A) = P(A \mid R)P(R) + P(A \mid R^c)P(R^c) = (0.4 \times 0.3) + (0.2 \times 0.7) = 0.26.$$

(b) We want to compute $P(R \mid A)$. The defnition of conditional proabability leads to

$$P(R \mid A) = \frac{P(A \mid R)P(R)}{P(A)} = \frac{0.4 \times 0.3}{0.26} = 0.46,$$

where we used the result in (a).

Exercise 8.9 Let R_i denote the event "the receiver gets an i" and E_i the event "the transmitter sends an i" ($i \in \{0, 1\}$).

(a) We want to find $P(R_0)$. By the Law of Total Probability,

$$P(R_0) = P(R_0 \mid E_0)P(E_0) + P(R_0 \mid E_1)P(E_1) = (0.8\times 0.45) + (0.1\times 0.55) = 0.415,$$

as $E_0 = E_1^c$.

(b) We want to compute $P(E_0 \mid R_0)$. The definition of conditional probability leads to

$$P(E_0 \mid R_0) = \frac{P(R_0 \mid E_0)P(E_0)}{P(R_0)} = \frac{0.8 \times 0.45}{0.415} = 0.867,$$

where we used the result in (a).

Exercise 8.10 Let I,L and C be the events "the voter is independent, democrat or republican", respectively. Let V be the event "he actually voted in the election".

(a) By the Law of Total Probability, we have

$$P(V) = P(V|I)P(I) + P(V|L)P(L) + P(V|C)P(C) = 0.4862.$$

(b) We first compute $P(I|V)$. By Bayes' theorem, we have

$$P(I|V) = \frac{P(V|I)P(I)}{P(V)} = \frac{0.35 \cdot 0.46}{0.4862} = 0.331.$$

Similarly, we have

$$P(L|V) = \frac{P(V|L)P(L)}{P(V)} = \frac{0.62 \cdot 0.30}{0.4862} = 0.383,$$

and

$$P(C|V) = \frac{P(V|C)P(C)}{P(V)} = \frac{0.58 \cdot 0.24}{0.4862} = 0.286.$$

Exercise 8.11 Let A_n be the event "John drives on the n-th day" and R_n be the event "he is late on the n-th day".

(a) Let's compute $P(A_n)$, we have

$$\begin{aligned} P(A_n) &= P(A_n|A_{n-1})P(A_{n-1}) + P(A_n|A^c_{n-1})P(A^c_{n-1}) \\ &= \frac{1}{2}P(A_{n-1}) + \frac{1}{4}(1 - P(A_{n-1})) \\ &= \frac{1}{4}P(A_{n-1}) + \frac{1}{4}, \end{aligned}$$

where the event A^c_n stands for "John takes the train on the n-th day." Iterating this formula $n-1$ times, we obtain

$$\begin{aligned} P(A_n) &= \left(\frac{1}{4}\right)^{n-1} P(A_1) + \sum_{i=1}^{n-1}\left(\frac{1}{4}\right)^i = \left(\frac{1}{4}\right)^{n-1} p + \frac{1}{4}\left(\frac{1-(\frac{1}{4})^{n-1}}{1-\frac{1}{4}}\right) \\ &= \left(\frac{1}{4}\right)^{n-1} p + \frac{1}{3}\left(1 - \left(\frac{1}{4}\right)^{n-1}\right). \end{aligned}$$

(b) By the Law of Total Probability, we have

$$\begin{aligned} P(R_n) &= P(R_n|A_n)P(A_n) + P(R_n|A^c_n)P(A^c_n) \\ &= \frac{1}{2}P(A_n) + \frac{1}{4}(1 - P(A_n)) \\ &= \frac{1}{4}P(A_n) + \frac{1}{4} = P(A_{n+1}). \end{aligned}$$

By (a), we then have

$$P(R_n) = \left(\frac{1}{4}\right)^n p + \frac{1}{3}\left(1 - \left(\frac{1}{4}\right)^n\right).$$

(c) Let's compute $\lim_{n\to\infty} P(A_n)$. We know that $\lim_{n\to\infty}(\frac{1}{4})^{n-1} = 0$. Hence, $\lim_{n\to\infty} P(A_n) = \frac{1}{3}$. Similarly, we have $\lim_{n\to\infty} P(R_n) = \lim_{n\to\infty} P(A_{n+1}) = \frac{1}{3}$.

Exercise 9.1

(a) The events "'getting a spade"' and "'getting a heart"' are disjoint but not independent.

(b) The events "'getting a spade"' and "'getting a king"' are independent (check the definition) and not disjoint: you can get the king of spades.

(c) The events "'getting a king"' and "'getting a queen and a jack"' are disjoint (obvious) and independent. As the probability of the second event is zero, this is easy to check.

(d) The events "'getting a heart"' and "'getting a red king"' are not disjoint and not independent.

Exercise 9.2 See Ash's exercise 1.5.4.

Exercise 9.3 The number of ones (resp. twos) is comprised between 0 and 6. Hence, we have the following possibilities : three ones and no two, four ones and one two. (Other possibilities are not compatible with the experiment.) Hence, noting A the event of which we want the probability, we have

$$\begin{aligned} P(A) &= P\{\text{three 1's, no 2 (and three others)}\} + P\{\text{four one's, one two (and one other)}\} \\ &= \frac{\binom{6}{3}\cdot 4^3}{6^6} + \frac{\binom{6}{4}\cdot\binom{2}{1}\cdot 4}{6^6}. \end{aligned}$$

Indeed, we have to choose 3 positions among 6 for the ones and four choices for each of the other values for the first probability and we have to choose 4 positions among 6 for the ones, one position among the two remaining for the two and we have 4 choices for the last value for the second probability. The total number of results is 6^6 (six possible values for each roll of a die).

Exercise 9.4

(a) If A is independent of itself, then $P(A) = P(A \cap A) = P(A)P(A) = P(A)^2$. The only possible solutions to this equation are $P(A) = 0$ or $P(A) = 1$.

(b) Let B be any event. If $P(A) = 0$, then $A \cap B \subset A$, hence $0 \leqslant P(A \cap B) \leqslant P(A) = 0$. As a consequence, $P(A \cap B) = 0 = P(A)P(B)$. On another hand, if $P(A) = 1$, then $P(A^c) = 0$. Hence, by the first part, A^c is independent of any event B. This implies that A is independent of any event B by the properties of independence.

Exercise 9.5 The sample space for this experiment is

$$\Omega = \{(P,P,P),(P,P,F),(P,F,P),(P,F,F),(F,P,P),(F,P,F),(F,F,P),(F,F,F)\}.$$

All outcomes are equally likely and then, $P\{\omega\} = \frac{1}{8}$, for all $\omega \in \Omega$. Moreover, counting the favorable cases for each event, we see that

$$\begin{aligned} P(G_1) &= \frac{4}{8} = \frac{1}{2} = P(G_2) = P(G_3) \\ P(G_1 \cap G_2) &= \frac{2}{8} = \frac{1}{4} = P(G_1)P(G_2). \end{aligned}$$

Similarly, we find that $P(G_1 \cap G_3) = P(G_1)P(G_3)$ and that $P(G_2 \cap G_3) = P(G_2)P(G_3)$. The events G_1, G_2 and G_3 are pairwise independent.

However,

$$P(G_1 \cap G_2 \cap G_3) = \frac{2}{8} = \frac{1}{4} \neq P(G_1) \cdot P(G_2) \cdot P(G_3) = \frac{1}{8},$$

hence G_1, G_2 and G_3 are *not* independent. Actually, it is to see that if G_1 and G_2 occur, then G_3 occurs as well, which explains the dependence.

Exercise 9.6 We consider that having 4 children is the result of 4 independent trials, each one being a success (girl) with probability 0.48 or a failure (boy) with probability 0.52. Let E_i be the event "the i-th child is a girl".

(a) Having children with all the same gender corresponds to the event { 4 successes or 0 success}. Hence, P("all children have the same gender") = P("4 successes") + P("0 success") = $(0.48)^4 + (0.52)^4$.

(b) The fact that the three oldest children are boys and the youngest is a girl corresponds to the event $E_1^c \cap E_2^c \cap E_3^c \cap E_4$. Hence P("three oldest are boys and the youngest is a girl") = $(0.52)^3(0.48)$.

(c) Having three boys comes back to having 1 success among the 4 trials. Hence, P("exactly three boys") = $\binom{4}{3}(0.52)^3(0.48)$.

(d) The two oldest are boys, the other do not matter. This comes back to having two failures among the first two trials. Hence, P("the two oldest are boys") = $(0.52)^2$.

(e) Let's first compute the probability that there is no girl. This equals the probability of no sucess, that is $(0.52)^4$. Hence, P("at least one girl") = 1 − P("no girl") = $1 - (0.52)^4$.

Exercise 10.1 The sample space is $\Omega = \{H, T\}^3 := \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. These eight outcomes are equally likely, hence the probability measure is given by $P\{\omega\} = \frac{1}{8}$ for all $\omega \in \Omega$. The random variable X can be defined by

$$X(\omega) = \sum_{i=1}^{3} 1_{\{H\}}(\omega_i) \quad \text{when} \quad \omega = (\omega_1\ \omega_2\ \omega_3).$$

Otherwise, one can define X this way :

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = HHH, \\ 1 & \text{if } \omega = HHT, HTH, THH, \\ 2 & \text{if } \omega = HTT, THT, TTH, \\ 3 & \text{if } \omega = TTT. \end{cases}$$

Exercise 10.2 The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}^3 := \{(\omega_1, \omega_2, \omega_3) : \omega_1, \omega_2, \omega_3 \in \{1, 2, 3, 4, 5, 6\}\}$. There are 216 equally likely outcomes, hence the probability measure is given by $P\{\omega\} = \frac{1}{216}$ for all $\omega \in \Omega$. The random variable X can be defined by

$$X(\omega) = \omega_1 \cdot \omega_2 \cdot \omega_3 \quad \text{when} \quad \omega = (\omega_1, \omega_2, \omega_3).$$

Exercise 11.1 See Ash's exercise 2.3.2.

Exercise 11.2 $P(X = k) = \binom{3}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{3-k}$, which gives

x	0	1	2	3
$f(x)$	$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$

$f(x) = 0$ for all $x \neq 0, 1, 2, 3$.

Exercise 11.3

(a)

$$f(x) = \begin{cases} \frac{1}{36} & \text{if } x = 1, 9, 16, 25 \text{ or } 36, \\ \frac{1}{18} & \text{if } x = 2, 3, 5, 8, 10, 15, 18, 20, 24 \text{ or } 30, \\ \frac{1}{12} & \text{if } x = 4, \\ \frac{1}{9} & \text{if } x = 6 \text{ or } 12, \\ 0 & \text{otherwise.} \end{cases}$$

(b)

x	1	2	3	4	5	6
$f(x)$	$\frac{1}{36}$	$\frac{1}{12}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{1}{4}$	$\frac{11}{36}$

$f(x) = 0$ for all $x \neq 0, \ldots, 6$.

Exercise 11.4 The random variable X counts the number of even outcomes, when we roll a fair die twice. Its probability mass function is

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$f(x) = 0$ for all $x \neq 0, 1, 2$.

Exercise 11.5

(a) There are $\binom{5}{3} = 10$ ways of picking the balls. The maximum number can only be $3, 4$ or 5.

x	1	2	3	4	5
$f(x)$	0	0	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{6}{10}$

$f(x) = 0$ for all $x \neq 1, \ldots, 5$.

(b) The minimum number can only be $1, 2$ or 3.

x	1	2	3	4	5
$f(x)$	$\frac{6}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	0	0

$f(x) = 0$ for all $x \neq 1, \ldots, 5$.

Exercise 12.1

(a) The random variable X has a geometric distribution with parameter p, hence $P\{X = n\} = p(1-p)^{n-1}$. Then,

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

by the standard formula for geometric series.

(b) The random variable Y has a Poisson distribution with parameter λ, hence $P\{Y = n\} = e^{-\lambda}\frac{\lambda^n}{n!}$. Then,

$$\sum_{n=0}^{\infty} P\{Y = n\} = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda} = 1,$$

by the standard series expansion for exponentials.

Exercise 12.2

(a) Let X be the r.v. counting the number of cars having an accident this day. The r.v. X has a binomial distribution with parameters $n = 10,000$ and $p = 0.002$. As p is small, n is large and np is not too large, nor too small, we can approximate X by a Poisson random variable with parameter $\lambda = np = 20$. Then, we have

$$P\{X = 15\} \simeq e^{-\lambda}\frac{\lambda^{15}}{15!} = e^{-20}\frac{20^{15}}{15!} = 5.16\%.$$

We notice that the exact value of $P\{X = 15\}$ would be precisely 5.16%.

(b) As above, let Y be the r.v. counting the number of gray cars having an accident this day. By a similar argument as in (a) and as one car out of 5 is gray, the r.v. Y follows a binomial random variable with parameters $n = 2,000$ and $p = 0.002$. We can again approximate by a Poisson distribution of parameter $\lambda = np = 4$. Then, we have

$$P\{Y = 3\} \simeq e^{-\lambda}\frac{\lambda^{3}}{3!} = e^{-4}\frac{4^{3}}{3!} = 19.54\%.$$

The exact value would be $P\{Y = 3\} = 19.55\%$.

Exercise 13.1

(a) We can easilly check that $F(x)$ is a non-decreasing function, that $\lim_{x\to\infty} F(x) = 1$, that $\lim_{x\to-\infty} F(x) = 0$ and that F is right-continuous. (A plot can help.) Hence, F is a cumulative distribution function.

(b) The random variable X is neither discrete, nor continuous. Indeed, F has jumps, which prevents it from being continuous and the portions between jumps are not always constant.

(c) We will use the properties of CDFs to compute the probabilities. Namely, we have

$$P\{X = 2\} = F(2) - F(2-) = \left(\frac{1}{6}\cdot 2 + \frac{1}{3}\right) - \frac{1}{3} = \frac{1}{3}.$$

(d) $P\{X < 2\} = F(2-) = \lim_{x\uparrow 2} \frac{1}{3} = \frac{1}{3}$.

(e) As the two events are disjoint, we have

$$\begin{aligned} P\left\{X = 2 \text{ or } \frac{1}{2} \leqslant X < \frac{3}{2}\right\} &= P\{X = 2\} + P\left\{\frac{1}{2} \leqslant X \leqslant \frac{3}{2}\right\} \\ &= P\{X = 2\} + \left(F\left(3/2-\right) - F\left(1/2-\right)\right) \\ &= \frac{1}{3} + \left(\frac{1}{3} - \frac{1}{12}\right) = \frac{7}{12} \end{aligned}$$

(f) As 2 is included in $[\frac{1}{2};3]$, we have

$$\begin{aligned} P\left\{X = 2 \text{ or } \frac{1}{2} \leqslant X \leqslant 3\right\} &= P\left\{\frac{1}{2} \leqslant X \leqslant 3\right\} \\ &= F(3) - F\left(1/2-\right) \\ &= \frac{5}{6} - \frac{1}{12} = \frac{3}{4}. \end{aligned}$$

Exercise 14.1

(a) We need to find c such that $\int_{-\infty}^{+\infty} f(x)dx = 1$. In order for f to be a pdf, we need $c > 0$. Moreover,

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)dx &= c\int_{-2}^{2}(4-x^2)dx = c\left(4x - \frac{x^3}{3}\right)\Big|_{-2}^{2} \\ &= c\left(\left(8-\frac{8}{3}\right) - \left(-8+\frac{8}{3}\right)\right) = \frac{32}{3}c.\end{aligned}$$

Hence, $c = \frac{3}{32}$.

(b) The cdf F of X is given by $F(x) = \int_{-\infty}^{x} f(u)du$. As $x \leqslant -2$, the integral vanishes. Moreover, as $x \geqslant 2$, we have $\int_{-\infty}^{x} f(u)du = \int_{-\infty}^{+\infty} f(u)du = 1$. Then, for $-2 < x < 2$,

$$\begin{aligned}\int_{-\infty}^{x} f(u)du &= \frac{3}{32}\int_{-2}^{x}(4-u^2)du = \frac{3}{32}\left(4u - \frac{u^3}{3}\right)\Big|_{-2}^{x} \\ &= \frac{3}{32}\left(\left(4x-\frac{x^3}{3}\right) - \left(-8+\frac{8}{3}\right)\right) = \frac{3}{32}\left(4x - \frac{x^3}{3} + \frac{16}{3}\right) \\ &= \frac{1}{32}\left(16+12x-x^3\right).\end{aligned}$$

Finally,

$$F(x) = \begin{cases} 0 & \text{if } x \leqslant -2, \\ \frac{1}{32}(16+12x-x^3) & \text{if } -2 < x < 2, \\ 1 & \text{if } x \geqslant 2. \end{cases}$$

Exercise 14.2

(a) We need to find c such that $\int_{-\infty}^{+\infty} f(x)dx = 1$. In order for f to be a pdf, we need $c > 0$. Moreover,

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)dx &= c\int_{0}^{\frac{\pi}{2}}\cos^2(x)\,dx = c\int_{0}^{\frac{\pi}{2}}\frac{1+\cos(2x)}{2}\,dx \\ &= c\left(\frac{x}{2} + \frac{\sin(2x)}{4}\right)\Big|_{0}^{\frac{\pi}{2}} = \frac{\pi c}{4}.\end{aligned}$$

Hence, $c = \frac{4}{\pi}$.

(b) The cdf F of X is given by $F(x) = \int_{-\infty}^{x} f(u)du$. As $x \leqslant 0$, the integral vanishes. Moreover, as $x \geqslant \frac{\pi}{2}$, we have $\int_{-\infty}^{x} f(u)du =$

$\int_{-\infty}^{+\infty} f(u)du = 1$. Then. for $0 < x < \frac{\pi}{2}$,

$$\begin{aligned}\int_{-\infty}^{x} f(u)du &= \frac{4}{\pi}\int_0^x \cos^2(u)\,du = \frac{4}{\pi}\left(\frac{u}{2}+\frac{\sin(2u)}{4}\right)\Big|_0^x \\ &= \frac{2}{\pi}\left(x+\frac{\sin(2x)}{2}\right).\end{aligned}$$

Finally,

$$F(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ \frac{2}{\pi}\left(x+\frac{\sin(2x)}{2}\right) & \text{if } 0 < x < \frac{\pi}{2}, \\ 1 & \text{if } x \geqslant \frac{\pi}{2}. \end{cases}$$

Exercise 14.3 For each question, we need to find the right set (union of intervals) and integrate f over it. The fact that for $a, b > 0$, $\int_a^b f(x)\,dx = \frac{1}{2}(e^{-a} - e^{-b})$ is used throughout.

(a) By symmetry around 0, we have

$$P\{|X| \leqslant 2\} = 2 \cdot P\{0 \leqslant X \leqslant 2\} = 2 \cdot \frac{1}{2}(1-e^{-2}) = 1 - e^{-2}.$$

(b) We have $\{|X| \leqslant 2 \text{ or } X \geqslant 0\} \Leftrightarrow \{X \geqslant -2\}$. Hence,

$$\begin{aligned}P\{|X| \leqslant 2 \text{ or } X \geqslant 0\} &= \int_{-2}^{\infty} f(x)\,dx = \frac{1}{2}\int_{-2}^{0} e^x\,dx + \frac{1}{2}\int_0^{\infty} f(x)\,dx \\ &= \frac{1}{2}(1-e^{-2}) + \frac{1}{2} = 1 - \frac{1}{2}e^{-2}.\end{aligned}$$

(c) We have $\{|X| \leqslant 2 \text{ or } X \leqslant -1\} \Leftrightarrow \{X \leqslant 2\}$. Moreover, by symmetry, $P\{X \leqslant 2\} = P\{X \geqslant -2\} = 1 - \frac{1}{2}e^{-2}$, by the result in (b).

(d) The condition $|X| + |X-3| \leqslant 3$ corresponds to $0 \leqslant X \leqslant 3$. Hence,

$$P\{|X| + |X-3| \leqslant 3\} = P\{0 \leqslant X \leqslant 3\} = \frac{1}{2}(1-e^{-3}).$$

(e) We have $X^3 - X^2 - X + 2 = (X-2)(X^2+X+1)$. Hence, $X^3 - X^2 - X + 2 \geqslant 0$ if and only if $X \geqslant 2$. Then, using the result in (c)

$$P\{X^3 - X^2 - X + 2 \geqslant 0\} = P\{X \geqslant 2\} = \frac{1}{2}e^{-2}.$$

(f) We have

$$\begin{aligned} e^{\sin(\pi X)} \geqslant 1 &\Leftrightarrow \sin(\pi X) \geqslant 0 \\ &\Leftrightarrow X \in [2k, 2k+1] \text{ for some } k \in \mathbb{Z}.\end{aligned}$$

Now by symmetry, $P\{-2k \leqslant X - 2k + 1\} = P\{2k - 1 \leqslant X \leqslant 2k\}$. Hence, $P\{X \in [2k, 2k+1] \text{ for some } k \in \mathbb{Z}\} = P\{X \geqslant 0\} = \frac{1}{2}$.

(g) As X is a continuous random variable,

$$P\{X \in \mathbb{N}\} = \sum_{n=0}^{\infty} P\{X = n\} = 0,$$

as $P\{X = x\} = 0$ for every x for a continuous random variable.

Exercise 14.4

(a) In order for f to be a pdf, we need $c > 0$. Let's compute $\int_{-\infty}^{+\infty} f(x)dx$:

$$\int_{-\infty}^{+\infty} f(x)dx = c\int_{1}^{+\infty} \frac{1}{\sqrt{x}}dx = c(2\sqrt{x})\Big|_{1}^{+\infty} = +\infty,$$

for all $c > 0$. Hence, there is no value of c for which f is a pdf.

(b) We need to check the properties of a cdf. First, F is non-decreasing. Indeed, if $0 < x \leqslant y$, then $0 \leqslant e^{-\frac{1}{x}} \leqslant e^{-\frac{1}{y}}$.

The function F is right-continuous. For $x \neq 0$, this is obvious. At $x = 0$, we have $\lim_{x\downarrow 0} e^{-\frac{1}{x}} = \lim_{y\uparrow\infty} e^{-y} = 0$.

Finally, $\lim_{x\to-\infty} F(x) = 0$ by definition and

$$\lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} e^{-\frac{1}{x}} = \lim_{z\to 0} e^{-z} = 1.$$

The function F is a cdf. The density function is given by $f(x) = F'(x)$. Hence, $f(x) = 0$ pour $x < 0$. For $x > 0$, we have

$$f(x) = \frac{d}{dx}e^{-\frac{1}{x}} = \frac{1}{x^2}e^{-\frac{1}{x}}.$$

At $x = 0$, we have $F'_{-}(0) = 0$ and

$$F'_{+}(0) = \lim_{h\downarrow 0} \frac{F(0+h) - F(0)}{h} = \lim_{h\downarrow 0} \frac{e^{-\frac{1}{h}}}{h} = \lim_{x\uparrow+\infty} xe^{-x} = 0.$$

Hence, $F'(0) = 0$ and we finally have

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ \frac{1}{x^2}e^{-\frac{1}{x}} & \text{if } x > 0. \end{cases}$$

Exercise 14.5

(a) In order for f to be a pdf, we need $c > 0$. Let's compute $\int_{-\infty}^{+\infty} f(x)dx$:

$$\int_{-\infty}^{+\infty} f(x)dx = c\int_{0}^{+\infty} \frac{1}{1+x^2}dx = c(\arctan(x))\Big|_{0}^{+\infty} = \frac{\pi c}{2}.$$

Then, taking $c = \frac{2}{\pi}$, f is a pdf. This is a Cauchy distribution.

(b) We need to check the properties of a cdf. First check that F is non-decreasing. It is enough to check that $g(x) = \frac{x}{\sqrt{1+x^2}}$ is non-decreasing. For $x \geqslant 0$, we have

$$g(x) = \sqrt{\frac{x^2}{1+x^2}} = \sqrt{1 - \frac{1}{1+x^2}},$$

which is non-decreasing on $[0, +\infty)$. The function g being an odd function, this is also true for $x < y \leqslant 0$. Finally, for $x < 0 < y$, we have $g(x) < 0 < g(y)$. Hence, F is non-decreasing.

It is obvious to see that F is right-continuous.

Finally,

$$\lim_{x \to +\infty} F(x) = \frac{1}{2}\left(1 + \lim_{x \to +\infty} \frac{x}{\sqrt{1+x^2}}\right) = \frac{1}{2}\left(1 + \sqrt{\lim_{x \to +\infty} \frac{x^2}{1+x^2}}\right) = 1.$$

and

$$\lim_{x \to -\infty} F(x) = \frac{1}{2}\left(1 + \lim_{x \to -\infty} \frac{x}{\sqrt{1+x^2}}\right) = \frac{1}{2}\left(1 - \lim_{x \to +\infty} \frac{x}{\sqrt{1+x^2}}\right) = 0.$$

The function F is a cdf. The density function is given by $f(x) = F'(x)$. We have

$$f(x) = \frac{d}{dx}\frac{1}{2}\left(1 + \frac{x}{\sqrt{1+x^2}}\right) = \frac{\sqrt{1+x^2} - x\frac{2x}{2\sqrt{1+x^2}}}{2(1+x^2)} = \frac{1+x^2-x^2}{2(1+x^2)^{\frac{3}{2}}} = \frac{1}{2(1+x^2)^{\frac{3}{2}}},$$

for all $x \in \mathbb{R}$.

Exercise 18.1 Let $g : (-1, 1) \to (\frac{1}{e}, e)$ defined by $g(x) = e^{-x}$. Then, g is one-to-one. Its inverse is the solution of $y = e^{-x}$ which is $x = -\log y$. As

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in (-1, 1), \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$f_Y(y) = f_X(x)|\frac{dx}{dy}| = f_X(-\log y)\frac{1}{y} = \begin{cases} \frac{1}{2y} & \text{if} \quad y \in (\frac{1}{e}, e), \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 18.2 Let $g : (0, \infty) \to (0, \infty)$ defined by $g(x) = x^2$. Then, g is one-to-one. Its inverse is the solution of $y = x^2$ which is $x = \sqrt{y}$ (not $-\sqrt{y}$ since $x > 0$). As

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

we have

$$f_Y(y) = f_X(x)|\frac{dx}{dy}| = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} = \begin{cases} \frac{\lambda e^{-\lambda\sqrt{y}}}{2\sqrt{y}} & \text{if} \quad y > 0, \\ 0 & \text{if} \quad y \leqslant 0. \end{cases}$$

Exercise 18.3

(a) We have $Y = g(X)$, with $g : \mathbb{R} \to (0, \infty)$ given by $g(x) = e^x$. Then, g is one-to-one. Its inverse is the solution to $y = e^x$ which is $x = \log y$. As X is a standard normal random variable, $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Then, for $y > 0$,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(\log y)^2}{2}}|(\log y)'| = \frac{1}{\sqrt{2\pi}\,y}e^{-\frac{(\log y)^2}{2}},$$

and $f_Y(y) = 0$ if $y \leqslant 0$.

(b) We have $h(Z) = X$, with $h : \mathbb{R} \to \mathbb{R}$ given by $h(z) = z^3 + z + 1$. This function is one-to-one. Indeed, $h'(z) = 3z^2 + 1 > 0$ for all $z \in \mathbb{R}$, hence h is strictly increasing. Moreover, $\lim_{z\to+\infty} h(z) = +\infty$ and $\lim_{z\to-\infty} h(z) = -\infty$, which ensures the bijectivity of h. The r.v. Z is then given by $Z = h^{-1}(X)$ which (conveniently!) has solution $X = h(Z)$.

The probability density function f_Z is then $f_Z(z) = f_X(h(z))|h'(z)|$. As X is a standard normal random variable,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}(3z^2 + 1)e^{-\frac{(z^3+z+1)^2}{2}},$$

for all $z \in \mathbb{R}$.

Exercise 18.4

(a) We have $Y = g(X)$, with $g : (0, \infty) \to \mathbb{R}$ given by $g(x) = \log x$. The function g is one-to-one. Its inverse is the solution to $y = \log x$ which is $x = e^y$. Hence :

$$f_Y(y) = f_X(e^y)|(e^y)'|.$$

As X is an exponential r.v. with parameter λ, $f_X(x) = \lambda e^{-\lambda x}$, for $x \geqslant 0$ and $f_X(x) = 0$ otherwise. Hence, for $y \in \mathbb{R}$,

$$f_Y(y) = \lambda e^{-\lambda e^y}|e^y| = \lambda e^{y - \lambda e^y}.$$

(b) We have $h(Z) = X$, with $h : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}$ given by $h(z) = z + \tan(z)$. This function is one-to-one. Indeed, $h'(z) = 1 + \frac{1}{\cos^2(z)} > 0$ for all $z \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, hence g is strictly increasing. Moreover, $\lim_{z \to \frac{\pi}{2}} h(z) = +\infty$ and $\lim_{z \to -\frac{\pi}{2}} h(z) = -\infty$, which ensures the bijectivity of h. The r.v. Z is then given by $Z = h^{-1}(X)$, the inverse of which is (conveniently) given by $X = h(Z)$.

The probability density function f_Z is then $f_Z(z) = f_X(h(z))|h'(z)|$. As X is a standard normal random variable,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \left(1 + \frac{1}{\cos^2(z)}\right) e^{-\frac{(z + \tan(z))^2}{2}},$$

for all $z \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

Exercise 18.5 Let $g : [1, \infty) \to [1, \infty)$ defined by

$$g(x) = \begin{cases} 2x & \text{if } x \geqslant 2, \\ x^2 & \text{if } x < 2, \end{cases}$$

Then, as it is increasing, the function g is one-to-one. As

$$f_X(x) = \begin{cases} \frac{1}{x^2} & \text{if } x \geqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$f_Y(y) = f_X(h(y))|h'(y)|.$$

We have to consider two cases. First, for $1 \leqslant y < 4$, we have

$$h(y) = g^{-1}(y) = \sqrt{y}, \qquad h'(x) = \frac{1}{2\sqrt{y}}.$$

Then, for $y > 4$, we have

$$h(y) = g^{-1}(y) = \frac{y}{2}, \qquad h'(x) = \frac{1}{2}.$$

Hence,

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

Exercise 18.6

(a) We have $Y = g(X)$, with $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = x^2$. The function g is not one-to-one from $\mathbb{R}$ into $\mathbb{R}$. First find the cumulative distribution function F_Y from the definition. For $y \leqslant 0$, $F_Y(y) = P\{Y \leqslant y\} = P\{X^2 \leqslant y\} = 0$. For $y > 0$,

$$F_Y(y) = P\{Y \leqslant y\} = P\{X^2 \leqslant y\} = P\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\} = F(\sqrt{y}) - F(-\sqrt{y}),$$

where F is the CDF of X. We then find the probability density function f_Y by taking the derivative. For $y < 0$, $f_Y(y) = 0$. For $y > 0$,

$$f_Y(y) = F_Y'(y) = \frac{1}{2\sqrt{y}} f(\sqrt{y}) + \frac{1}{2\sqrt{y}} f(-\sqrt{y}) = \frac{1}{2\sqrt{y}} (f(\sqrt{y}) + f(-\sqrt{y})).$$

Finally,

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} (f(\sqrt{y}) + f(-\sqrt{y})) & \text{if } y > 0, \\ 0 & \text{if } y < 0. \end{cases}$$

Alternatively, we could have applied the formula for f_Y once to each solution of $x^2 = y$ and then added the two. So

$$f_Y(y) = f(\sqrt{y})|(\sqrt{y})'| + f(-\sqrt{y})|(-\sqrt{y})'|$$

which gives the same answer as above.

(b) We will apply the formula obtained in (a) when X is a standard normal r.v., namely with $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. In this case, for $y > 0$,

$$f_Y(y) = \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} (e^{-\frac{y}{2}} + e^{-\frac{y}{2}}) = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}},$$

which corresponds to a Gamma density function with parameters $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$. Indeed, $y^{-\frac{1}{2}} e^{-\frac{y}{2}}$ appears in the Gamma density and the constant is necessarily the right one, determined by the property $\int_0^{+\infty} f_Y(y) dy = 1$. In particular, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Exercise 18.7 Let $g : [0, \frac{\pi}{2}] \to [0, \frac{v_0^2}{g}]$ be defined by $g(x) = \frac{v_0^2}{g} \sin(2\theta)$. The function g is not one-to-one as every possible sine value can be obtained in two ways, namely as $\frac{1}{2}\arcsin(\frac{gR}{v_0^2})$ and $\frac{\pi}{2} - \frac{1}{2}\arcsin(\frac{gR}{v_0^2})$. Hence, for $0 \leqslant r \leqslant \frac{v_0^2}{g}$,

$$\begin{aligned}
F_R(r) &= P(R \leqslant r) = P\left(\left\{0 \leqslant \theta \leqslant \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right\} \text{ or } \left\{\frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right) \leqslant \theta \leqslant \frac{\pi}{2}\right\}\right) \\
&= \frac{2}{\pi}\left(\frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right) + \frac{2}{\pi}\left(\frac{\pi}{2} - \left(\frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right)\right) \\
&= \frac{2}{\pi}\arcsin\left(\frac{gr}{v_0^2}\right).
\end{aligned}$$

Now, we find that for $0 \leqslant r \leqslant \frac{v_0^2}{g}$,

$$\begin{aligned}
f_R(r) &= F_R'(r) = \frac{2}{\pi}\frac{g}{v_0^2}\arcsin'\left(\frac{gr}{v_0^2}\right) \\
&= \frac{2g}{\pi v_0^2}\frac{1}{\sqrt{1 - \frac{g^2r^2}{v_0^4}}} \\
&= \frac{2g}{\pi}\frac{1}{\sqrt{v_0^4 - g^2r^2}}
\end{aligned}$$

Finally,

$$f_R(r) = \begin{cases} \frac{2g}{\pi}\frac{1}{\sqrt{v_0^4 - g^2r^2}} & \text{if } 0 \leqslant r \leqslant \frac{v_0^2}{g}, \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 21.1 We have

$$P\{X=+1\}=\frac{18}{38}, \qquad P\{X=-1\}=\frac{20}{38}.$$

Hence,

$$E[X]=(+1)\times\frac{18}{38}+(-1)\times\frac{20}{38}=-\frac{2}{38}\simeq -0.0526.$$

This means that *on average* you lose 5.26 cents per bet.

Exercise 21.2 Let the sample space Ω be $\{1,2,5,10,20,\text{“0”},\text{“00”}\}$. Denote by ω the outcome of the wheel. The probability measure P is given by

ω	1	2	5	10	20	0	00
$P\{\omega\}$	$\frac{22}{52}$	$\frac{15}{52}$	$\frac{7}{52}$	$\frac{4}{52}$	$\frac{2}{52}$	$\frac{1}{52}$	$\frac{1}{52}$

(a) Let H be the random variable given the profit of the player when he bets \$1 on each of the possible numbers or symbols. The possible values for H are

ω	1	2	5	10	20	0	00
$H(\omega)$	-5	-4	-1	4	14	34	34

The probability mass function of H is

x	-5	-4	-1	4	14	34
$P\{H=x\}$	$\frac{22}{52}$	$\frac{15}{52}$	$\frac{7}{52}$	$\frac{4}{52}$	$\frac{2}{52}$	$\frac{2}{52}$

Hence, the expectation is

$$E[H]=(-5)\cdot\frac{22}{52}+(-4)\cdot\frac{15}{52}+(-1)\cdot\frac{7}{52}+4\cdot\frac{4}{52}+14\cdot\frac{2}{52}+34\cdot\frac{2}{52}=-\frac{65}{52}=-1.25.$$

(b) For $m\in\{1,2,5,10,20,\text{“0”},\text{“00”}\}$, let H_m be the profit of the player when he bets \$1 on the number or symbol m. Then, H_m can only take two values and its mass function is

x	-1	m
$P\{H_m=x\}$	$1-p_m$	p_m

, if $m\in\{1,2,5,10,20\}$,

x	-1	40
$P\{H_m=x\}$	$\frac{51}{52}$	$\frac{1}{52}$

, if $m\in\{0,00\}$,

where $p_m=P\{\omega=m\}$. Hence, $E[H_m]=mp_m+(-1)(1-p_m)$. The numerical results are presented in the following table:

m	1	2	5	10	20	0	00
$E[H_m]$	$-\frac{8}{52}$	$-\frac{7}{52}$	$-\frac{10}{52}$	$-\frac{8}{52}$	$-\frac{10}{52}$	$-\frac{11}{52}$	$-\frac{11}{52}$

Hence, betting on "0" or "00" gives the worst expectation and bet on "2" gives the best. We notice that the expected values are all negative and, hence, this game is always in favor of the organiser.

Exercise 21.3 We know that for a Geometric random variable, $f(k) = P\{X = k\} = p(1-p)^{k-1}$ for $k \geqslant 1$. Hence, we have

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p \sum_{k=1}^{\infty} kq^{k-1},$$

with $q := 1 - p$. The trick to compute this sum is to remark that kq^{k-1} is the derivative with repsect to q of q^k. Hence, we can write

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

Exercise 23.1 For an exponential random variable, we have $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, $f(x) = 0$ otherwise. Hence, by an integration by parts, we have

$$\begin{aligned} E[X] &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\ &= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\ &= \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Then, again using integration by parts and the results above, we have

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \\ &= \left[-x^2 e^{-\lambda x}\right]_0^\infty + 2\int_0^\infty xe^{-\lambda x}\,dx \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

Exercise 23.2 First of all, if n is odd, we have

$$E[X^n] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x^n e^{-\frac{x^2}{2}}\,dx = 0,$$

by the symmetry of the function $x \mapsto x^n e^{-\frac{x^2}{2}}$ (the function is odd). Moreover, if $n = 2$, we have seen in class that $E[X^2] = 1$. Let's prove the result by induction. Assume the result is true for all even numbers up to $n-2$ and let's compute $E[X^n]$. Using an integration by parts, we have

$$\begin{aligned} E[X^n] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x^n e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty x^{n-1} x e^{-\frac{x^2}{2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\left(-x^{n-1}e^{-\frac{x^2}{2}}\right)\Bigg|_{-\infty}^\infty + \frac{(n-1)}{\sqrt{2\pi}}\int_{-\infty}^\infty x^{n-2}e^{-\frac{x^2}{2}}\,dx \\ &= (n-1)E[X^{n-2}] = (n-1)\cdot(n-3)(n-5)\cdots 1. \end{aligned}$$

Hence, by induction the result is true for every even integer n.

Exercise 23.3 By the definition of expectation and using integration by parts, we have

$$\begin{aligned}
E[c(X)] &= \int_{-\infty}^{\infty} c(x)f(x)\,dx = 2\int_0^3 xe^{-x}\,dx + \int_3^{\infty}(2+6(x-3))xe^{-x}\,dx \\
&= 2\left(-xe^{-x}\right)\Big|_0^3 + 2\int_0^3 e^{-x}\,dx + \left(-(2+6(x-3))xe^{-x}\right)\Big|_3^{\infty} + \int_3^{\infty}(12x-16)e^{-x}\,dx \\
&= -6e^{-3} + 2\left(-e^{-x}\right)\Big|_0^3 + 6e^{-3} + \left(-(12x-16)e^{-x}\right)\Big|_3^{\infty} + \int_3^{\infty} 12e^{-x}\,dx \\
&= 2 - 2e^{-3} + 20e^{-3} + \left(-12e^{-x}\right)\Big|_3^{\infty} \\
&= 2 + 18e^{-3} + 12e^{-3} = 2 + 30e^{-3}.
\end{aligned}$$

Exercise 24.1 We remind that if X is exponentially distributed with parameter 1, then $E[X] = 1$.

(a) We have $E[XY] = E[X]E[Y] = 1 \cdot 1 = 1$, since X and Y are independent.
(b) We have $E[X - Y] = E[X] - E[Y] = 1 - 1 = 0$.
(c) This is Example **??** on page **??** in the Lecture Notes.

Exercise 24.2 This corresponds to Examples 23.5 on page 115 and 24.6 on page 120 in the Lecture Notes.

Exercise 24.3 This corresponds to Example 24.9 on page 121 in the Lecture Notes.

Exercise 24.4 First of all, let's notice that $Y^2 + Z^2 = \cos^2(X) + \sin^2(X) = 1$ and that $YZ = \cos(X)\sin(X) = \frac{\sin(2X)}{2}$. Hence, we have

$$E[YZ] = \frac{1}{2}E[\sin(2X)] = \frac{1}{4\pi}\int_0^{2\pi} \sin(2x)\,dx = 0$$

Moreover,

$$E[Y] = E[\cos(X)] = \frac{1}{2\pi}\int_0^{2\pi} \cos(x)\,dx = 0.$$

Similarly, $E[Z] = 0$ and $E[YZ] = E[Y]E[Z]$. Then, as $E[Y] = 0$,

$$\mathrm{Var}(Y) = E[Y^2] = E[\cos^2(X)] = \frac{1}{2\pi}\int_0^{2\pi} \cos^2(x)\,dx = \frac{1}{4\pi}(x - \frac{\sin(2x)}{2})\Bigg|_0^{2\pi} = \frac{1}{2}.$$

Simlarly, we can show that $\mathrm{Var}(Z) = \frac{1}{2}$. Moreover, as $E[Y + Z] = 0$,

$$\mathrm{Var}(Y + Z) = E[(Y + Z)^2] = E[Y^2 + Z^2 + 2YZ] = 1 + 2E[YZ] = 1.$$

Hence, $\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + \mathrm{Var}(Z)$. Nevertheless, we have,

$$P(Y > 1/2) = P(\cos(X) > 1/2) = P(-\pi/3 < X < \pi/3) = \frac{1}{3},$$

$$P(Z > 1/2) = P(\sin(X) > 1/2) = P(\pi/6 < X < 5\pi/6) = \frac{1}{3},$$

and

$$P(Y > 1/2, Z > 1/2) = P(\pi/6 < X < \pi/3) = \frac{1}{12} \neq \frac{1}{9},$$

which proves that Y and Z are not independent.

Exercise 24.5 See Ash's exercise 3.2.8.

Exercise 24.6 This corresponds to Theorem 24.1 on page 119 in the Lecture Notes.

Exercise 25.1

(a) No, X and Y are *not* independent. For instance, we have $f_X(1) = 0.4 + 0.3 = 0.7$, $f_Y(2) = 0.3 + 0.1 = 0.4$. Hence, $f_X(1)f_Y(2) = 0.7 \cdot 0.4 = 0.28 \neq 0.3 = f(1,2)$.

(b) We have

$$P(XY \leqslant 2) = 1 - P(XY > 2) = 1 - P(X = 2, Y = 2) = 1 - 0.1 = 0.9.$$

Exercise 25.2

(a) The set of possible values for X_1 and X_2 is $\{1, \ldots, 6\}$. By definition, we always have $X_1 \leqslant X_2$. We have to ccompute $f(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\}$. If $x_1 = x_2$, both outcomes have to be the same, equal to x_1. There is only one possible roll for this, namely (x_1, x_1) and $f(x_1, x_2) = \frac{1}{36}$. If $x_1 < x_2$, one dice has to be x_1, the other one x_2. There are two possible rolls for this to happen, namely (x_1, x_2) and (x_2, x_1). We obtain $f(x_1, x_2) = \frac{1}{18}$. Then, for $x_1, x_2 \in \{1, 2, 3, 4, 5, 6\}$,

$$f(x_1, x_2) = \begin{cases} \frac{1}{36} & \text{if } x_1 = x_2, \\ \frac{1}{18} & \text{if } x_1 < x_2, \\ 0 & \text{otherwise.} \end{cases}$$

(b) In order to find the density of X_1, we have to add all the probabilities for which X_1 takes a precise value (i.e. $f_{X_1}(x_1) = \sum_{i=1}^{6} f(x_1, i)$). The following table sums up the results (as in the example in class).

$x_1 \mid x_2$	1	2	3	4	5	6	$f_{X_1}(x_1)$
1	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{11}{36}$
2	0	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{9}{36}$
3	0	0	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{7}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{5}{36}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{3}{36}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$f_{X_2}(x_2)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	

(c) They are not independent. Namely, $f(x_1, x_2) \neq f_{X_1}(x_1)f_{X_2}(x_2)$. For instance, $f(6, 1) = 0 \neq \frac{1}{36^2} = f_{X_1}(6)f_{X_2}(1)$.

Exercise 25.3

(a) The set of possible values for X_1 is $\{4, 5, 6, 7, 8\}$ and the set of possible values for X_2 is $\{4, 6, 8, 9, 12, 16\}$. We can see that the values of X_1 and X_2 only correspond to one exact possible draw (up to the symmetry). Hence, possible values $(4,4),(6,9)$ and $(8,16)$ for (X_1,X_2) respectivley correspond to the draws of $(2,2),(3,3)$ and $(4,4)$. Their probability is $\frac{1}{9}$. Possible values $(5,6),(6,8)$ and $(7,12)$ for (X_1,X_2) respectivley correspond to the draws of $(2,3),(2,4)$ and $(3,4)$ (and their symmetric draws). Their probability is $\frac{2}{9}$. Other pairs are not possible and have probability 0.

(b) In order to find the density of X_1, we have to add all the probabilities for which X_1 takes a precise value (i.e. $f_{X_1}(x_1) = \sum_{i=1}^{6} f(x_1,i)$). The following table sums up the results (as in the example in class).

$x_1 \vert x_2$	4	6	8	9	12	16	$f_{X_1}(x_1)$
4	$\frac{1}{9}$	0	0	0	0	0	$\frac{1}{9}$
5	0	$\frac{2}{9}$	0	0	0	0	$\frac{2}{9}$
6	0	0	$\frac{2}{9}$	$\frac{1}{9}$	0	0	$\frac{3}{9}$
7	0	0	0	0	$\frac{2}{9}$	0	$\frac{2}{9}$
8	0	0	0	0	0	$\frac{1}{9}$	$\frac{1}{9}$
$f_{X_2}(x_2)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	

(c) They are not independent. Namely, $f(x_1,x_2) \neq f_{X_1}(x_1)f_{X_2}(x_2)$. For instance, $f(5,4) = 0 \neq \frac{2}{81} = f_{X_1}(5)f_{X_2}(4)$.

Exercise 26.1 We can see that the distribution of (X, Y) is uniform on the square $[-1, 1]^2$. Hence, we can use a ratio of surfaces to compute the probabilities. (In most cases a drawing of the domain can help.)

(a) We have $P\{X+Y \leqslant \frac{1}{2}\} = P\{Y \leqslant \frac{1}{2} - X\} = 1 - P\{Y > \frac{1}{2} - X\}$. Now the surface corresponding to $\{Y \geqslant \frac{1}{2} - X\}$ is a triangle and we have

$$P\{X+Y \leqslant \frac{1}{2}\} = 1 - \frac{\frac{1}{2}\cdot\left(\frac{3}{2}\right)^2}{4} = \frac{23}{32}.$$

(b) The domain corresponding to $\{X - Y \leqslant \frac{1}{2}\}$ has exactly the same shape as the one in (a). Hence, $P\{X - Y \leqslant \frac{1}{2}\} = \frac{23}{32}$.

(c) We have $XY > \frac{1}{4} \Leftrightarrow Y > \frac{1}{4X}$ if $X \geqslant 0$ and $XY > \frac{1}{4} \Leftrightarrow Y < \frac{1}{4X}$ if $X < 0$. Now, we can write the surface of the domain corresponding to $XY > \frac{1}{4}$ as

$$2\int_{\frac{1}{4}}^{1} dx \int_{1/4x}^{1} dy = 2\int_{\frac{1}{4}}^{1} dx \left(1 - \frac{1}{4x}\right) = 2\left(x - \frac{\ln(x)}{4}\right)\Bigg|_{\frac{1}{4}}^{1} = \frac{3-\ln(4)}{2}.$$

Hence, $P\{XY \leqslant \frac{1}{4}\} = 1 - P\{XY > \frac{1}{4}\} = 1 - \frac{3-\ln(4)}{8} = \frac{5+\ln(4)}{8}$.

(d) We have $\frac{Y}{X} \leqslant \frac{1}{2} \Leftrightarrow Y \leqslant \frac{X}{2}$ if $X \geqslant 0$ and $\frac{Y}{X} \leqslant \frac{1}{2} \Leftrightarrow Y \geqslant \frac{X}{2}$ if $X < 0$. Hence, the surface corresponding to $\{\frac{Y}{X} \leqslant \frac{1}{2}\}$ is the union of two trapezoids with surface $\frac{5}{4}$ each. Hence, $P\{\frac{Y}{X} \leqslant \frac{1}{2}\} = 2 \cdot \frac{5/4}{4} = \frac{5}{8}$.

(e) We have $P\left\{\left|\frac{Y}{X}\right| \leqslant \frac{1}{2}\right\} = P\left\{\frac{Y^2}{X^2} \leqslant \frac{1}{4}\right\} = P\left\{Y^2 \leqslant \frac{X^2}{4}\right\} = P\left\{-\frac{|X|}{2} \leqslant Y \leqslant \frac{|X|}{2}\right\}$. We can easilly identify the surface as the union of two triangles of surface $\frac{1}{2}$ each and, hence,

$$P\left\{\left|\frac{Y}{X}\right| \leqslant \frac{1}{2}\right\} = 2 \cdot \frac{1/2}{4} = \frac{1}{4}.$$

(f) We have $P\{|X| + |Y| \leqslant 1\} = P\{|Y| \leqslant 1 - |X|\} = P\{|X| - 1 \leqslant Y \leqslant 1 - |X|\}$. The surface is then a square with corners $(0, 1), (-1, 0), (0, -1)$ and $(1, 0)$. The sides have length $\sqrt{2}$ and

$$P\{|X| + |Y| \leqslant 1\} = \frac{(\sqrt{2})^2}{4} = \frac{1}{2}.$$

(g) We have $P\{|Y| \leqslant e^X\} = P\{-e^X \leqslant Y \leqslant e^X\}$. This condition only matters when $X < 0$. Hence,

$$P\{|Y| \leqslant e^X\} = \frac{1}{2} + \int_{-1}^{0} dx \int_{-e^x}^{e^x} dy\, \frac{1}{4} = \frac{1}{2} + \frac{1}{2}\int_{-1}^{0} dx\, e^x = \frac{1}{2} + \frac{1}{2}(1 - e^{-1}) = 1 - \frac{1}{2e}.$$

Exercise 27.1

(a) We must choose c such that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1.$$

But,

$$\begin{aligned}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dxdy &= c\int_0^1\int_0^1 (x+y)\,dxdy = c\int_0^1\left[\frac{x^2}{2}+xy\right]_{x=0}^{x=1} dy\\ &= c\int_0^1\left(\frac{1}{2}+y\right)\,dy = c\left[\frac{y}{2}+\frac{y^2}{2}\right]_0^1 = c\left(\frac{1}{2}+\frac{1}{2}\right) = c.\end{aligned}$$

Hence, $c = 1$.

(b) Observe that

$$\begin{aligned}P\{X<Y\} &= \iint_{\{(x,y)\,:\,x<y\}} f(x,y)\,dxdy = \int_0^1\int_0^y (x+y)\,dxdy\\ &= \int_0^1\left[\frac{x^2}{2}+xy\right]_{x=0}^{x=y} dy = \frac{3}{2}\int_0^1 y^2\,dy = \frac{3}{2}\left[\frac{y^3}{3}\right]_0^1 = \frac{1}{2}.\end{aligned}$$

(c) For $x\notin[0,1]$, $f_X(x)=0$. For $x\in[0,1]$,

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy = \int_0^1 (x+y)\,dy = \left[xy+\frac{y^2}{2}\right]_0^1 = \frac{1}{2}+x.$$

By symmetry, $f_Y(y) = f_X(y)$ for all $y\in\mathbb{R}$.

(d) We can write $P\{X=Y\}$ as

$$P\{X=Y\} = \iint_{\{(x,y)\,:\,x=y\}} f(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_y^y f(x,y)\,dxdy = 0,$$

for all density function f. Hence, $P\{X=Y\}=0$ for all jointly continuous random variables.

Exercise 27.2

(a) First of all, observe that $f_X(x)=0$ if $x\notin[0,1]$. Then, for $0\leqslant x\leqslant 1$,

$$f_X(x) = \int_{\mathbb{R}} f(x,y)dy = \int_0^x 4xy\,dy + \int_x^1 6x^2\,dy = 2x^3+6x^2(1-x) = 6x^2-4x^3.$$

Moreover, $f_Y(y)=0$ for $y\notin[0,1]$. Then, for $0\leqslant y\leqslant 1$,

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)dx = \int_0^y 6x^2\,dx + \int_y^1 4xy\,dx = 2y^3+2y(1-y^2) = 2y.$$

(b) We have

$$\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
&= P\{X \leqslant 1/2\} + P\{Y \leqslant 1/2\} - P\{X \leqslant 1/2, Y \leqslant 1/2\} \\
&= \int_0^{1/2} (6x^2 - 4x^3)\,dx + \int_0^{1/2} 2y\,dy - \int_0^{1/2} dx \left(\int_0^x 4xy\,dy + \int_x^{1/2} 6x^2\,dy \right) \\
&= (2x^3 - x^4)\Big|_0^{1/2} + y^2\Big|_0^{1/2} - \int_0^{1/2} dx(3x^2 - 4x^3) \\
&= \left(\frac{1}{4} - \frac{1}{16}\right) + \frac{1}{4} - (x^3 - x^4)\Big|_0^{1/2} \\
&= \frac{7}{16} - \left(\frac{1}{8} - \frac{1}{16}\right) = \frac{6}{16} = \frac{3}{8}.
\end{aligned}$$

Exercise 27.3 First of all, $f_X(x) = 0$ if $x < 0$. Now, for $x \geqslant 0$,

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = 2\int_0^x e^{-(x+y)}\,dy = 2e^{-x}(-e^{-y})\Big|_0^x = 2e^{-x}(1 - e^{-x}).$$

We have $f_Y(y) = 0$ for $y < 0$. For $y \geqslant 0$,

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx = 2\int_y^\infty e^{-(x+y)}\,dx = 2e^{-y}(-e^{-x})\Big|_y^\infty = 2e^{-2y}.$$

Exercise 27.4 See Ash's exercise 2.7.3.

Exercise 27.5 See Ash's exercise 2.7.8.

Exercise 28.1 There are two ways to proceed.

One way is to first compute the CDF. There are two cases to distinguish. If $0 < z \leqslant 1$, then the parabola $y = zx^2$ intersects the square at the right-most side. Then in that case,

$$P\{Y/X^2 \leqslant z\} = P\{Y \leqslant zX^2\} = \int_0^1 zx^2\,dx = z/3.$$

But when $z > 3$ then the parabola intersects the square at its top side at $x = 1/\sqrt{z}$ and

$$P\{Y/X^2 \leqslant z\} = P\{Y \leqslant zX^2\} = \int_0^{1/\sqrt{z}} zx^2\,dx + \int_{1/\sqrt{z}}^1 1\,dx = \frac{z}{3(\sqrt{z})^3} + 1 - \frac{1}{\sqrt{z}} = 1 - \frac{2}{3\sqrt{z}}.$$

Therefore, the pdf is $1/3$ when $z \in (0,1]$ and $1/(3z^{3/2})$ when $z > 1$.

Alternatively, one can use the pdf method: let $W = X$ and $Z = Y/X^2$. Then $X = W$ and $Y = ZW^2$. The Jacobian matrix is

$$\begin{bmatrix} 1 & 0 \\ 2zw & w^2 \end{bmatrix}$$

and its determinant is w^2. So

$$f_{Z,W}(z,w) = 1 \times w^2 = w^2.$$

The crucial thing though is the domain! The above formula is valid if $0 < x < 1$ and $0 < y < 1$ which becomes $0 < w < 1$ and $0 < zw^2 < 1$. The pdf is 0 otherwise. So if $0 < z < 1$ then $0 < w < 1$ and while if $z > 1$ we gave $0 < w < 1/sqrt(z)$. Finally, the pdf of Z is

$$f_Z(z) = \int_0^1 w^2\,dw = \frac{1}{3} \quad \text{if } 0 < z < 1$$

and

$$f_Z(z) = \int_0^{1/\sqrt{z}} w^2\,dw = \frac{1}{3z^{3/2}} \quad \text{if } z > 1.$$

Exercise 28.2 First of all, by independence,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \begin{cases} e^{-(x+y)} & \text{if } x \geqslant 0, y \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

(a) We will use the transformation $U = X$, $Z = X + Y$. This transformation is bijective with inverse given by $X = U$, $Y = Z - U$. The Jacobian of this transformation is given by

$$J(u,z) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial z} \end{pmatrix} = \det\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} = 1.$$

Now,

$$f_{U,Z}(u,z) = f_{X,Y}(x(u,z), y(u,z))|J(u,z)| = e^{-(u+(z-u))} = e^{-z},$$

for $u \geqslant 0$, $z \geqslant 0$ and $u \leqslant z$. The latter condition comes from $y \geqslant 0$. Hence,

$$f_{U,Z}(u,z) = \begin{cases} e^{-z} & \text{if } u \geqslant 0, z \geqslant 0 \text{ and } u \leqslant z, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, $f_Z(z) = 0$ if $z < 0$ and, for $z \geqslant 0$,

$$f_Z(z) = \int_{\mathbb{R}} f_{U,Z}(u,z)\,du = \int_0^z e^{-z}\,du = ze^{-z}.$$

(b) Similarly as above, we will consider $V = X$, $W = \frac{Y}{X}$. This transformation is bijective with inverse $X = V$, $Y = VW$. The Jacobian of this transformation is given by

$$J(v,w) = \det\begin{pmatrix} \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \end{pmatrix} = \det\begin{pmatrix} 1 & 0 \\ w & v \end{pmatrix} = v.$$

Now,

$$f_{V,W}(v,w) = f_{X,Y}(x(v,w), y(v,w))|J(v,w)| = e^{-(v+vw)} \cdot v = ve^{-(1+w)v},$$

for $v \geqslant 0$, $w \geqslant 0$. Hence,

$$f_{V,W}(v,w) = \begin{cases} ve^{-(1+w)v} & \text{if } v \geqslant 0, w \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, $f_W(w) = 0$ if $w < 0$ and, for $w \geqslant 0$,

$$f_W(w) = \int_{\mathbb{R}} f_{V,W}(v,w)\,dv = \int_0^\infty ve^{-(1+w)v}\,dv = \frac{1}{(1+w)^2},$$

where we used the properties of Gamma integrals on p.73 of the Lecture Notes.

Exercise 28.3 First of all, by independence,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}.$$

We will consider $U = X$, $Z = \frac{Y}{X}$. This transformation is bijective with inverse $X = U$, $Y = UZ$. The Jacobian of this transformation is given by

$$J(u,z) = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial z} \end{pmatrix} = \det\begin{pmatrix} 1 & 0 \\ z & u \end{pmatrix} = u.$$

Now,

$$f_{U,Z}(u,z) = f_{X,Y}(x(u,z), y(u,z))|J(u,z)| = \frac{1}{2\pi}e^{-\frac{1}{2}(u^2+u^2z^2)} \cdot |u| = \frac{1}{2\pi}|u|e^{-\frac{1}{2}(1+z^2)u^2},$$

for all $u, z \in \mathbb{R}$. Finally, for $z \in \mathbb{R}$,

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_{U,Z}(u,z)\,du = \frac{1}{2\pi}\int_{-\infty}^{\infty} |u| e^{-\frac{1}{2}(1+z^2)u^2}\,du = \frac{1}{\pi}\int_0^{\infty} u e^{-\frac{1}{2}(1+z^2)u^2}\,du \\ &= -\frac{1}{\pi(1+z^2)} e^{-\frac{1}{2}(1+z^2)u^2}\Big|_0^{\infty} = \frac{1}{\pi(1+z^2)}. \end{aligned}$$

Exercise 28.4 See Ash's exercise 2.8.5.

Exercise 28.5 See Ash's exercise 2.8.6.

Exercise 28.6 See Ash's exercise 2.8.8.

Exercise 28.7 First of all, by independence,

$$f(x,y,z) = f_X(x) f_Y(y) f_Z(z) = \begin{cases} e^{-(x+y+z)} & \text{if } x \geqslant 0, y \geqslant 0, z \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

Now, letting $A = \{(x,y,z) \in \mathbb{R}^3 : x \geqslant 2y \geqslant 3z\}$, we have

$$\begin{aligned} P\{X \geqslant 2Y \geqslant 3Z\} &= \int\!\!\int\!\!\int_A f(x,y,z)\,dx\,dy\,dz \\ &= \int_0^{\infty} dz \int_{\frac{3}{2}z}^{\infty} dy \int_{2y}^{\infty} dx\, e^{-(x+y+z)} \\ &= \int_0^{\infty} dz\, e^{-z} \int_{\frac{3}{2}z}^{\infty} dy\, e^{-y} \int_{2y}^{\infty} dx\, e^{-x} \\ &= \int_0^{\infty} dz\, e^{-z} \int_{\frac{3}{2}z}^{\infty} dy\, e^{-3y} \\ &= \frac{1}{3}\int_0^{\infty} dz\, e^{-\frac{11}{2}z} = \frac{2}{33}. \end{aligned}$$

Exercise 28.8 Let X (resp. Y) be the number of minutes after 10 am at which the woman (resp. man) arrives. The random variables X and Y are independent and both uniformly distributed on the interval $[0,60]$. Then, the vector (X,Y) is uniformly distributed on the square $[0,60]^2$. We want to find the probability p that both people arrive within an interval of 10 minutes one from the other. In other words, we want $p = P(\{X \leqslant Y \leqslant X+10\} \cup \{Y < X \leqslant Y+10\})$. Both events are disjoint and, by symmetry, have the same probability. Hence, $p = 2P\{X \leqslant Y \leqslant X+10\} = 2P\{(X,Y) \in A\}$, with $A = \{(x,y) \in [0,60]^2 : x \leqslant y \leqslant x+10\}$. We know that

$$p = 2 \cdot \frac{\text{Area}(A)}{\text{Aire}([0,60]^2)} = \frac{\text{Area}(A)}{1800}.$$

Let's compute the area of A (a picture can help). The set A is a trpezoid made of the triangle $T_1 = \{(x,y) \in [0,60]^2 : x \leqslant y\}$ minus the triangle $T_2 = \{(x,y) \in [0,60]^2 : y \leqslant x+10\}$. We have $\text{Area}(A) = \text{Area}(T_1) - \text{Area}(T_2)$.

Triangle T_1 (resp. T_2) has size of length 60 (resp. 50). Then, we find $\text{Area}(A) = \frac{60^2}{2} - \frac{50^2}{2} = \frac{1100}{2} = 550$. Finally, $p = \frac{550}{1800} = \frac{11}{36} = 0.3056$.

Exercise 28.9

(a) Let X (resp. Y) the number of minutes John will have to wait for the bus (resp. the train). As John doesn't know the exact schedules, we assume that the random variables X and Y are independent and unifomrly distributed on the interval $[0,20]$ (resp. $[0,10]$). Hence, the random vector (X,Y) is uniformly distributed on the rectangle $[0,20] \times [0,10]$. We want to find the probability p that the total travel time with public transportation is larger than 27. In other words, $p = P\{X+Y+12 > 27\} = P\{X+Y > 15\}$. Hence, $p = P\{(X,Y) \in A\}$, with $A = \{(x,y) \in [0,20] \times [0,10] : x+y > 15\}$. We know that

$$p = \frac{\text{Area}(A)}{\text{Area}([0,20] \times [0,10])} = \frac{\text{Area}(A)}{200}.$$

Let's compute $\text{Area}(A)$. The set A is a trapezoid with small base 5, large base 15 and height 10. Hence, $\text{Area}(A) = \frac{(5+15)\cdot 10}{2} = 100$. Finally, $p = \frac{100}{200} = \frac{1}{2}$.

(b) If we know that the buses are systematically 2 minutes late, it doesn't change anything to the problem above. As John doesn't know the exact schedule, the uniform assumption remains unchanged.

Exercise 28.10 First of all, by independence,

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x^2+y^2)}.$$

We will consider the transformation $X = R\cos(\Theta)$, $Y = R\sin(\Theta)$. This transformation is bijective, it is the polar change of coordinates. The Jacobian of this transformation is given by

$$J(r,\theta) = \det\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \det\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} = r.$$

Now,

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x(r,\theta), y(r,\theta))|J(r,\theta)| = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(r^2\cos^2(\theta)+r^2\sin^2(\theta))} \cdot |r| = \frac{1}{2\pi\sigma^2} r e^{-\frac{r^2}{2\sigma^2}},$$

for $r \geqslant 0$ and $0 \leqslant \theta < 2\pi$. We can immediately conclude that R and Θ are independent. Namely, it is easy to see that we can write $f_{R,\Theta}(r,\theta) = g(r)h(\theta)$ for suitable functions g and h. Finally, a direct integration shows

that
$$f_R(r) = \begin{cases} \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} & \text{if } r \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$
and
$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi} & \text{if } 0 \leqslant \theta < 2\pi, \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 28.11 See Ash's exercise 2.8.16.

Exercise 29.1 See Ash's exercise 3.4.1.

Exercise 29.2 See Ash's exercise 3.4.3.

Exercise 29.3 See Ash's exercise 3.4.4.

Exercise 29.4 Recall that a direct computation shows that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

This proves the result when $n = 2$. Now, we proceed by induction. Let us assume the result is true for n and prove it for $n+1$. Noting $S_n = X_1 + \cdots + X_n$, we have

$$\begin{aligned}
&\mathrm{Var}(X_1 + \cdots + X_{n+1}) = \mathrm{Var}(S_n + X_{n+1}) \\
&= \mathrm{Var}(S_n) + \mathrm{Var}(X_{n+1}) + 2\mathrm{Cov}(S_n, X_{n+1}) \\
&= \sum_{i=1}^{n+1} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathrm{Cov}(X_i, X_j) + 2\mathrm{Cov}(X_{n+1}, X_1 + \cdots + X_n) \\
&= \sum_{i=1}^{n+1} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathrm{Cov}(X_i, X_j) + 2\sum_{j=1}^{n} \mathrm{Cov}(X_{n+1}, X_j) \\
&= \sum_{i=1}^{n+1} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n+1}\sum_{j=1}^{i-1} \mathrm{Cov}(X_i, X_j).
\end{aligned}$$

Exercise 30.1 See Ash's exercise 4.4.4.

Exercise 30.2 See Ash's exercise 4.4.7.

Exercise 31.1 See Ash's exercise 4.2.5.

Exercise 31.2 See Ash's exercise 4.3.1.

Exercise 31.3 See Ash's exercise 4.3.2.

Exercise 31.4 See Ash's exercise 4.3.4.

Exercise 31.5 See Ash's exercise 4.4.1.

Exercise 31.6 See Ash's exercise 4.4.3.

Exercise 31.7 See Ash's exercise 4.4.5.

Exercise 31.8 See Ash's exercise 4.4.9.

Exercise 31.9 See Ash's exercise 4.4.10.

Exercise 31.10 See Ash's exercise 4.4.11.

Exercise 31.11 See Ash's exercise 4.4.16.

Exercise 31.12 See Ash's exercise 4.4.17.

Exercise 32.1 We use the formula developped in class. We have $n = 10,000$, $a = 7940$, $b = 8080$, $p = 0.8$. Hence, $np = 8,000$, $np(1-p) = 1,600$ and $\sqrt{np(1-p)} = 40$. Now,

$$\begin{aligned} P\{7940 \leqslant X \leqslant 8080\} &= \Phi\left(\frac{8,080-8,000}{40}\right) - \Phi\left(\frac{7,940-8,000}{40}\right) = \Phi(2) - \Phi(-1.5) \\ &= \Phi(2) - 1 + \Phi(1.5) = 0.9772 + 0.9332 - 1 = 0.9104. \end{aligned}$$

Hence, there is 91.04% probability to find between 7,940 and 8,080 successes.

Exercise 33.1 See Ash's exercise 3.5.2.

Exercise 33.2 Consider an element x. If it does not belong to any A_i, then all of the indicator functions in the formula take the value 0 and the formula says $0 = 0$, which is true.

If x does belong to at least one A_i, then consider all sets A_i to which x belongs. There is no loss in generality when assuming these sets are $A_1, \cdots, A_r$ for some $r \geqslant 1$. (Otherwise, simply rename the sets!) The left-hand side of the formula is 1. So we need to show that the right-hand side is also 1.

The indicator functions on the right-hand side take the value 0 unless all the indices $j_1, \ldots, j_i$ are among $\{1, \ldots, r\}$, in which case the indicator function takes the value 1. Moreover, for a given $i \leqslant r$ the number of possible choices of distinct integers $j_1, \ldots, j_i$ from $\{1, \ldots, r\}$ is $\binom{r}{i}$. Hence, the right-hand side in fact equals

$$\sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i}.$$

Now since $(1-1)^r = 0$ we can use the binomial formula to write

$$\sum_{i=0}^{r} (-1)^{i} \binom{r}{i} = 0.$$

But then

$$\sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} = 1 - \sum_{i=0}^{r} (-1)^{i} \binom{r}{i} = 1.$$

Exercise 33.3 See Ash's exercise 3.5.6.

Exercise 33.4 See Ash's exercise 3.7.1.

Exercise 33.5 See Ash's exercise 3.7.3.

Exercise 34.1 By Example 34.13,

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha} \qquad \text{and} \qquad M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\beta}$$

Then,

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}$$

and $X + Y$ follows a Gamma distribution with parameters $\alpha + \beta$ and λ.

Exercise 34.2 For $i = 1, \ldots, n$, we have $M_{X_i}(t) = \exp(\mu_i t + \frac{\sigma_i^2 t^2}{2})$. Then,

$$\begin{aligned} M_{X_1+\cdots+X_n}(t) &= M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t) \\ &= \exp((\mu_1 + \cdots + \mu_n)t + (\sigma_1^2 + \cdots + \sigma_n^2)\frac{t^2}{2}). \end{aligned}$$

Identifying the moment generating function, this proves the result.

Exercise 34.3

(a) It's not an mgf, it can take negative values.

(b) It's not an mgf, $M(0) \neq 1$.

(c) It is the mgf of an exponential random variable with parameter $\lambda = 1$. (See Example 34.13)

(d) It is the mgf of a discrete random variable taking values $-2, 0, 2, 13$ with respective probabilities $\frac{1}{12}, \frac{1}{3}, \frac{1}{2}, \frac{1}{12}$.

Exercise 34.4

$$M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = E[e^{bt}e^{taX}] = e^{bt}E[e^{taX}] = e^{bt}M_X(at).$$

Exercise 34.5 To be added in the future.

Exercise 35.1 For $i = 1, \ldots, n$, we have $M_{X_i}(t) = \exp(\lambda(e^t - 1))$. Hence,

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t) = \exp(n\lambda(e^t - 1)).$$

As a consequence, $X_1 + \ldots + X_n$ has a Poisson distribution with parameter $n\lambda$.

Then, setting $Z_n = \frac{X_1+\cdots+X_n - n\lambda}{\sqrt{n\lambda}}$, we have

$$\begin{aligned} M_{Z_n}(t) = E[e^{tZ_n}] &= e^{-t\sqrt{n\lambda}} E[e^{\frac{t}{\sqrt{n\lambda}}(X_1+\cdots+X_n)}] \\ &= e^{-t\sqrt{n\lambda}} M_{X_1+\cdots+X_n}(\frac{t}{\sqrt{n\lambda}}) = e^{-t\sqrt{n\lambda}} \exp(n\lambda(e^{\frac{t}{\sqrt{n\lambda}}} - 1)). \end{aligned}$$

Using de l'Hospital's rule, we can prove that, as $n \to \infty$, this function converges to $\exp(\frac{t^2}{2})$, hence to a standard normal distribution.

Exercise 35.2 For $t < 1$, the mgf is given by

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-2}^{\infty} e^{tx} e^{-(x+2)}\, dx = \int_{-2}^{\infty} e^{(t-1)x-2}\, dx \\ &= \frac{1}{t-1}\, e^{(t-1)x-2}\Big|_{-2}^{\infty} = \frac{1}{1-t}\, e^{-2t}. \end{aligned}$$

Then,

$$M'_X(t) = \frac{(2t-1)e^{-2t}}{(1-t)^2} \qquad \text{and} \qquad E[X] = M'_X(0) = -1,$$

and

$$M''_X(t) = \frac{2(2t^2 - 2t + 1)e^{-2t}}{(1-t)^3} \qquad \text{and} \qquad E[X^2] = M''_X(0) = 2.$$

Exercise 35.3 We have $M_{X_n}(t) = \frac{\frac{\lambda}{n} e^t}{1-(1-\frac{\lambda}{n})e^t}$. We have

$$M_{X_n/n}(t) = M_{X_n}(\frac{t}{n}) = \frac{\frac{\lambda}{n} e^{\frac{t}{n}}}{1 - (1 - \frac{\lambda}{n})e^{\frac{t}{n}}}.$$

Then,

$$\lim_{n\to\infty} M_{X_n/n}(t) = \lim_{h\to 0} \frac{\lambda h e^{ht}}{1 - (1 - \lambda h)e^{ht}} = \frac{\lambda}{\lambda - t}.$$

Identifying the mgf, we see that $\frac{X_n}{n}$ converges in distribution to an exponential random variable of parameter λ.

Exercise 35.4 To be added in the future.

Exercise 36.1 The probability is approximately 0.8512.

Exercise 36.2 (a) The probability is approximately 0.9393; (b) $a = 11.65$.

Exercise 36.3 The probability is approximately 0.4359.

Exercise 36.4 The random variable X has binomial distribution with parameters $n = 10,000$ and $p = 0.8$. Hence, $E[X] = np = 8000$ and $\text{Var}(X) = np(1-p) = 1600$. Now, by the Central Limit Theorem, we know that $Z = (X - np)/\sqrt{np(1-p)} = \frac{X-8000}{\sqrt{1600}}$ follows approximately a $N(0,1)$ distribution. Hence,

$$\begin{aligned} P(7940 \leqslant X \leqslant 8080) &= P\left(\frac{7940-8000}{40} \leqslant \frac{X-8000}{40} \leqslant \frac{8080-8000}{40}\right) \\ &= P\left(-\frac{3}{2} \leqslant Z \leqslant 2\right) \simeq \Phi(2) - \Phi\left(-\frac{3}{2}\right) = 0.977 - 0.067 = 0.910. \end{aligned}$$

The probability that the player scores between 7940 and 8080 baskets is approximately 91%.

Appendix C

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8314	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

γ	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995
z_γ	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417

Figure C.1. ©1991 Introduction to Probability and Mathematical Statistics, 2nd Edition, by Bain & Engelhardt