Math 5010
Introduction to Probability

Davar Khoshnevisan and Firas Rassoul-Agha
University of Utah

Last Scribed November 15, 2022

Acknowledgments: Earlier versions of the notes were based on D. Stirzaker?s book then in part also on R.B. Ash's book. The current version was substantially rewritten and is based on *Introduction to Probability* by Anderson, Seppäläinen, and Valkó, Cambridge University Press.

We thank Daniel Conus for considerably expanding the list of exercises and Ryan Viertel for catching a large number of typos in an earlier version. We thank Chris Janjigian for a substantial amount of useful comments and suggestions on the previous version of the notes, many of which we incorporated in the current version. We thank Daniel Conus, Stewart Ethier, Nicos Georgiou, Kunwoo Kim, and Ryan Viertel for valuable comments.

Syllabus

Here is a rough outline for a 15-week course that meets three times a week for 50-minute lectures.

Week	Lecture	Topic
Week 1	Lecture 1 Lecture 2 Lecture 3	Probability models Rule of probability Properties of probability
Week 2	Lectures 4 & 5 Lecture 6	Counting problems Infinite state space
Week 3	Lecture 7 Lecture 8	Conditional probability Independence
Week 4	Lecture 9 Lecture 10 **Test 1**	Random variables Examples of discrete random variables Lectures 1-8
Week 5	Lecture 11	Continuous random variables and examples
Week 6	Lecture 12 Lecture 13	Mathematical expectation: discrete case Mathematical expectation: continuous case, and Variance
Week 7	Lecture 14	Law of large numbers and central limit theorem for Binomials
Week 8	Lecture 15 **Test 2**	Joint probability mass functions Lectures 9-14
Week 9	Lecture 16 Lecture 17	Joint probability density functions Covariance and correlation
Week 10	Lecture 18 Lecture 19	Moment generating function Gamma random variables
Week 11	Lecture 20 Lecture 21	Distributional convergence The central limit theorem and the law of large numbers
Week 12	Lecture 22 **Test 3**	Cumulative distribution functions Lectures 14-21
Week 13	Lecture 23	Distribution of functions of a random variable
Week 14	Lecture 24	Conditioning
Week 15	**Test 4**	Lectures 20-24

Lecture 1

1. Probability models

Probability theory is the field of mathematics that studies "randomness". More precisely, a probabilist develops and studies models that describe experiments involving processes that are too complicated to predict exactly. The goal is to be able to predict the outcomes of such experiments.

One starts by defining the space of all possible *outcomes* of the experiment we are modeling. We will denote this space by Ω and will denote a generic element of Ω by ω. In statistics, Ω is referred to as the *population* and the elements of Ω are called *samples*. In statistical mechanics, the elements of Ω are the possible *states* of the system and Ω is called the *state space* or the *phase space*.

Let us for now consider the case of a finite state space. This means that Ω has finitely many elements, i.e. our experiment has only finitely many possible outcomes. Later we will see how to treat the more general case of a countable or even uncountable state space. But starting with the finite case allows us to build some intuition for what is going on.

We are modeling an experiment in which the outcomes come out through a very complicated mechanism, so unpredictable that we consider it to be random. Hence, the next step in building the probability model is to specify the probability for each of the outcomes to occur when we run the experiment. This means to assign to each outcome a number in $[0, 1]$. Intuitively, this number gives the fraction of time we observe this particular outcome if we repeat the experiment a very large number of times. Consequently, this assignment of probabilities must be such that adding the numbers over all possible outcomes results in a total of one.

Example 1.1. Tossing a coin. A natural sample space is

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

To model a fair coin we assign the numbers $1/2$ to each of the two outcomes. We write $P(H) = P(T) = 1/2$. But we can also model biased coins. For example assigning $P(H) = 0.8$ and $P(T) = 0.2$ means we are modeling a coin toss where the coin comes up heads 80% of the time. More generally, for any $p \in [0, 1]$ a p-coin is one where $P(H) = p$ and $P(T) = 1 - p$.

Example 1.2. Rolling a six-sided die. A natural sample space is:

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

[0] Last modified on April 22, 2020 at 23:31:53 -06'00'

Then the model of a fair die consists of assigning a probability of 1/6 to each of the six outcomes. Again, we can also model a loaded die by assigning uneven probabilities to the different outcomes.

Example 1.3. Tossing two coins. A natural sample space is

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

Tossing two fair coins independently means we assign a probability of 1/4 to each of the four outcomes. We can also model a situation where we cheat and first toss a fair coin and then place the second coin on the table so that it matches the outcome of the first coin. This corresponds to assigning the probabilities $P(H_1, H_2) = P(T_1, T_2) = 1/2$ and $P(H_1, T_2) = P(T_1, H_2) = 0$. We will return to this example at the end of the next lecture.

Once we assigned the probabilities to the different outcomes we can ask about the probabilities of more complicated things. For example, what is the probability that our die lands an even number of pips? More mathematically, a subset of the state space is called an event. In our example, the event in question is $A = \{2, 4, 6\}$.

Intuition says that the probability of A could be computed by running the experiment a large number of times and computing the fraction of time the outcome was in A. But then this means that we can compute this probability by adding the probabilities of the different outcomes in A. That is,

$$P(A) = P(2) + P(4) + P(6).$$

And more generally,

$$P(A) = \sum_{\omega \in A} P(\omega). \tag{1.1}$$

To summarize: a probability model consists of defining the set of all possible (finitely many) outcomes then assigning probabilities to these outcomes (numbers between 0 and 1 that add up to 1) and then defining the probability of an event (a set of outcomes of interest) to be the sum of the probabilities of the individual outcomes.

2. Equally-likely outcomes

One way to assign the probabilities to the different outcomes is to assume a symmetry that says that in fact all outcomes of the experiment are alike and thus have the same probability to occur. Mathematically, suppose Ω has N distinct elements ("N distinct outcomes of the experiment"). If we know all outcomes have the same probability, then it must be that

$$P(\omega) = \frac{1}{N}.$$

And then if we denote by $|A|$ the number of elements of A (so for example $|\Omega| = N$), then

$$P(A) = \sum_{\omega \in A} \frac{1}{N} = \frac{|A|}{N} = \frac{|A|}{|\Omega|}.$$

Example 1.4. Consider the experiment of tossing two fair coins independently. We have seen that

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

and that in this model it is assumed that all outcomes are equally likely. Then,

$$\begin{aligned}\mathrm{P}(\text{“coin \#1 comes up Heads”}) &= \mathrm{P}\Big(\{H_1, H_2\} \cup \{H_1, T_2\}\Big) \\ &= \mathrm{P}(\{H_1, H_2\}) + \mathrm{P}(\{H_1, T_2\}) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.\end{aligned}$$

So coin #1 is fair. Similarly we get that coin #2 is fair. (Check this!) We will see later that in this model the two coins are also independent. (We will define what this means.) Hence we say this is a model of two fair coins tossed independently.

Example 1.5. Let us continue with the sample space of the previous example, but assign probabilities differently. Here, we define $\mathrm{P}(\{H_1, H_2\}) = \mathrm{P}(\{T_1, T_2\}) = 1/2$ and $\mathrm{P}(\{H_1, T_2\}) = \mathrm{P}(\{T_1, H_2\}) = 0$. We compute, as we did before, to find that both coins are again fair. (Do the computation yourself!) But now the coins are not tossed independently. 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 placed on the table to match the result of the first coin. Thus, if someone only observes the second coin it will appear fair to them (because it is a copy of the first coin, which is a fair coin!). But the second toss does depend on the first one and is not hence an independent toss.

3. Word of caution

Consider the experiment of tossing two identical fair coins. What is 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. This would be a perfectly fine mathematical model. But it would not be modeling tossing two fair coins independently. For example, if we do the tossing a large number of times and observe the fraction of time we got two different faces, this fraction will not be close to 1/3. It will in fact be close to 1/2. (Do the experiment and check for yourself!)

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

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

This matches what an empirical experiment would give, and hence is the more accurate model of a toss of two fair coins.

The upshot is that outcomes in a state space do not have to be equally likely. (Of course, we have already seen this when we modeled tossing a loaded coin.)

4. Rolling dice

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

4.1. 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 have already seen we can assign $\mathrm{P}(A) = |A|/|\Omega|$ for any event A. Therefore, the first question we address is, "how many items are in Ω?" We can think of Ω as a 6-by-6 table; so $|\Omega| = 6 \times 6 = 36$.

Before we proceed with our example, let us document this observation more abstractly.

Proposition 1.6 (The second principle of counting)**.** *If we have* m *distinct forks and* n *distinct knives, then* mn *distinct knife–fork combinations are possible.*

. . . not to be mistaken with . . .

Proposition 1.7 (The first principle of counting)**.** *If we have* m *distinct forks and* n *distinct knives, then there are* $m + n$ *utensils.*

. . . back to our problem . . .

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

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

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

4.3. 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 $\mathrm{P}(A) = |A|/36 = 4/36 = 1/9$.

4.4. 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 thus find $\mathrm{P}(A) = 10/36$.

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

$$A = \begin{Bmatrix} (1,1), & (1,3), & (1,5) \\ (3,1), & (3,3), & (3,5) \\ (5,1), & (5,3), & (5,5) \end{Bmatrix}.$$

And $\mathrm{P}(A) = 9/36 = 1/4$.

5. Easy cards

There are 52 cards in a deck. You deal two cards, all pairs equally likely.

The state space Ω is the collection of all pairs, where the pair "ace of hearts" and "king of diamonds" is the same as the pair "king of diamonds" and "ace of hearts". The selection process is thus called *unordered*, since order does not matter.

If instead we were picking one card from the deck and then a second card, then the pair would be *ordered* because getting an ace of hearts first and then a king of diamonds is different from getting the king of diamonds first. To emphasize that "ace of hearts" and "ace of hearts" is not an admissible pair we say that this selection is *without replacement*. The idea being that we select the first card and then the second card is selected without putting the first card back into the deck.

What is $|\Omega|$ in the unordered case? To answer this note that the ordered and unordered cases are related. Indeed, to pick two cards out of the deck in an unordered manner, we could pick them one at a time and then ignore the order. This means that the number of ordered pairs is twice the number of unordered pairs. But the number of ordered pairs is 52×51 since we have 52 options for the first card and then 51 options for the second card. Therefore, in the unordered case

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

Now we can compute the probabilities of many events.

Example 1.8. There are 48 unordered pairs where one card is the ace of hearts and the other is not an ace. Similarly, there are 48 unordered pairs where one card is the ace of spades and the other is not an ace. Therefore, there are $4 \times 48 = 192$ unordered pairs with exactly one ace. Thus, the probability that exactly one card is an ace is $192/1326 \simeq 0.1448$.

If we would rather enforce order, then our state space would have $52 \times 51 = 2652$ outcomes and the event in question would have $4 \times 48 + 48 \times 4 = 384$ out comes (first card is an ace and the second is not or the first card is not an ace and the second is an ace). So the probability that exactly one card is an ace is $384/2652 \simeq 0.1448$.

Note that what was important here is not whether order mattered or not but rather to track down what the state space is and what the event in question is.

Example 1.9. Similarly to how we computed $|\Omega|$, the number of unordered pairs of cards that are both aces is $(4 \times 3)/2 = 6$. Thus, the probability both cards are aces equals $6/1326 \simeq 0.0045$.

Example 1.10. Similarly, the probability there is exactly one heart equals $13 \times 39/1326 \simeq 0.38235$ and the probability both cards are hearts is $(13 \times 12/2)/1326 \simeq 0.05882$.

Example 1.11. The probability none of the two cards is a heart equals $(39 \times 38/2)/1326 \simeq 0.55882$.

Example 1.12. The probability that both cards are the same is

$$\mathrm{P}\{\text{ace and ace}\} + \cdots + \mathrm{P}\{\text{king and king}\} = 13 \times 6/1326 \simeq 0.0588.$$

Read sections 1.1 and 1.2 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 1.1. We roll a die 3 times. Give a sample space Ω.

Exercise 1.2. An urn contains three chips: one black, one green, and one red. We draw one chip at random. Give a sample space Ω.

Exercise 1.3. A fair die is rolled 5 times and the sequence of scores recorded.

(a) How many outcomes are there?

(b) Find the probability that the first and last rolls are 6.

Exercise 1.4. 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 1.5. 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 with the alphabetical part starting with an A?

Exercises 1.1, 1.2, 1.3, 1.4.a, 1.4.b, 1.5.a on pages 29 and 30 in the textbook by Anderson, Sepäläinen, and Valkó.

1. Rules of probability

The procedure of assigning probabilities to events, as described in Lecture 1, respects certain rules.

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

Proof. Each $P(\omega)$ is nonnegative and so (1.1) implies $P(A)$ is always nonnegative. Furthermore, by design, if we add $P(\omega)$ over all $\omega \in \Omega$ we would get one. So if we add over only $\omega \in A$ we will get at most one. In other words, $P(A) \leqslant 1$ always. □

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

Proof. Recall (1.1) and plug in Ω for A. Then recall that we assigned the probabilities $P(\omega)$ to add up to one when we add over all $\omega \in \Omega$. □

Recall that the empty set (the set with no outcomes in it, or the event that "nothing happened") is denoted by $\varnothing$.

Rule 3 (Addition rule)**.** If A and B are disjoint events [i.e., they do not have outcomes in common, or mathematically $A \cap B = \varnothing$], then the probability that "the outcome belongs to at least one of the two events" is the sum of the two probabilities. Mathematically put,

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

Proof. This is simply the associative property of addition:

$$P(A \cup B) = \sum_{\omega \in A \cup B} P(\omega) = \sum_{\omega \in A} P(\omega) + \sum_{\omega \in B} P(\omega) = P(A) + P(B).$$

The second equality is correct because A and B do not have any elements in common. So adding over the union $A \cup B$ can be done by first adding over the elements of A then over the elements of B. □

[0] Last modified on September 04, 2020 at 16:30:36 -06'00'

Note that $\Omega \cap \varnothing = \varnothing$ and hence these two events are disjoint. Furthermore, $\Omega \cup \varnothing = \Omega$. So Rule 3, when applied to the two disjoint events Ω and $\varnothing$, implies the following:

$$P(\Omega) = P(\Omega) + P(\varnothing).$$

Canceling $P(\Omega)$ on both sides gives that $P(\varnothing) = 0$. This makes sense: the probability that nothing happens is zero. (We ARE running the experiment and it MUST result in an outcome.)

Remark 2.1. Recall the notion of intersection of two sets $A \cap B$. The event $A \cap B$ is the event that "the outcome of the experiment belongs to both A and B". The probability $P(A \cap B)$ is sometimes written as $P(A, B)$.

As one can see from the above, it is important to know some set-theoretical notation and facts.

2. 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 $A \setminus B$ 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 $A \subset B$ when A is inside B; i.e. $x \in A$ implies $x \in B$. It is clear that if $A \subset B$, then $A \cap B = A$ and $A \cup B = B$. Thus, if $A_1 \subset A_2 \subset \cdots \subset A_n$, 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 pairwise disjoint if $A_i \cap A_j = \varnothing$, for all $i \neq j$.

Example 2.2. The sets $\{1,2\}$, $\{2,3\}$, and $\{1,3\}$ are disjoint but not pair-wise 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 $B \cap (A \cup C) \setminus D$ 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. Let $A = \{1,3,7,13\}$, $B = \{2,3,4,13,15\}$, $C = \{1,2,3,4,17\}$, and $D = \{13,17,30\}$. Then, $A \cup C = \{1,2,3,4,7,13,17\}$, $B \cap (A \cup C) = \{2,3,4,13\}$, and $B \cap (A \cup C) \setminus D = \{2,3,4\}$. Similarly, $A \cap B^c = \{1,7\}$, $C \cap D = \{17\}$, and $(A \cap B^c) \cup (C \cap D) = \{1,7,17\}$.

We have the following distributive relation.

Lemma 2.5. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *and* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof. First, we show that $A \cup (B \cap C) \subset (A \cup B) \cap (A \cup C)$. 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 $(A \cup B) \cap (A \cup C) \subset A \cup (B \cap C)$. 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

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

Homework Problems

Exercise 2.1. 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 2.2. Let A, B and C be events in a sample space Ω. Justify 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 2.3. 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 2.4. 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 2.5. If $A_n \subset A_{n-1} \subset \cdots \subset A_1$, show that $\cap_{i=1}^n A_i = A_n$ and $\cup_{i=1}^n A_i = A_1$.

Exercise 2.6. A public opinion poll (fictional) consists of the following three questions:

(1) Are you a registered Democrat?

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

(3) Do you favor Bill X?

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 yes answers and one no). 100 people answer yes to all three questions. 125 registered Democrats approve of the President's performance. How many of those who favor Bill X do not approve of the President's performance and in addition are not registered Democrats? (Hint: use a Venn diagram.)

Exercise 2.7. Let A and B be events in a sample space Ω. We remind that $A \setminus B = A \cap B^c$. Explain why the following hold.

(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 2.8. Show that if $A_1, \ldots, A_n$ are pairwise disjoint then A_1 and $A_2 \cup \cdots \cup A_n$ are disjoint.

Lecture 3

1. Properties of probability

Rules 1–3 have some consequences.

Lemma 3.1. *Choose and fix an integer $n \geqslant 1$. If $A_1, A_2, \ldots, A_n$ are pairwise disjoint events, then*

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

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 $n = 3$. But now that 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. (Check this!) 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. □

The above result is usually applied as follows: we have a complicated event that turns out to be decomposable into a union of simpler pairwise disjoint events. Then to compute the probability of the complicated event we compute the probabilities of the simpler events and add them up. A baby version of this is the definition (1.1) of the probability of an event. (Indeed, the events $\{\omega\}$, where ω varies in A, are pairwise disjoint!)

Example 3.2. Consider a deck of 52 cards. Suppose you pick two cards out of the deck. What is the probability there is at least one heart among the two cards?

To compute this probability we can decompose the event "there is at least one heart" into the union of the two clearly disjoint events: "there is exactly one heart" and "both cards are hearts".

[0] Last modified on September 04, 2020 at 16:30:58 -06'00'

The probabilities of these two events were computed in Example 1.10. The probability of the event of interest is then the sum of these two probabilities and equals approximately 0.44118.

The above lemma itself has a few consequences.

Example 3.3. Let $A \subset B$. 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

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

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

$$P(A^c) = 1 - P(A). \tag{3.1}$$

"The probability A does NOT happen is one minus the probability it DOES happen."

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

Example 3.4. One can use (3.1) to give an alternate solution to Example 3.2. Namely, in Example 1.11 we computed the probability of "none of the two cards is a heart" to be approximately 0.55882. Thus, the probability of the event in question is approximately $1 - 0.55882 = 0.44118$.

Note that if we were to pick 10 cards from the deck, instead of only two, then this second method is faster than the one presented in Example 3.2 since here we only need to compute the probability of the event "none of the 10 cards is a heart" as opposed to computing the probabilities of 10 events ("exactly one card is a heart", "exactly two cards are hearts", etc).

Example 3.5. Since $P(B \setminus A) \geqslant 0$, the above also shows another physically-appealing property:

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

"If all of the outcomes in A are included in B then B is at least as likely to happen as A is."

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

Lemma 3.6 (Another addition rule)**.** *If A and B are events (not necessarily disjoint), then*

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

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,

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

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

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

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

$$P(B) = P(A^c \cap B) + P(A \cap B). \tag{3.5}$$

Add (3.4) and (3.5) and solve to find that

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

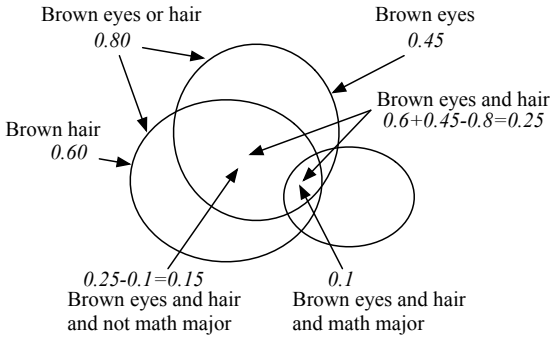


Figure 3.1. Venn diagram for Example 3.7.

Plug this into the right-hand side of (3.3) to finish the proof. □

Example 3.7. 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,

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

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

Formula (3.2) has a generalization. The following is called the "inclusion-exclusion" principle.

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

For example,

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ & - P(B \cap C) + P(A \cap B \cap C). \end{aligned} \tag{3.6}$$

Proving the inclusion-exclusion formula is deferred to a later exercise.

As a corollary of the above lemma we have the following useful fact, called a *union bound*.

Lemma 3.8. *For any integer $n \geqslant 1$, if $A_1, \ldots, A_n$ are events (not necessarily disjoint), then*

$$P(A_1 \cup \cdots \cup A_n) \leqslant P(A_1) + \cdots + P(A_n).$$

The proof of the lemma goes by induction using the previous lemma. (Do it!)

Such inequalities are useful because often times one cannot compute a probability exactly and yet one would like to get an estimate that says the probability of a certain event is not too high. If one can rewrite this event of interest as a union of some other events whose probabilities can be computed, then using the above union bound one gets an upper bound on the probability of the original event.

Read section 1.4 of the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 3.1. 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.

Exercise 3.2. Give an example to show that $P(A \setminus B)$ does not need to equal $P(A) - P(B)$.

Exercise 3.3. In a town 15% of the population is blond, 25% of the population has blue eyes and 2% of the population is blond with blue eyes. What is the probability that a randomly chosen individual from the town is not blond and does not have blue eyes? (We assume that each individual has the same probability to be chosen.)

Exercises 1.13, 1.14 on page 31 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 4

1. The birthday problem

There are n people in a room. Suppose 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. Let us start by first finding $p(2)$. So there are two people in the room and the sample space is the collection of all pairs of the form (D_1, D_2), where D_1 and D_2 are birthdays. Then $|\Omega| = 365^2$.

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 distinct birthdays. 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, $|A| = 365 \times \cdots \times (365 - n + 1)$.

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, once $n \geqslant 23$, we have $p(n) < 0.5$.

2. Ordered selection with replacement

Theorem 4.1. *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. There are n options for the first ball. For each choice of the first ball there are n options for the second ball. Thus there are n^2 options for the first two balls. For each choice of the first two balls there are n options for the third ball. So there are n^3 options for the first three balls. By induction, there are n^k options for the k balls. □

[0] Last modified on February 27, 2020 at 09:33:35 -07'00'

Example 4.2. What is the number of functions from a set A to a set B? To see the answer suppose $|A| = k$ and $|B| = n$. A function from A to B consists of assigning an element in B to each element in A. Repetition is allowed (so we can assign the same element of B to multiple elements in A). Hence, there are n possible options to assign to each element in A, making the total number of functions n^k.

Example 4.3. What is the probability that 10 people, 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 4.4. 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 4.5. We roll a fair die, draw a card from a standard deck, and toss a fair coin. Then, the probability that the die score is even, the card is a heart, and the coin is heads is equal to $\frac{3\times13\times1}{6\times52\times2} = 1/16$.

Example 4.6. 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

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

3. Ordered selection without replacement: Permutations

The following comes directly from the second principle of counting.

Theorem 4.7. *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 4.8. 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 4.9. 6 dice are rolled. What is the probability that they all show different faces? A sample space is the set of all 6-tuples of numbers from 1 to 6. That is, $\Omega = \{1,\dots,6\}^6$. Then $|\Omega| = 6^6$. If A is the event in question, then $|A| = 6\times5\times4\times3\times2\times1 = 6!$. The probability is $\frac{6!}{6^6}$.

Example 4.10. We roll a fair die five times. What is $P(A)$, where A is the event that all five rolls 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 4.11. The number of permutations of cards in a regular 52-card deck is $52! > 8\times10^{68}$. 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 4.12. 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 $4!\times2^4$ ways to seat couples together. The probability is thus $\frac{4!\times2^4}{8!} = 1/105$.

Example 4.13. n purple and n orange balls are in an urn. All balls are identical except for their color. You select one ball at random and then another ball, also at random, without replacing the first ball back into the urn. What are the chances that they have different colors?

Let us number the purple balls 1 through n and the orange balls $n+1$ through $2n$. This is only for convenience, so that we can define a sample space. The balls of course do not know they are numbered!

The sample space Ω is then the collection of all pairs of distinct numbers 1 through $2n$. Note that $|\Omega| = 2n(2n-1)$. Since the balls are identical, the person pulling the balls out of the urn does not differentiate between the $2n$ balls and does not know the numbers nor the colors of the balls until they look at the balls that came out. This means that all outcomes are equally likely.

We have

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

Furthermore, decomposing according to the color we have

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

where O_j denotes the event that the jth ball is orange, and P_k the event that the k-th 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,

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

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

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

Homework Problems

Exercise 4.1. 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 4.2. 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 4.3. 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 4.4. 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 4.5. 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 4.6. 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 4.7. 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 4.8. Eight cards are drawn without replacement from an ordinary deck. Find the probability of obtaining exactly three aces or exactly three kings (or both).

Exercises 1.6, 1.7, 1.8, 1.15 on pages 30 and 31 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 5

1. Unordered selection without replacement: Combinations

Theorem 5.1. *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 5.2. 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 $n-1$ ways 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 $n(n-1)/2 = \binom{n}{2}$.

Proof of Theorem 5.1. 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. □

[0] Last modified on September 04, 2020 at 16:26:44 -06'00'

Example 5.3. 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 5.4. 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. (That is, it matters if b is a heart and c is a diamond or if it is the other way around.) 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 on the position of the card. (So now, the dealer is giving the cards one at a time, and we are taking into account which card came first, which came second, and so on.) 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 5.5. Let A denote the event that we get two pairs $[a, a, b, b, c]$ in a poker hand. Then,

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

Another way to compute this (which some may find more intuitive) is as: $\binom{13}{3}$ is to pick the face values, times 3 to pick which face value is the single card and which are the two pairs, and then times $\binom{4}{2}^2 \times 4$ to deal the cards. Check that this gives the same answer as above.

In any case,

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

Homework Problems

Exercise 5.1. A lottery is played as follows: the player picks six numbers, without replacement, out of {1,2,...,54}. Then, six numbers are drawn at random out of the 54. You win the first prize 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 5.2. Another lottery is played as follows: the player picks five numbers, without replacement, out of {1,2,...,50} and two other numbers, without replacement, from the list {1,...,9}. An example is $\{1, 5, 7, 10, 50\}$ and $\{1, 8\}$. 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 5.3. 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).

Exercise 5.4. 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 5.5. 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 5.6. 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.

Exercise 5.7. 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 strictly increasing order (from left to right)?

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

Optional exercises:

Exercise 5.8. Prove the following two facts.

(a) For any integers $0 \leqslant k \leqslant n$

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

(b) For $1 \leqslant k \leqslant n-1$ integers we have

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \tag{5.1}$$

Exercise 5.9. Answer the following questions.

(a) Show that there are 2^n subsets of $\{1, \ldots, n\}$ (including $\varnothing$)?
Hint: Assign to each element of $\{1, \ldots, n\}$ a zero ["not in the subset"] or a one ["in the subset"].

(b) Prove that

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

Hint: An alternative way to compute the number of subsets of $\{1, \ldots, n\}$ is by computing the number of subsets of a given size and then adding over the size. For example, the empty set has 0 elements and there is only one such set. On the other hand, there are n subsets of size one. How many are there of size 2? 3? etc.

(c) Prove the binomial theorem (part (b) is then the special case $x = y = 1$):

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

Hint: Either use induction and do the necessary algebra, or think of a combinatorial proof as follows: $(x+y)^n$ is $(x+y)$ multiplied n times. The term $x^j y^{n-j}$ therefore comes from picking x in j of these $(x+y)$'s and y in the remaining $n-j$ ones. In how many ways can you do that?

(d) What is the coefficient in front of x^3y^4 in $(2x-4y)^7$?

Optional read:

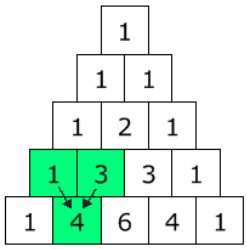
1. Similarly to the combinatorial proof of the binomial theorem, one can work out the coefficients in the multinomial theorem.

Theorem 5.6 (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 5.1.*

2. Formula (5.1) 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)]:



which gives the coefficients in the binomial theorem!

Lecture 6

1. Infinite state spaces

So far, all of our examples had finite state spaces. This is quite restrictive since often times it is not possible to model an experiment using a finite state space. For example, consider the experiment where we toss a fair coin repeatedly until it lands Heads and write down the number of tosses it took us. Clearly, all positive integers are possible outcomes of this experiment and hence a finite state space will not do. Another experiment in which the state space must be infinite is one in which we pick a number at random from the interval $[0, 1]$.

We distinguish two kinds of infinite state spaces: the countable ones and the uncountable ones.

Definition 6.1. A set A is said to be countable if it is either finite or it is infinite and we can find a bijection (i.e. a one-to-one and onto function) from $\mathbb{N}$ to A. Otherwise, A is said to be uncountable.

The above means that A is countable if we can write it in the form $\{a_1, a_2, a_3, \cdots\}$.

The typical countable state spaces we will see are $\mathbb{N}$ (positive integers), $\mathbb{Z}_+$ (nonnegative integers), and $\mathbb{Z}$ (integers). The typical uncountable state spaces we will see are $[a, b]$, $[a, b)$, $(a, b]$, (a, b) (intervals) with $a < b$, $(0, \infty)$ (positive real numbers), $\mathbb{R}_+ = [0, \infty)$ (nonnegative real numbers), and $\mathbb{R}$ (the real numbers).

1.1. Countable state space. Not much changes in the way we built probability models when the space is countable. One can still start by defining probabilities $P(\omega)$ on the outcomes $\omega \in \Omega$ and then define

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

Rules 1–3 from Lecture 2 and their consequences from Lecture 3 still all hold.

Example 6.2. Let us work out the probability assignment for the example mentioned at the beginning of the lecture. Namely, we are interested in the experiment where a fair coin is tossed repeatedly until it lands Heads. The outcome of the experiment is the number of tosses that were required. So all positive integers are possible outcomes. In principle, it may also happen that

[0] Last modified on September 04, 2020 at 16:26:44 -06'00'

we never get heads. In that case, we will say that we needed an infinite number of tosses. So a natural state space is $\Omega = \{\infty\} \cup \mathbb{N} = \{\infty, 1, 2, 3, \cdots\}$. (Here, think of ∞ as just a symbol. Saying that the outcome of the experiment was ∞ just conveys the message that we never got Heads.)

Now that we figured out the state space, let us figure out what the correct probability assignments should be. What is the probability that the outcome is n? To answer this question we can consider that we are tossing the coin n times and looking for the probability we get $n-1$ Tails followed by Heads on the n-th toss. But now that we put it this way, we can reduce our problem of computing the probability the outcome of the experiment is n to an equivalent computation that involves a smaller state space that has a total of 2^n possible outcomes (n tosses with two outcomes each) and only one desirable outcome ($n-1$ Tails followed by one Heads). The probability of this is $\frac{1}{2^n}$.

To summarize, our experiment has the outcome n with probability $\frac{1}{2^n}$.

Since $\sum_{n \geqslant 1} \frac{1}{2^n} = 1$ (see the remark below) we see that the remaining outcome ∞ has probability 0! In other words, we have shown that the coin will always eventually land Heads, i.e. there is zero chance it will keep landing Tails forever. So we actually did not need to include ∞ in our state space.

Of course, by symmetry, we also now know that there is zero chance the coin will land Heads forever. And putting the two facts together we see that the coin will in fact land infinitely many Heads and infinitely many Tails if we keep tossing it forever. An intuitive fact that we have just shown is supported by our model.

Remark 6.3. In the above we used the fact that $1/2 + 1/4 + 1/8 + \cdots = 1$. This comes from the geometric series computation. Namely, for any real number a and integers $n \geqslant m$

$$\begin{aligned}(1-a)(a^m + a^{m+1} + a^{m+2} + \cdots + a^{n-1} + a^n) \\ &= a^m + a^{m+1} + a^{m+2} + \cdots + a^{n-1} + a^n \\ &\qquad - a^{m+1} - a^{m+2} - \quad \cdots\cdots \quad - a^n - a^{n+1} \\ &= a^m - a^{n+1}.\end{aligned}$$

From this we get

$$a^m + a^{m+1} + \cdots + a^n = \frac{a^m - a^{n+1}}{1-a}.$$

If now $|a| < 1$ and we take $n \to \infty$ we get

$$\sum_{k=m}^{\infty} a^k = \frac{a^m}{1-a}. \tag{6.1}$$

Applying this with $m = 1$ and $a = 1/2$ we get what was claimed.

1.2. An uncountable state space. Suppose we want to model an experiment where we choose a number from $\Omega = [0, 1]$ with all outcomes being equally likely. Intuitively, this should be possible to model. But if we start by saying we assign a probability to each outcome $\omega \in \Omega$ and then define the probability of an event as the sum of the assignments over all the outcomes in the event, then we have a problem. Indeed, since we want all the outcomes to be equally likely we should assign the same number to each outcome. Say this number is p. Since we have infinitely many outcomes in $[0, 1]$ and we want the assignments to add up to one we would have $p \cdot \infty = 1$ which implies that $p = 0$. But then this means that none of our outcomes have a positive probability of ocurring!! Clearly, we are approaching things the wrong way.

The problem is that there are "too many outcomes" when the state space is uncountable. The right way to do things is not to define the probability model starting with assigning probabilities to outcomes. Instead, one goes directly to events. For example, if all outcomes are equally likely, then the probability we get an outcome in the interval $[0, 1/2]$ must be $1/2$, since this interval is exactly half the size of the state space $[0, 1]$. Similarly, the probability we get an outcome in the interval $[1/4, 1/2]$ should equal $1/4$ because the length of this interval is one fourth of the length of the original interval (the state space).

So a good probability model for picking a number in $[0, 1]$ at random with all outcomes being equally likely is to say that for a subset $A \subset [0, 1]$ the probability the outcome is in A equals the "length" of A. It is clear what we mean by "length" when A is made out of a union of a bunch of intervals. For example, the length of $[1/8, 1/4] \cup [1/2, 3/4]$ (the union of the two intervals) is $3/4-1/2+1/4-1/8 = 3/8$. The length of $[1/8, 1/2] \cup (1/4, 3/4) \cup [7/8, 1]$ is computed by rewriting this as $[1/8, 3/4) \cup [7/8, 1]$, which is now a union of two disjoint intervals and so the total length is $3/4 - 1/8 + 1 - 7/8 = 6/8 = 3/4$. There is a way to define the length of more complicated sets A, but to understand this one needs to learn a subject called *measure theory*. We will not go there because in this course we will not need to compute the length of sets that are more complicated than the union of a bunch of intervals.

We can generalize the above example to the situation where we pick a number from $[a, b]$ at random with all outcomes being equally likely. (Here, a and b are numbers such that $a < b$.) This is done by saying that the probability of an event $A \subset [a, b]$ is equal to the "length" of A divided by the length of $[a, b]$, which is $b - a$. If we denote the length of a set A by $|A|$ then we get the familiar formula:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{b - a}.$$

Note that Rules 1–3 from Lecture 2 are satisfied for the model we just defined. Indeed, for $A \subset \Omega$, the length of A is smaller than the length of Ω and both are nonnegative numbers. Hence, $P(A)$ is between 0 and 1. Also, $P(\Omega) = |\Omega|/|\Omega| = 1$. And for the third rule we observe that if A and B are disjoint the the length of $A \cup B$ is the sum of the lengths of the two sets. Dividing by the length of Ω we get that

$$P(A \cup B) = \frac{|A \cup B|}{|\Omega|} = \frac{|A| + |B|}{|\Omega|} = \frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|} = P(A) + P(B).$$

Since Rules 1–3 are satisfied, their consequences from Lecture 3 are all true as well.

The above can all be generalized to higher dimensions. For example, one can take Ω to be a subset of $\mathbb{R}^2$ (the two-dimensional plane) and model the experiment of picking a point in Ω at random with all outcomes being equally likely by assigning the probability

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

where now $|A|$ is the area of A (and $|\Omega|$ is the area of Ω). And in three dimensions we just replace "area" by "volume".

Example 6.4. The playable area of a dartboard has diameter $13\frac{1}{4}$ inches. The bullseye region in the center has diameter $1\frac{1}{4}$ inches. Suppose a player is throwing a dart so that they always hit the playable area on the board but the point where the dart lands is random, with all points equally likely. What is the probability of hitting the bullseye? The answer is the ratio of the area of the bullseye region to the area of the playable area. This is $\frac{1.25^2\pi}{13.25^2\pi} \simeq 0.0089$, i.e. less than 1% chance.

Read section 1.3 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 6.1. Explain why $\mathbb{Z}_+$ and $\mathbb{Z}$ are countable.

Exercise 6.2. We roll a fair die repeatedly until we see the number four appear and then we stop. The outcome of the experiment is the number of rolls. Describe a state space for this experiment. For a positive integer n, what is the probability the outcome equals n? What is the probability a four never appears?

Exercises 1.9, 1.11 on pages 30 and 31 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 7

1. Conditional probabilities

Example 7.1. 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

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

Someone has looked at the result of the sample and tells us that the person sampled is employed. We can then consider that the sample space has changed. It is now the set of the 12 employed people. And now, the probability of picking a female is equal to

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

To emphasize that we are computing this probability knowing the information that the person is employed, we write $P(\text{female} \,|\, \text{employed}) = 1/4$ and we read this as "the probability of choosing a female, given that the person is employed."

In light of the above example we make the following definition.

Definition 7.2. 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

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

Example 7.3. If we deal two cards fairly from a standard deck, the probability of $K_1 \cap K_2$ [$K_j = \{\text{King on the } j \text{ draw}\}$] 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 $|\Omega| = 52 \times 51$.

[0] Last modified on September 04, 2020 at 16:31:08 -06'00'

Similarly,

$$\begin{aligned}\mathrm{P}(\mathrm{K}_1 \cap \mathrm{K}_2 \cap \mathrm{K}_3) &= \mathrm{P}(\mathrm{K}_1) \times \frac{\mathrm{P}(\mathrm{K}_1 \cap \mathrm{K}_2)}{\mathrm{P}(\mathrm{K}_1)} \times \frac{\mathrm{P}(\mathrm{K}_3 \cap \mathrm{K}_1 \cap \mathrm{K}_2)}{\mathrm{P}(\mathrm{K}_1 \cap \mathrm{K}_2)} \\ &= \mathrm{P}(\mathrm{K}_1)\mathrm{P}(\mathrm{K}_2 \,|\, \mathrm{K}_1)\mathrm{P}(\mathrm{K}_3 \,|\, \mathrm{K}_1 \cap \mathrm{K}_2) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}.\end{aligned}$$

Or for that matter,

$$\mathrm{P}(\mathrm{K}_1 \cap \mathrm{K}_2 \cap \mathrm{K}_3 \cap \mathrm{K}_4) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}.$$

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

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

If, in addition, $0 < \mathrm{P}(B) < 1$*, then*

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

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

Example 7.5. Once again, we draw two cards from a standard deck. The probability $\mathrm{P}(\mathrm{K}_2)$ (second draw is a king, regardless of the first) is best computed by splitting it into the two disjoint cases: $\mathrm{K}_1 \cap \mathrm{K}_2$ and $\mathrm{K}_1^c \cap \mathrm{K}_2$. Thus,

$$\begin{aligned}\mathrm{P}(\mathrm{K}_2) &= \mathrm{P}(\mathrm{K}_2 \cap \mathrm{K}_1) + \mathrm{P}(\mathrm{K}_2 \cap \mathrm{K}_1^c) = \mathrm{P}(\mathrm{K}_1)\mathrm{P}(\mathrm{K}_2 \,|\, \mathrm{K}_1) + \mathrm{P}(\mathrm{K}_1^c)\mathrm{P}(\mathrm{K}_2 \,|\, \mathrm{K}_1^c) \\ &= \frac{4}{52} \times \frac{3}{51} + \frac{48}{52} \times \frac{4}{51}.\end{aligned}$$

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).

Example 7.6. 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 $\mathrm{P}(\Theta)$. The result of Theorem 7.4 gets replaced with

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

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

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

Figure 7.1. Thomas Bayes (1702 – Apr 17, 1761, England)

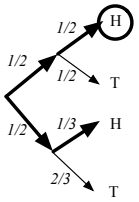


Figure 7.2. Boldface arrows indicated paths giving heads. The path going to the boldface circle corresponds to choosing the first coin and getting heads. Probabilities multiply along paths by Bayes' formula.

2. Bayes' Theorem

The following question arises from time to time: Suppose A and B are two events of positive probability. If we know $P(B \mid A)$ but want $P(A \mid B)$, then 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 7.4, and obtain

Theorem 7.7 (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 7.8. 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 7.9. 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 7.2),

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

Remark 7.10. 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 7.6 to compute the probability $P(\Theta)$). The corresponding diagram (analogous to Figure 7.2) could then have more than two branches out of each node. But the methodology is the same. See Exercise 7.6 for an example of this.

Example 7.11. Suppose that a test for a rare disease is 95% accurate in that 95% of people with this disease will test positive and 95% of people without this disease test negative. Suppose that 1% of the population actually has this disease. You go in for a routine checkup and when the doctor administers the test you test positive. What is the probability that you actually have the disease?

Let S denote the event I am sick and H denote the event I am healthy. Let $+$ denote the event I tested positive and $-$ denote the event I tested negative. Then I want to calculate the probability $P(S \mid +)$. This equals

$$P(S \mid +) = \frac{P(S \text{ and } +)}{P(+)}.$$

By the law of total probability we get

$$\begin{aligned} P(+) &= P(S \text{ and } +) + P(H \text{ and } +) = P(+ \mid S)P(S) + P(+ \mid H)P(H) \\ &= 0.95 \times 0.01 + 0.05 \times 0.99. \end{aligned}$$

And notice that while doing this computation we also computed

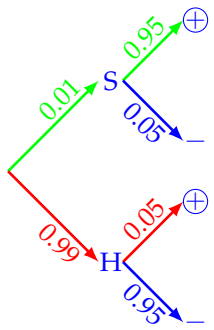
$$P(S \text{ and } +) = P(+ \mid S)P(S) = 0.95 \times 0.01.$$

So the answer is

$$P(S \mid +) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \approx 0.161.$$

So there is about 16.1% chance I actually have the disease.

We can also draw the tree that summarizes the information given in the problem and compute the probabilities from that diagram:



The answer then is the ratio of the product of the probabilities along the green arrows (along the path going through S AND +) divided by the sum of the product of probabilities along the green arrows and along the red arrows (the paths leading to a +).

Note how the probability of actually being sick given I tested positive is rather small (16.1%), even though the test seems quite accurate. The reason is that the disease is quite rare and the test is in fact not accurate enough. If for example the test had a 1% chance of a false positive, i.e. given that I am healthy there is a 1% chance of a positive test, then the probability of being sick given that the test is positive would have been

$$\mathrm{P}(S\,|\,+) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.01 \times 0.99} \approx 0.4897,$$

which is quite higher than the 16.1% we had when the probability of false positive was 5%, though it is still close to a fair coin flip. If on the other hand, the test were only 95% accurate, as in the original question, but it were not administered to the whole population, but only to those who are high risk (e.g. those who have certain genetic markers that say that there is a high chance of contracting the disease), then the probability of being sick in this high risk population is not 1%. Suppose for example it is 40% (i.e. 40% of those who have the genetic markers end up having the disease). Then the probability of being sick given that the test is positive becomes

$$\mathrm{P}(S\,|\,+) = \frac{0.95 \times 0.4}{0.95 \times 0.4 + 0.05 \times 0.6} \approx 0.9268.$$

This demonstrates why it is best to target the high risk group when testing for a rare disease.

Read sections 2.1 and 2.2 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 7.1. 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 7.2. 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 7.3. 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 7.4. 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 7.5. 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 7.6. 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 7.7. 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).

Exercises 2.1, 2.3, 2.5, 2.9 on pages 72 and 73 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 8

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. For example, subtracting both sides of the first equality from 1 gives the second equality. (Check that the third equality also follows from the first. Then the fourth comes again by subtracting both sides of the third equality from 1.) The first equality is also equivalent to:

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

Note that this is statement is now symmetric in A and B. So we see that if A is independent of B then also B is independent of A. Note also that this equation makes sense even if $P(B) = 0$ or $P(A) = 0$. Thus this is what we will use to say that A and B are independent.

Example 8.1. 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 8.2. 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 if A is independent of itself then $P(A) = P(A \cap A) = P(A)P(A)$.

It is noteworthy that disjoint events are in fact dependent. Indeed, if A and B are disjoint and we are told the outcome belongs to A then we know for a fact that the outcome does not belong to B. This is demonstrated in the following example.

Example 8.3. 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, let C be the event of getting a number less than or equal to 2. Then, $P(A \cap C) =$

[0] Last modified on February 27, 2020 at 09:35:31 -07'00'

$P\{2\} = 1/6$ 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 $\mathcal{E}_1$ and $\mathcal{E}_2$ are *independent* if A_1 and A_2 are independent for all choices of events A_1 and A_2 of experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively.

Example 8.4. 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 $P(T_1 \cap T_2) = P(T_1)P(T_2) = 1/4$, $P(T_1 \cap H_2) = P(T_1)P(H_2) = 1/4$, and $P(H_1 \cap H_2) = P(H_1)P(H_2) = 1/4$.

This computation works in the reverse direction as well. Namely, two fair coins are tossed independently, then all possible outcomes are equally likely to occur. (Check!)

Example 8.5. What if the coins are not fair, say $P(H_1) = P(H_2) = 1/4$? In this case, $P(H_1 \cap H_2) = P(H_1)P(H_2) = 1/16$ while for example $P(T_1 \cap T_2) = P(T_1)P(T_2) = 9/16$. So outcomes are no longer equally likely.

Note how in Examples 8.4 and 8.5 independence came from the assumption on the model. Namely, we are told that the two coins are independent and so any event A that involves only the first coin and any event C that involves only the second coin are independent. However, in Example 8.3 the independence of A and C is not that obvious and to prove it we needed to compute $P(A \cap C)$ and $P(A)P(C)$ and check that they are equal.

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

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \tag{8.1}$$

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

Example 8.6. 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 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 (8.1) does not hold and hence A_1, A_2, and A_3 are not independent.

Example 8.7. 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 8.8. 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 8.9. We now give an alternative solution to Example 4.6. First, decompose the event in question into a union of disjoint events by writing

$$P(A) = P(D1 \cap H_1) + P(D2 \cap H_1 \cap H_2) + \cdots.$$

But since the die and the coins are all independent we have $P(D_1 \cap H_1) = P(D_1)P(H_1)$, $P(D_1 \cap H_1 \cap H_2) = P(D_1)P(H_1)P(H_2)$ and so on. Hence,

$$P(A) = P(D1)P(H_1) + P(D2)P(H_1)P(H_2) + \cdots = \frac{1}{6}\Big(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^6}\Big).$$

Read sections 2.3 and 2.4 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 8.1. 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 8.2. 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 8.3. Argue for the correctness of 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 8.4. 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". Explain why these events are pairwise independent but not independent.

Exercise 8.5. 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.

Exercise 8.6. 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. The die roll and the coin tosses are all independent of each other. If the first n throws result in heads, what is the probability that the fair coin is being used?

Exercise 8.7. 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*?

Exercises 2.12, 2.15, 2.19 on pages 73 and 74 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 9

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. The state space is the set of students. Each student is a possible outcome. Assigning a value to each possible outcome is what a random variable does (the height, in this example).

Definition 9.1. 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 $\mathbb{R}$, or d-dimensional space $\mathbb{R}^d$.

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

Example 9.2. Define the sample space,

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}.$$

Then, the random variable $X(\boxdot) = 1$, $X(\boxdot) = 2$, ..., $X(\boxdot) = 6$ indicates the number of pips in a roll of a fair six-sided die. With the help of this random variable we can now express events such as "the die rolled an even number of pips" by writing "X is even".

Example 9.3. With the same state space as above, the random variable $Y(\boxdot) = Y(\boxdot) = 5$, $Y(\boxdot) = Y(\boxdot) = Y(\boxdot) = 2$, and $Y(\boxdot) = Y(\boxdot) = -1$ 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.

Note that 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 9.4. In the previous two examples 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

$$\mathrm{P}\left(\{\omega \in \Omega : X(\omega) = 3\}\right) = \mathrm{P}\left(\{\boxdot\}\right) = \frac{1}{6}. \tag{9.1}$$

The same holds if 3 is replaced by any of the numbers 1 through 6. The probability is zero if we replace 3 by a number other than the integers $\{1, \ldots, 6\}$, since X does not take such values.

[0] Last modified on February 27, 2020 at 09:35:55 -07'00'

Usually, we write $X \in A$ in place of the longer notation $\{\omega \in \Omega : X(\omega) \in A\}$. In this notation, we have

$$P\{X = k\} = \begin{cases} \frac{1}{6} & \text{if } k = 1, \ldots, 6, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

Also,

$$P(X \text{ is even}) = P(X \in \{2,4,6\}) = P\left(\{\boxdot, \boxdot, \boxdot\}\right) = \frac{1}{2}.$$

Similarly,

$$\begin{aligned} P\{Y = 5\} &= P\left(\{\boxdot, \boxdot\}\right) = \frac{1}{3}, \\ P\{Y = 2\} &= P\left(\{\boxdot, \boxdot, \boxdot\}\right) = \frac{1}{2}, \text{ and} \\ P\{Y = -1\} &= P\left(\{\boxdot\}\right) = \frac{1}{6}. \end{aligned} \tag{9.3}$$

Observe that we could have chosen our state space as

$$\Omega = \left\{ \text{𓅃}, \text{𓆑}, \text{𓅃}, \text{𓆛}, \text{𓅞}, \text{𓅬} \right\}.$$

If we then define the random variable as $Y(\text{𓅃}) = Y(\text{𓅃}) = 5$, $Y(\text{𓆑}) = Y(\text{𓅞}) = Y(\text{𓅬}) = 2$, and $Y(\text{𓆛}) = -1$, then we are still modeling the same game and (9.3) still holds.

Example 9.5. In fact, if we change the weights of our die so that $\boxdot$, $\boxdot$, $\boxdot$, $\boxdot$, $\boxdot$, and $\boxdot$ come up with probabilities, respectively, 1/12, 5/24, 1/4, 1/6, 1/24, and 1/4, then (9.3) still holds. For example

$$P\{Y = 5\} = P\left(\{\boxdot, \boxdot\}\right) = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}.$$

Thus, Y still models the same game even though we are using a different die!

Example 9.6. Here is another example demonstrating the above point. Suppose we toss a coin once. Suppose the coin is loaded so that heads come out twice as often as tails. A natural state space is $\Omega = \{H, T\}$. The probability that corresponds to the loaded coin is the one with the assignments $P(H) = 2/3$ and $P(T) = 1/3$. (Heads come out twice as often as tails.) Let X be the random variable with $X(H) = 1$ and $X(T) = 0$. Then we have

$$P(X = 0) = 1/3 \quad \text{and} \quad P(X = 1) = 2/3.$$

Now say we roll a fair die once. Then a natural state space is $\Omega' = \{1,2,3,4,5,6\}$. The probability that corresponds to a fair die is the one with the equal assignments of $1/6$ to each of the six outcomes. Let Y be the random variable with $Y(1) = Y(2) = Y(3) = Y(4) = 1$ and $Y(5) = Y(6) = 0$. Then we have

$$P(Y = 0) = 1/3 \quad \text{and} \quad P(Y = 1) = 2/3.$$

If someone gives us data from consecutive fair coin tosses where they recorded the values of X and then another set of data where they rolled the die many times and recorded the values of Y then there is no statistical test that would be able to tell which data set is which. This is because the frequency with which we will see 0's is the same in both cases and the frequency with which we will see 1's is the same in both cases.

The point is that what matters is not quite the state space nor the probability on it but the values X takes and the corresponding probabilities of X taking these various values.

2. Discrete random variables

In this section we will introduce one type of random variables that is encountered in many applications. Namely, a random variable X that takes values in a finite or countably-infinite set is said to be a *discrete* random variable. Its distribution is called a discrete distribution. The function

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

is then called the *mass function* of X. (If there are more than one random variable involved, e.g. X and Y, then we write f_X and f_Y to make it clear whose mass function we are talking about.)

To compute the probability that X is in some set A we simply write

$$P\{X \in A\} = \sum_{x \in A} P\{X = x\} = \sum_{x \in A} f(x). \tag{9.4}$$

Here are two important properties of mass functions to keep in mind:

- $0 \leqslant f(x) \leqslant 1$ for all x.
- $\sum_{x \in D} f(x) = 1$.

The first one is clear because $f(x) = P(X = X)$. The second one comes by taking $A = D$ in (9.4) and noting that of course $P(X \in D) = 1$ since D has all the possible values X can take.

As was explained in the previous section, knowledge of the mass function is sufficient to determine the distribution of X. In other words, if two random variables have the same mass function then they are statistically indistinguishable. Or, as one of the students in the class put it nicely: "If you run an experiment and collect data on a certain random variable and I run a different experiment and collect data on a different random variable and then we discover that the two random variables happen to have the same mass function, then whatever conclusions apply to your random variable apply to mine and vice versa (even though they came from two different settings!)."

Remark 9.7. Since the mass function is nonnegative and adds up to 1, it defines a probability on the space of possible values D. So the random variable "pushes" the probability measure from the sample space Ω to a probability measure on the space of possible values D.

Read sections 1.5 and 3.1 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 9.1. 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 9.2. 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.

Exercise 9.3. Consider a sequence of five independent tosses of a fair coin. Let X be the number of times that a head is followed immediately by a tail. For example, if the outcome is $\omega = \text{HHTHT}$ 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 9.4. 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 9.5. 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 9.6. Let $\Omega = \{1, \ldots, 6\}^2 = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \ldots, 6\}\}$ 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 9.7. 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?

Exercises 1.16 to 1.19 on pages 31 and 32 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 10

We now study a few frequently encountered discrete random variables. As we learned in the previous lecture, to describe these random variables we just need to specify their probability mass functions.

1. The Bernoulli distribution

Suppose we perform an experiment once and the possible outcomes are either a "success" or a "failure". Let $p \in [0, 1]$ be the probability of success. This is called a *parameter*. So the state space is $\Omega = \{\text{success}, \text{failure}\}$. Let $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. The probability 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 random variable with this mass function is said to have a *Bernoulli distribution* with parameter p and we write $X \sim \text{Bernoulli}(p)$ or $X \sim \text{Ber}(p)$.

2. The binomial distribution

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 tossing n independent 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 $0, \ldots, n$. Let us find the probability mass function of X. That is, we seek to find $f(k) = P(X = k)$, where $k = 0, \ldots, n$. This is the probability of getting exactly k successes and $n-k$ failures. There are $\binom{n}{x}$ ways to choose which k trials are successes. Moreover, by independence, the probability of getting any specific combination of exactly k successes (e.g. first k trials were successes and the rest were failures, or first we got a success then two failures then $k-1$ successes then the rest were failures, etc) is the same number: $p^k(1-p)^{n-k}$. Therefore,

[0] Last modified on March 18, 2020 at 09:29:53 -06'00'

Figure 10.1. Jacob Bernoulli (also known as James or Jacques) (Dec 27, 1654 – Aug 16, 1705, Switzerland)

the probability of getting exactly k successes equals $p^k(1-p)^{n-k}$, added $\binom{n}{k}$ times. In other words, the probability mass function is

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

Note that $\sum_{k=0}^{n} f(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1$ by the binomial theorem.

A random variable with the above mass function is said to have a *binomial distribution* with parameters n and p and we write $X \sim \text{Binomial}(n, p)$ or $X \sim \text{Bin}(n, p)$. Note that $\text{Bin}(1, p) = \text{Ber}(p)$!

Example 10.1. Ten percent of a certain (large) population smoke. If we take a random sample with 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 $n = 5$ ["smoker"=1]. Therefore,

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

Alternatively, we can follow the longer route and write

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

Compute this yourself and check that you get the same answer as above.

The more natural thing is to sample without replacement. In that case, X is not exactly binomial. For example, if the population size is $1,000,000$ and there are $100,000$ of them that are smokers (10 percent) then the probability to get a smoker is 0.1. But if our first sampled person is indeed a smoker, then the probability our next pick is again a smoker is $99,999/999,999$ which is not exactly 0.1 but it is very close to that. So we see that although X is not exactly binomial, it is approximately binomial, as long as the sample size n is small compared to the population size.

The upshot is that if we are taking a sample from a very large population and wondering about the number of elements in our sample that satisfy a certain trait, then using a binomial

random variable is appropriate. The parameters of this binomial would then be $n =$ the sample size (5 in our example) and $p =$ the fraction in the population that has this trait (0.1 in the example).

Example 10.2. This evening I will be playing a card game with three other friends where each round consists of dealing 13 cards to each player out of a regular deck of 52 cards. In this game, the term acing a round means getting at least two aces at hand. We plan on playing 20 rounds. What is the probability I ace at least five rounds? To answer this question, let X be the number of rounds in which I ace. Then I want to calculate $P(X \geqslant 5)$. The possible values that the random variable X can take are the integers between 0 (I do not ace any of the rounds) and 20 (I ace all 20 rounds). It is also reasonable to assume that the contents of my hand are independent from round to round. So if we think of acing as a success, then X counts the number of successes in 20 independent trials. Hence, it is a binomial with parameters $n = 20$ and $p =$ the probability of acing a round. Now we need to compute p.

Presumably, the cards are well shuffled and come at a completely random order. That is, any ordering of the 52 cards is equally likely. If so, then the probability of me getting all four aces in my hand should not depend on the order in which the cards are dealt (whether 13 cards to each player in turn or one card per player until all players receive 13 cards, and whether I am dealt first or second or third or last). So we can assume I am the first to be dealt their hand and I am dealt all 13 cards right away. It is easier to compute $1-p$, i.e. the probability I get at most one ace in a given round. For this I compute the probability of not getting any ace and the probability of getting exactly one ace and I add them up.

To not get any ace I need to be dealt all 13 cards from the 48 cards that are not aces and so the probability of not getting any ace equals $\binom{48}{13}/\binom{52}{13}$. Similarly, to get exactly one ace I need to be dealt one of the four aces and then 12 cards from the 48 that are not aces. The probability of that is $4 \times \binom{48}{12}/\binom{52}{13}$. Thus,

$$p = 1 - \frac{\binom{48}{13}}{\binom{52}{13}} - \frac{4 \times \binom{48}{12}}{\binom{52}{13}} \simeq 0.2573.$$

Now I can compute the probability I ace in at least five rounds. It is equal to one minus the probability I ace in at most four rounds, i.e.

$$1 - \binom{20}{0}(0.7427)^{20} - \binom{20}{1}(0.2573)(0.7427)^{19} - \binom{20}{2}(0.2573)^2(0.7427)^{18} - \binom{20}{3}(0.2573)^3(0.7427)^{17} - \binom{20}{4}(0.2573)^4(0.7427)^{16} \simeq 0.6144.$$

3. The geometric distribution

Suppose we now keep running the trials until the first success. Another way to think about this is as follows. Fix a parameter $0 < p < 1$. 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.

Evidently, if n is an integer greater than or equal to one, then

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

Therefore, the probability mass function of X is given by

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

The fact that the above probability mass function adds up to one comes by applying equation (6.1) with $m = 0$ and $a = 1 - p$. (Check that!)

A random variable with this probability mass function is a said to be a geometric random variable with parameter p and we write $X \sim \text{Geometric}(p)$ or $X \sim \text{Geo}(p)$. This generalizes the situation we had in Exercise 6.2 where the coin was fair, i.e. $p = 1/2$, and the one in Exercise 6.2 where $p = 1/6$.

Example 10.3. Suppose in Example 10.2 I say I do not want to play 20 rounds. Instead, I will keep playing the game until I ace. What is the probability I will end up having to play exactly 20 rounds? Since rounds are independent the number of rounds I have to play until I ace is a geometric random variable with parameter $p =$ the probability I ace, which we computed to be approximately 0.2573. Playing exactly 20 rounds has then probability $(0.2573)(0.7427)^{19} \simeq 0.0009$.

Example 10.4. 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, for a random variable X the probability $P(X \geqslant n)$, when n is large, is called the *tail* of the random variable X. This is the probability of outliers, i.e. of X taking large values. In the case of $X \sim \text{Geometric}(p)$ saying $X \geqslant n$ is the same as saying that the first $n-1$ trials were failures. So

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

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

Example 10.5. In Example 10.3 the probability of ending up having to play at least five rounds is $(0.7427)^4 \simeq 0.3042$ while the probability of having to play at least 10 rounds is $(0.7427)^9 \simeq 0.0687$. The probability of having to play 20 rounds or more is $(0.7427)^{19} \simeq 0.0009$.

Here is now an interesting fact about geometric random variables.

Example 10.6. 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

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

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 heads, 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" or "a geometric random variable is memoryless".

4. 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. Alternatively, we are running independent trials and looking for the time when we get our r-th success.

X is a discrete random variable with possible values $r, r+1, r+2, \ldots$. When $r=1$, then X is Geometric(p). To compute its probability mass function note that $X=k$ means that the k-th toss gave heads and the first $k-1$ tosses have $r-1$ heads and the remaining $k-r$ tosses were heads. Again, any specific combination of heads and tails that satisfies these conditions has probability $p^r(1-p)^{k-r}$ (r heads and $k-r$ tails). And we have to add this number once for every such combination. There are as many such combinations as there are ways to specify the positions of $r-1$ heads in $k-1$ tosses. That is $\binom{k-1}{r-1}$. Hence,

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

One still needs to prove that this adds up to one, to rule out the possibility that we toss the coin indefinitely and never get the r-th success. For now we will just take it for granted, but one can see it at an intuitive level like this: all one has to prove is that there is probability 100% that we will get the r-th success at some point, i.e. that it will not happen that we keep tossing the coin and never get r heads. We already proved this fact in the case $r=1$ (waiting for one success) and waiting for r successes is like repeating the experiment of "waiting for one success" r times. Each time the experiment will terminate (we will get the one success we are waiting for) and so we will eventually get the r successes. All this is just an intuition and needs proof. We will see a proof of this later, in Lecture 18. (The solution of Exercise 12.3 shows has the case $r=2$.)

A random variable with this mass function is said to have a *negative binomial distribution* with parameters r and p. Note that a negative binomial with parameters 1 and p is the same as a geometric with parameter p.

5. 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 ($X \sim \text{Poisson}(\lambda)$ or $X \sim \text{Poi}(\lambda)$) if its probability mass function is

$$f(k) = \begin{cases} \dfrac{e^{-\lambda}\lambda^k}{k!} & \text{if } k = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases} \tag{10.1}$$


Figure 10.2. Left: Siméon Denis Poisson (Jun 21, 1781 – Apr 25, 1840, France). Right: Sir Brook Taylor (Aug 18, 1685 – Nov 30, 1731, England)

In order to make sure that this makes sense, it suffices to prove that $\sum_{k=0}^{\infty} f(k) = 1$, 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.

5.1. Law of rare events. Suppose there is a very large number n of people in the supermarket and at some point in time they all, independently of each other, make a decision on whether or not to go check out. Suppose the probability any given person decides to check out is a very small number $p \in (0,1)$. Let X be the number of people at the check out register. Then $X \sim \text{Binomial}(n,p)$.

Poisson's "law of rare events" states that if n (the number of people in the supermarket) is large and p is small, but np is "not too large nor too small" then the binomial random variable with parameters n and p can be approximated by a Poisson random variable with parameter np.

So if n is large and p is small but np is not too large nor too small, then instead of saying X is binomial with parameters n and p we can say that X is Poisson with parameter np.

Going back to the supermarket example we see that the number of people in line for check out can be thought of as a Poisson random variable. This explains why Poisson random variables are used to model the length of a waiting list.

All that is good, but in order to make Poisson's law precise one still has to explain what is meant by "not too large nor too small." The precise statement is the following: let $\lambda > 0$ be some fixed real number. Consider a random variable $X \sim \text{Binomial}(n, \lambda/n)$. So here, $p = \lambda/n$ and $np = \lambda$. As n grows large p becomes small, but np remains equal to λ the whole time (so it is not too large nor too small - in fact, it is held fixed here!). We will next show that the probability mass function of X converges, as $n \to \infty$, to the probability mass function of a Poisson(λ) random variable.

For a given integer k between 0 and n the probability mass function of X equals

$$\begin{aligned} f(k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

As we take $n\to\infty$ we get that $\lambda/n\to 0$ and so $(1-\lambda/n)^{-k}\to 1$. Also,

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} = \frac{n}{n}\times\frac{n-1}{n}\times\cdots\times\frac{n-k+1}{n}\underset{n\to\infty}{\longrightarrow} 1.$$

And lastly, we compute the limit of $(1-\lambda/n)^n$. Note that this is not the same as when the power was $-k$ and hence not changing with n. To get the above limit one can use the fact that the exponential function is continuous and hence

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \exp\Big\{\lim_{n\to\infty}\ln\Big(1-\frac{\lambda}{n}\Big)^n\Big\} = \exp\Big\{\lim_{n\to\infty} n\ln\Big(1-\frac{\lambda}{n}\Big)\Big\}.$$

To get the last limit change variables $h = 1/n$ which now goes to 0 and then the limit becomes

$$\lim_{n\to\infty} n\ln\Big(1-\frac{\lambda}{n}\Big) = \lim_{h\to 0}\frac{\ln(1-\lambda h)}{h}.$$

Plugging $h = 0$ gives $0/0$. Applying de l'Hôpital's rule we get that this limit equals

$$\lim_{h\to 0}\frac{\frac{d\ln(1-\lambda h)}{dh}}{1} = \lim_{h\to 0}\frac{-\lambda}{1-\lambda h} = -\lambda.$$

Going back to the above we have

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \exp\Big\{\lim_{n\to\infty} n\ln\Big(1-\frac{\lambda}{n}\Big)\Big\} = e^{-\lambda}.$$

Putting everything together we see that as n (the population) grows large and k is held fixed, the probability mass function $f(k)$ (the probability the queue length is k) approaches the limit

$$\frac{\lambda^k}{k!}e^{-\lambda},$$

which is precisely the formula for the probability mass function of a Poisson random variable.

Example 10.7. Here is a natural situation where a Poisson distribution is used (other than to model the length of a waiting list). We would like to model the number of people with a rare disease in a large sample from an even larger population. But we do not know (or care to know) the size of the sample nor the size of the population. Then we can use a Poisson random variable to do the job. We just need to estimate its parameter and we would have a probability model that approximates the situation pretty well. Here are some more details to explain.

Suppose we are studying a large population of people. Suppose that a very small fraction p of the population has a rare disease. Suppose we take a large sample n and we are wondering about the number of people in the sample that do have the disease. More specifically, say we are wondering about the probability at most 50 people have the disease. In reality, we do not know the size of the population nor the size of the sample. We only know that the sample size is very large and the population size is much larger. But to be concrete let us say we are considering a

city with a $1,000,000$ inhabitants of whom 0.5% have a rare disease and we are taking a sample of $1,000$ people.

Since the sample is small relative to the population, the number of people in with the disease, in our sample, is a binomial random variable with parameters $n = 1,000$ and $p = 0.005$. But since n is quite large and $np = 5$ is not too large and not too small, we can instead say that the number of people in our sample is a Poisson random variable with parameter $\lambda = np = 5$. So if we are asked e.g. about the probability we get at most 50 people with this disease, then instead of computing

$$\sum_{k=0}^{50} \binom{1,000}{k} (0.005)^k (0.995)^{1,000-k},$$

which is what we would have to do if we use a binomial and for which we need to know that the sample size is $1,000$, we compute

$$\sum_{k=0}^{50} e^{-5} \frac{5^k}{k!},$$

which is what we would do if we use a Poisson and for which we do not need to know the sample size.

Can you write a code to compute both sums and compare?!

Example 10.8. A soccer player scores at least one goal in about half of the games he plays. We want to estimate the percentage of games in which he scores a hat-trick (exactly three goals). To build a simplified mathematical model we will assume that during each game the player is involved in a large number of independent attempts to score and that in each attempt he has a certain not too large chance to score. This last assumption is probably reasonable. The first two (the large number of attempts and more so, the independence of the attempts) are questionable. But let us assume them for simplification. Then, this is like running a large number of independent trials where success in each trial has a small probability. So the number of successes (goals) is a bernoulli with a large parameter n and a small parameter p and can hence be approximated by a Poisson with some parameter λ. To estimate λ we recall that we are told that the player scores at least one goal in half his games. This means that the probability the player scores at least one goal on a given game is $1/2$. So $P(X \geqslant 1) = 1/2$. But if $X \sim \text{Poisson}(\lambda)$, then we know that

$$P(X \geqslant 1) = 1 - P(X = 0) = 1 - e^{-\lambda}.$$

So $1 - e^{-\lambda} = 1/2$ and solving for λ gives $\lambda = \ln 2 \simeq 0.6931$. But the the probability of a hat-trick is

$$P(X = 3) = e^{-0.6931} \frac{(0.6931)^3}{3!} \simeq 0.0278 = 2.78\%.$$

Go through the examples in the lecture and make sure you understand them well.

Read sections 2.4, 2.5, and 4.4 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 10.1. Some day, 10,000 cars are traveling 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.

Exercises 2.20, 2.21, 2.22, 2.23, 2.28(b), 2.28(c), 4.9, and 4.11 on pages 74, 75, and 172 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 11

1. Continuous random variables

So far we learned how to describe random variables that have countably many possible values. But what about random variables that can take uncountably many values? For example, how do we describe picking a random number from the interval $[0, 1]$? We learned in Section 1.2 how to do this for equally likely outcomes. In the language of random variables, this is expressed by saying that the probability that X takes a value in say the interval $[a, b]$ (i.e. our random number turned out to be between a and b) is equal to $b - a$. In mathematical notation

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

And more generally,

$$P(X \in A) = |A|,$$

where $|A|$ is the "length" of A. But how do we model a situation where the random variable does not take all its possible value equally likely?

As we explained in Lecture 9, describing a random variable can be done by specifying its distribution. That is, specifying $P(X \in A)$ (the probability X is in A) for all events A. And just like in Section 1.2 we saw that assigning probabilities to individual outcomes and then deducing probabilities of events from that would not work in an uncountable setting, here too working with a mass function (i.e. with $P(X = x)$) will not do the job. Instead, we need to define $P(X \in A)$ directly, in a way that respects Rules 1–3 from Lecture 2. Here is one way to do this, which uses calculus to mimic the way a mass function works.

Suppose we are given a function $f : \mathbb{R} \to [0, \infty)$ such that

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

Note how the function takes only nonnegative values. Set

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

This says that to calculate the probability that X is in A we integrate the function f over the set A.

[0] Last modified on May 12, 2020 at 13:06:20 -06'00'

Notice how all three rules of probability are satisfied. Indeed, since we are integrating a nonnegative function we will always have $P(X \in A) \geqslant 0$. And since the function integrates to one on the whole real line we have $P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\,dx = 1$. Finally, if A and B are disjoint, then integrating over $A \cup B$ amounts to integrating over each separately and then adding them up and so Rule 3 holds:

$$P(X \in A \cup B) = \int_{A\cup B} f(x)\,dx = \int_A f(x)\,dx + \int_B f(x)\,dx = P(X \in A) + P(X \in B).$$

A special event that we often encounter is the one that says X falls in some given interval: $\{X \in [a, b]\}$ or $\{a \leqslant X \leqslant b\}$. The probability of this is computed as follows:

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

Note that we also have

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

And as a special case,

$$P(X = a) = P(a \leqslant X \leqslant a) = \int_a^a f(x)\,dx = 0.$$

This is why a mass function would not be of any help here. It would be identically zero! However, one can think of $f(x)$ as the "likelihood" of X taking values near x. This is because if ε is some very small positive number, then for x in the interval $(a - \varepsilon, a + \varepsilon)$ the value of $f(x)$ is close to that of $f(a)$ and so

$$P\{X \in (a - \varepsilon, a + \varepsilon)\} = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx \simeq f(a) \int_{a-\varepsilon}^{a+\varepsilon} dx = 2\varepsilon f(a).$$

So we see that $2\varepsilon f(a)$ is approximately the probability of X being within distance ε of a and that the larger $f(a)$ is the more likely X to be near a and the smaller $f(a)$ is the less likely X is to be near a.

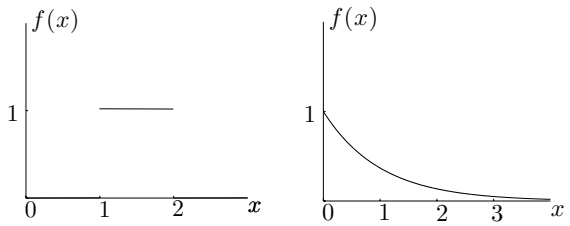
The function f is called the *probability distribution function* (pdf) of the random variable X and the random variable X is said to be a *continuous random variable*. The pdf uniquely characterizes the continuous random variable much like the mass function characterizes a discrete random variable.

Example 11.1. Say X is a continuous random variable with probability density function

$$f(x) = \frac{1}{4x^2} \text{ if } |x| > 1/2 \text{ and } f(x) = 0 \text{ otherwise.}$$

Let us first double check this is a legitimate pdf: it is clearly non-negative and we have

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{-1/2} \frac{1}{4x^2}\,dx + \int_{1/2}^{\infty} \frac{1}{4x^2}\,dx = 2\int_{1/2}^{\infty} \frac{1}{4x^2}\,dx = 2 \cdot \frac{-1}{4x}\Big|_{1/2}^{\infty} = 2 \cdot \frac{1}{2} = 1.$$



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

Figure 11.1. pdf for certain continuous distributions

Now that we are sure this is indeed a pdf, we can compute things like

$$\begin{aligned} P\{X < -1 \text{ or } 0 < X \leqslant 1 \text{ or } X \geqslant 2\} &= P\{X < -1\} + P\{0 < X \leqslant 1\} + P\{X \geqslant 2\} \\ &= \int_{-\infty}^{-1} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{2}^{\infty} f(x)\,dx \\ &= \int_{-\infty}^{-1} \frac{1}{4x^2}\,dx + \int_{1/2}^{1} \frac{1}{4x^2}\,dx + \int_{2}^{\infty} \frac{1}{4x^2}\,dx \\ &= \frac{5}{8}. \end{aligned}$$

Notice how the second integral goes from $1/2$ to 1 instead of from 0 to 1. This is because $f(x) = 0$ for $|x| \leqslant 1/2$.

2. Standard examples

We will now go through some well-known examples of continuous random variables. As was the case for discrete random variables, a random variable X is in general completely (statistically) determined by knowing the probabilities $P(a \leqslant X \leqslant b)$ for all $a < b$. Since for a continuous random variable this probability is given by integrating the probability density function over the interval $[a, b]$ we see that a continuous random variable is completely determined by its probability density function. In other words, if two different experiments yield two random variables that happen to have the exact same pdf, then there is no statistical test that can distinguish the data from one experiment from the data from the other experiment. Therefore, in all the examples that follow we will define the random variables by giving their probability density function.

Example 11.2 (Uniform random variable)**.** If $a < b$ are fixed, then the uniform density function 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 11.1(a). A random variable with this density takes any value in $[a, b]$ "equally likely" and has 0 likelihood of taking values outside $[a, b]$. We write $X \sim \text{Uniform}(a, b)$ to say that X has this pdf.

If $a \leqslant c \leqslant d \leqslant b$, then we get the familiar

$$P\{c \leqslant X \leqslant d\} = \int_{c}^{d} \frac{1}{b-a}\,dx = \frac{d-c}{b-a}.$$

Example 11.3 (Exponential random variable)**.** 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 pdf, and is called the *exponential probability density with parameter* λ. See Figure 11.11.1(b). 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 exponential random variables 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 got served by time x, that is $P\{X \leqslant x\}$, is the same as the probability that there was at least one toss that came up heads, among the first nx. This is equal to one minus the probability all nx tosses landed tails, i.e. $1 - (1 - \lambda/n)^{nx}$. By a similar calculation as in Section 5.1 we see that this converges to $1 - e^{-\lambda x}$. So in the limit, the random variable that we get satisfies

$$P(X \leqslant x) = 1 - e^{-\lambda x}$$

for all $x > 0$. But if this random variable has pdf f, then we would have

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

Furthermore, since X is the amount of time it took to be served it is non-negative and so $f(x) = 0$ for $x < 0$. Therefore, we have established that

$$\int_{0}^{x} f(y)\, dy = 1 - e^{-\lambda x} \quad \text{for all } x > 0.$$

This means that $1 - e^{-\lambda x}$ is an anti-derivative of f and so $f(x)$ is the derivative of this function and equals

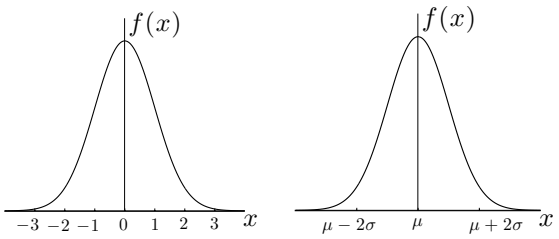
$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x > 0.$$

This is indeed the pdf of an exponentially distributed random variable with parameter λ.

Example 11.4. Just as we have seen for a geometric random variable, an exponential random variable does not recall history; see Example 10.6. 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?



(a) Standard Normal (b) Normal(μ,σ^2)

Figure 11.2. pdf of normal distributions.

3. The Normal distribution: the bell curve

In the next two examples, we present one of the most important random variables.

Example 11.5 (Standard normal density)**.** I claim that

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

defines a density function; see Figure 11.2(a). 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?]

Note that if we are interested in the probability X is in an interval, then we can compute it like this:

$$P(a < X < b) = \int_a^b \phi(x)\, dx = \int_{-\infty}^b \phi(x)\, dx - \int_{-\infty}^a \phi(x)\, dx.$$

So if someone gives us a formula for the function

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

then we can use it to compute probabilities for a standard normal random variable to be in an interval:

$$P(a < X < b) = \Phi(b) - \Phi(a).$$

So is there a formula for Φ?

Figure 11.3. Johann Carl Friedrich Gauss (Apr 30, 1777 – Feb 23, 1855, Germany)

Note that Φ is has a certain symmetry: using the change of variables $y = -x$ and the fact that $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ integrates to 1 we have

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

Plugging in $z = 0$ we get $\Phi(0) = 1/2$.

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 exactly. for any value of z other than $z = 0$. 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 B. (Note that the table only has $z \geqslant 0$. The reason is that one can use the above symmetry to compute Φ at negative z values.)

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 11.6 (Normal or Gaussian density (Gauss, 1809))**.** Given two numbers $-\infty < \mu < \infty$ and $\sigma > 0$, a random variable X is said to have the Normal density (X ~Normal(μ,σ^2)) if it has pdf

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

see Figure 11.2(b). Using a change of variables, one can relate this distribution to the standard

normal one, denoted N(0,1). Indeed, for all $-\infty < a \leqslant b < \infty$,

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\,dx \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \qquad [z = (x-\mu)/\sigma] \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z)\,dz \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).\end{aligned} \tag{11.1}$$

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*.

(11.1) can be interpreted as follows. If X is a Normal(μ,σ^2) random variable, then the probability it is between a and b can be computed as follows: first "standardize" a and b, i.e. replace them with $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$, respectively, and then compute the probability a standard Normal(0,1) random variable is between these two standardized values. This last bit is done using the function Φ as explained above.

For example, the probability a random variable with distribution $N(-3,8)$ is between -3.5 and 5.2 is the same as the probability a standard normal $N(0,1)$ is between $(-3.5-(-3))/\sqrt{8} \simeq -0.18$ and $(5.2-(-3))/\sqrt{8} \simeq 2.9$. The table tells us that the probability a standard normal is below 2.9 is about 0.9981. The table does not give the probability a standard normal is below -0.18. But by the symmetry of the bell curve, this probability is the same as that of a standard normal being above 0.18, which is one minus the probability it is below 0.18, which the table says is approximately 0.5714. So the probability the standard normal is below -0.18 is approximately $1 - 0.5714 = 0.4286$. Therefore the probability the $N(-3,8)$ random variable is between -3.5 and 5.2 is approximately $0.9981 - 0.4286 = 0.5695$. is below -0.18.

Read sections 3.1, 3.5, and 4.5 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 11.1. Let X be a random variable with probability density function given by

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

What is the value of c?

Exercise 11.2. Let X be a random variable with probability density function given by

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

What is the value of c?

Exercise 11.3. Let X be a random variable with probability density function

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

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 11.4. 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?

Exercise 11.5. 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?

Exercises 3.1, 3.2, 3.3, 3.4, 3.17, 3.18, and 4.13 on pages 126, 128, 129, and 173 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 12

1. Mathematical Expectation: Discrete random variables

Suppose we are given a list of ten measurements and are asked about their mean or average. We would add up all the numbers and divide by the number of data points. For example, say the measurements are 1, 5.5, 2.3, 6.7, 5.5, 1, 5.5, 5.5, 1, 2.3. Then the average would be given by

$$\frac{1+5.5+2.3+6.7+5.5+1+5.5+5.5+1+2.3}{10},$$

which we can rearrange as

$$1\times\frac{3}{10}+5.5\times\frac{4}{10}+2.3\times\frac{2}{10}+6.7\times\frac{1}{10}.$$

Note that this amounts to an average over all the values we observed, weighted by the proportion of time each value occurred.

Recalling that in a probability model the probability of an outcome is supposed to replace the "proportion of time that outcome occurs" we are led to the following notion: 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 as the average of the possible values of X, weighted by their corresponding probabilities:

$$E[X]=\sum_x x\,f(x). \tag{12.1}$$

Example 12.1. 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 12.2. We roll a loaded die that 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. 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

[0] Last modified on February 27, 2020 at 09:53:54 -07'00'

Adding the elements in the last row gives

$$E[X] = -0.8 + 0.2 + 0.8 = 0.2 \text{ (20 cents).}$$

Intuitively, if we play 1000 times we expect to win about \$200. Thus, 20 cents is a fair participation fee for each attempt. We will see later how the math confirms this intuition.

Example 12.3. 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 at least a dollar and 75 cents each time!

Example 12.4 (Bernoulli and Binomial)**.** 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.

If X has infinitely-many possible values, then the sum in (12.1) must be defined. If X only takes nonnegative values, then (12.1) is a sum of nonnegative numbers and is thus always defined [though could be ∞]. Similarly, if X only takes nonpositive values, then the sum is that of nonpositive numbers and is always defined [though could be $-\infty$].

Example 12.5 (Geometric and Negative Binomial)**.** 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 12.6 (Poisson)**.** Suppose X is a Poisson random variable with parameter λ. Then

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

Some explanation: the first equality is the definition of expected value, the second one is because when $k = 0$ the corresponding term is 0 and so this term can be ignored, the fourth equality is just a change of the index of summation from $k = 1, 2, \ldots$ to $j = k - 1$ which now goes from 0 and on. The last equality comes by recognizing that the terms being summed at the end of the

second line are the values of the mass function of the Poisson random variable and hence add up to one.

The upshot of this example is that the parameter λ of the Poisson random variable is exactly the mean of the random variable. So for example when modeling the length of a waiting line with a Poisson random variable, the parameter λ is the average length of the line.

Example 12.7 (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 12.8. 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.

Read the first part of section 3.3 (the one up to Expectation of continuous random variables) in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 12.1. 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 12.2. In the game *Wheel of Fortune*, you have 52 possible outcomes: one "0", one "00", two "20", four "10", seven "5", fifteen "2" and twenty-two "1". If you bet \$1 on some number, you receive this amount of money if the wheel stops on this number. (In particular, you do NOT lose the \$1 you bet.) If the wheel stops at a different number, you lose the \$1 you bet. If you bet \$1 on "0" or "00", you receive \$40 if the wheel stops on this number (and in this case you do not lose the \$1 you bet). For example, say you bet \$1 on 10. If the wheel stops on 10, your profit is \$10. If it stops on something other than 10, your profit is -\$1 because you lose the \$1 you bet.

(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?

Remark: Try to redo the exercise with the assumption that you always lose the \$1 you bet. See how part (b) changes drastically, with just this small change in the rules of the game!

Exercise 12.3. Let X be a Geometric random variable with parameter $p \in [0, 1]$. Compute $E[X]$.

Exercises 3.8(a) and 3.12 on pages 127 and 128 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 13

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

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx. \tag{13.1}$$

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 13.1 (Uniform)**.** Suppose X is uniform on (a, b). Then the pdf equals the constant $\frac{1}{b-a}$ when $a < x < b$ and it is 0 otherwise. Therefore

$$E[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{1}{2}\frac{(b-a)(b+a)}{b-a} = \frac{b+a}{2}.$$

Example 13.2 (Exponential)**.** If X is Exponential(λ), then the pdf equals $\lambda e^{-\lambda x}$ when $x > 0$ and is 0 otherwise. Therefore,

$$E[X] = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx.$$

Since $\lambda e^{-\lambda x}$ is the derivative of $-e^{-\lambda x}$ we can use integration by parts we write

$$\int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx = -x e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\, dx = (0-0) + \frac{1}{\lambda}.$$

Therefore, the mean of an exponential random variable with parameter λ is equal to $E[X] = 1/\lambda$.

[0] Last modified on April 22, 2020 at 23:11:49 -06'00'

Example 13.3 (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,

$$\begin{aligned} E[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-z^2/2} \, dz \qquad (z = (x-\mu)/\sigma) \\ &= \mu \underbrace{\int_{-\infty}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz}_{1} + \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z e^{-z^2/2} \, dz}_{0, \text{ by symmetry}} \\ &= \mu. \end{aligned}$$

So the parameter μ is in fact the mean of the normal random variable.

Figure 13.1. Baron Augustin-Louis Cauchy (Aug 21, 1789 – May 23, 1857, France)

The next example introduces a random variable that does not have a well-defined mean.

Example 13.4 (The Cauchy density (Cauchy, 1827))**.** Define for all real numbers x,

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

Because

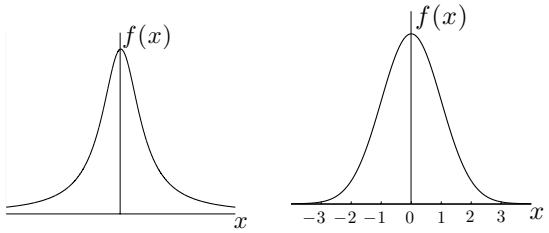
$$\frac{d}{dx} \arctan x = \frac{1}{1+x^2},$$

we have

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

Hence, f is a probability density; see Figure 13.2.

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



(a) Cauchy (b) Standard Normal

Figure 13.2. A comparison of the Cauchy pdf and the standard normal pdf.

Does this distribution have a mean? To answer this, we want to compute

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

For this, let us first compute the integral from 0 to ∞ and use the change of variables $u = 1 + x^2$. Then $du = 2x\, dx$ and so

$$\int_0^{\infty} \frac{x}{\pi(1+x^2)}\, dx = \frac{1}{2\pi}\int_1^{\infty} \frac{1}{u}\, du = \frac{1}{2\pi}\log u\Big|_1^{\infty} = \infty.$$

Similarly we can find that the integral from $-\infty$ to 0 equals $-\infty$. Therefore, $E[X]$ does not exist.

2. The mean of a function of a random variable

Suppose we have a random variable X and we are interested in computing the mean of X^2. One way to go about this would be by treating $Y = X^2$ as a new random variable, find its probability mass function (if X is discrete) or probability density function (if X is continuous) and then compute the mean as we learned in the previous sections. But there is a faster and more direct way.

Theorem 13.5. *If X is a discrete random variable with mass function $f(x)$ and if g is a function defined on the set of possible values of X, then*

$$E[g(X)] = \sum_x g(x)f(x).$$

Proof. Let $Y = g(X)$. This is a new random variable. If f is the mass function of X then the mass function of Y is given by

$$f_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P(X = x) = \sum_{x:g(x)=y} f(x).$$

Then, by definition

$$\begin{aligned} E[g(X)] = E[Y] &= \sum_y y\, f_Y(y) = \sum_y y \sum_{x:g(x)=y} f(x) \\ &= \sum_y \sum_{x:g(x)=y} y\, f(x) = \sum_y \sum_{x:g(x)=y} g(x)f(x) = \sum_x g(x)f(x), \end{aligned}$$

as desired. □

The above has a version for continuous random variables. But the proof of that version requires tools beyond the scope of this class. Therefore, we take the following for granted.

Theorem 13.6. *If X is a continuous random variable with density function $f(x)$ and if g is a function defined on the set of possible values of X, then*

$$E[g(X)] = \int g(x)f(x)\,dx.$$

Here are some consequences of the above theorems.

Theorem 13.7. *Let X be a random variable. Let a be any (nonrandom) number and let g and h be any two functions. Then:*

(1) $E[ag(X)] = aE[g(X)]$; *(constants can come out of the expectation)*
(2) $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$; *(expectation is additive)*
(3) $E[b] = b$; *(the mean of a constant is the constant itself)*
(4) *If* $P\{X \geqslant 0\} = 1$ *and* $E[X] = 0$, *then* $P\{X = 0\} = 1$; *(If a nonnegative random variable has mean zero, then the random variable itself is zero with probability one.)*

Proof. We show the proofs in the discrete case. The proofs in the continuous case are similar. We just need to replace sums by integrals.

For the first claim we can write

$$E[ag(X)] = \sum ag(x)f(x) = a\sum g(x)f(x) = aE[g(X)].$$

For the second claim write

$$\begin{aligned} E[g(X) + h(X)] &= \sum (g(x) + h(x))f(x) = \sum (g(x)f(x) + h(x)f(x)) \\ &= \sum g(x)f(x) + \sum h(x)f(x) = E[g(X)] + E[h(X)]. \end{aligned}$$

For the third claim the function being considered takes everything to the value b. So

$$E[b] = \sum bf(x) = b\sum f(x) = b.$$

(Recall that $\sum f(x) = 1$.) For the fourth claim, if $P\{X \geqslant 0\} = 1$ it means that X does not take negative values. So we can write

$$E[X] = \sum_{x \geqslant 0} xf(x).$$

But now the terms under the sum are all nonnegative (because $f(x)$ is always nonnegative and now we are adding only over nonnegative x). The only way for them to add up to 0 is for all the terms to equal 0. But if $x > 0$ then $xf(x) = 0$ implies $f(x) = 0$. So the only possible value the random variable can take is $x = 0$. That is: the random variable is really not random. It equals the constant 0. □

A special function that we will care about in the next lecture is $g(x) = x^2$. The above says that for a discrete random variable with mass function f we have

$$E[X^2] = \sum x^2 f(x)$$

and for a continuous random variable with pdf f we have

$$E[X^2] = \int_{-\infty}^{\infty} x^2\, f(x)\,dx.$$

Let us compute these quantities for a few examples.

Example 13.8. 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 $\text{Binomial}(n-2, p)$ 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 13.7 to find that

$$E[X^2] = E[X(X-1)] + E[X] = n(n-1)p^2 + np = (np)^2 + np(1-p).$$

Example 13.9. Suppose $X \sim \text{Geometric}(p)$ distribution. We have seen already that $E[X] = 1/p$ (Example 12.5). Let us find a new computation for this fact, and then go on and compute also $E[X^2]$.

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

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 since this is beyond the scope of this course.

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

$$E[X^2] = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2}.$$

Example 13.10. Suppose $X \sim \text{Poisson}(\lambda)$. We saw earlier 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. Indeed, change variables ($j = k-2$) to get a sum from $j = 0$ to ∞ and recognize the j-th term $e^{-\lambda}\lambda^j/j!$ as the probability that $\text{Poisson}(\lambda) = j$. 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,

$$E[X^2] = \lambda^2 + \lambda.$$

Example 13.11. Suppose $X \sim \text{Exponential}(\lambda)$. Then

$$E[X^2] = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}\, dx.$$

To compute this integral we use integration by parts. Namely, let $u = x^2$ and $v = -e^{-\lambda x}$. Then what we have is the integral of uv' and the product rule says that

$$(uv)' = uv' + u'v.$$

So

$$\int_0^{\infty} uv'\, dx = \int_0^{\infty} ((uv)' - u'v)\, dx = uv\Big|_0^{\infty} - \int_0^{\infty} u'v\, dx.$$

Plugging in the actual functions u and v we have $u' = 2x$ and so

$$\int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}\, dx = -2xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} (2x)e^{-\lambda x}\, dx.$$

The first term vanishes at $x = 0$ and also goes to 0 as $x \to \infty$. So it is gone. In the second term can be rewritten as

$$\frac{2}{\lambda}\int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx.$$

We have already calculated this integral. It is the mean of the exponential random variable and we already found that it equals $1/\lambda$. So we now have

$$E[X^2] = \frac{2}{\lambda^2}.$$

Example 13.12. Let $X \sim N(0,1)$. We have seen that $E[X] = 0$. What is $E[X^2]$? We need to compute

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx.$$

We use integration by parts. Letting $u = x$ and $v = -e^{-x^2/2}$ we have $uv' = x^2 e^{-x^2/2}$ and since $\int uv'\, dx = uv - \int vu'\, dx$ we get

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = -\frac{1}{\sqrt{2\pi}} x e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx.$$

The first term is 0 and the second is 1 (why?). Thus, $E[X^2] = 1$.

3. Variance

Suppose X is a random variable with mean $E[X] = \mu$. How random is this random variable? In other words, we want to measure how far "on average" X is from its mean. Recall the formula for the distance between two points in Euclidean space: the distance between $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$ is

$$\sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}.$$

Inspired by this analogy we define the *variance* of a random variable X with mean μ to be

$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right].$$

Thus:

(1) We predict the as-yet-unseen value of X by the nonrandom number $\mu = E[X]$ (its average value);

(2) $\mathrm{Var}(X)$ is the expected squared-error in this prediction. Note that $\mathrm{Var}(X)$ is also a nonrandom number.

The variance measures the amount of variation in the random variable. Note how if the random variable is not really random, then it equals its mean the whole time and so the variance then is zero. In fact, the other direction of this assertion is also true: if a random variable has zero variance then it is not random. Here is the mathematical statement and its proof.

Theorem 13.13. *If* $\mathrm{Var}(X) = 0$, *then* X *is a nonrandom constant.*

Proof. By Theorem 13.7, $0 = \mathrm{Var}(X) = E[(X - \mu)^2]$ implies that $P\{(X - \mu)^2 = 0\} = 1$. (The random variable $(X - \mu)^2$ is nonnegative and has mean 0.) But this means that $X = \mu$ with probability 100%. Recall that μ is not random. So we have shown that X equals a nonrandom number. □

The square root of the variance is the quantity that plays the role of the distance of the random variable X to its mean. (Recall how the formula for the Euclidean distance has the square root of a sum of squares.) The square root of the variance is called the *standard deviation*.

Here are some useful (and natural) properties of the variance.

Theorem 13.14. *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) $\mathrm{Var}(X + a) = \mathrm{Var}(X)$.

(1) is obvious (as the variance is the mean of a square). (2) says that nonrandom quantities have no variation, a fact we have already observed. (3) and (4) come by direct inspection of the definition of the variance. (3) says that changing units by a factor of a scales the variance like a^2. (4) says that shifting all the data by a nonrandom amount does not change the amount of variation in it.

Let us now compute the variance of a few random variables. But first, here is another useful way to write the variance

$$\begin{aligned} \mathrm{E}\left[(X-\mu)^2\right] &= \mathrm{E}\left[X^2 - 2X\mu + \mu^2\right] = \mathrm{E}[X^2] - 2\mu\mathrm{E}[X] + \mu^2 \\ &= \mathrm{E}[X^2] - 2\mu^2 + \mu^2 = \mathrm{E}[X^2] - \mu^2. \end{aligned}$$

Example 13.15 (Bernoulli and Binomial)**.** Suppose $X \sim \mathrm{Bernoulli}(p)$. Then, $X^2 = X$ (because X is either 0 or 1) and so $\mathrm{E}[X^2] = \mathrm{E}[X] = p$. But then, $\mathrm{Var}(X) = p - p^2 = p(1-p)$. More generally, if $X = \mathrm{Binomial}(n, p)$, then we have seen that $\mathrm{E}[X] = np$ and $\mathrm{E}[X^2] = (np)^2 + np(1-p)$. Therefore, $\mathrm{Var}(X) = np(1-p)$.

Example 13.16 (Geometric)**.** Suppose $X \sim \mathrm{Geometric}(p)$ distribution. We have seen in Example 13.9 that $\mathrm{E}[X] = 1/p$ and $\mathrm{E}[X^2] = \frac{2-p}{p^2}$. Consequently,

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Example 13.17 (Poisson)**.** We have already seen that if $X \sim \mathrm{Poisson}(\lambda)$, then $\mathrm{E}[X] = \lambda$ and $\mathrm{E}[X^2] = \lambda^2 + \lambda$. Therefore, $\mathrm{Var}(X) = \lambda$.

Example 13.18 (Uniform)**.** If X is $\mathrm{Uniform}(a, b)$, then $\mathrm{E}[X] = \frac{a+b}{2}$ and

$$\mathrm{E}[X^2] = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{b^2 + ab + a^2}{3}.$$

In particular, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Example 13.19 (Normal)**.** We can now compute the variance of a $\mathrm{N}(\mu, \sigma^2)$ random variable. Recall that we already found out that the parameter μ is actually the mean of the random variable. Therefore,

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

Use the change of variable $z = (x-\mu)/\sigma$ to get $dx = \sigma dz$ 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 equality we applied the computation from Example 13.12.

So the parameter σ^2 is actually the variance of the random variable. This is why one usually says that X is a normal random variable with mean μ and variance σ^2. And σ is in fact the standard deviation of the random variable.

Read the rest of section 3.3 and section 3.4 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 13.1. Let X be a uniform random variable on the interval $[a, b]$. Compute $E[X]$ and $E[X^2]$.

Exercise 13.2. 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 13.3. 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.

Exercises 3.8(b), 3.9, 3.10, 3.11, 3.14, 3.15, 3.16, 3.21, 3.22, 3.24, 3.31(all except (e) and (h)), 3.32(all except (b)), and 4.14 on pages 127, 128, 129, 131, and 173 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 14

1. The binomial distribution, the golden theorem, and a normal approximation

What makes probability theory useful (and elegant) is that once the sample size (or number of trials, or number of degrees of freedom, etc) increases, surprising and rather universal patterns appear out of the mess and chaos caused by the randomness! This section will touch on two such situations.

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 better be a consequence of the probability models we have been learning about. We will later prove that this is indeed a fact. Here is the mathematical statement.

Theorem 14.1 (Bernoulli's golden theorem a.k.a. the law of large numbers, 1713)**.** *Suppose* $0 \leqslant p \leqslant 1$ *is fixed. Then, with high probability, as* $n \to \infty$,

$$\frac{\textit{Number of heads}}{n} \approx p.$$

A bit more precisely: for any margin of error $\varepsilon > 0$

$$P\left\{\left|\frac{\textit{Number of heads}}{n} - p\right| < \varepsilon\right\} \underset{n\to\infty}{\longrightarrow} 1.$$

In words: In a large sample (n large), the probability is nearly one that the percentage in the sample (number of heads divided by n) is quite close to the percentage in the population (p); i.e. with high probability, random sampling works well for large sample sizes!

Next, a natural question comes to mind: for some given integers a and 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:

[0] Last modified on October 04, 2022 at 13:09:34 -06'00'

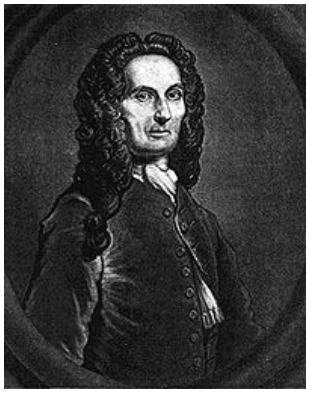



Figure 14.1. Left: Abraham de Moivre (May 26, 1667 – Nov 27, 1754, France). Right: Pierre-Simon, marquis de Laplace (Mar 23, 1749 – Mar 5, 1827, France).

Theorem 14.2 (The De Moivre–Laplace central limit theorem (CLT), 1733)**.** *Suppose* $0 < p < 1$ *is fixed. Then, as* $n \to \infty$,

$$\textit{Number of heads} \approx N\big(np, np(1-p)\big).$$

If you choose to get only one thing out of this course, it should be to understand this theorem and how to apply it!!

In words: Suppose we perform n independent trials and each trial has a probability of success p. Then as the number of trials increases (n becomes large) the random number of successes (which is a Binomial(n,p)) becomes approximately a normal random variable. The mean and variance of this normal random variable are the same as the Binomial, i.e. np and $np(1-p)$, respectively.

The way the above theorem is applied is via the following formulas: As n grows we have

$$\begin{aligned} P\big\{\text{Between } a \text{ and } b \text{ successes}\big\} &\approx P\big\{a \leqslant N\big(np, np(1-p)\big) \leqslant b\big\} \\ &= \Phi\Big(\frac{b-np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{a-np}{\sqrt{np(1-p)}}\Big), \end{aligned}$$

$$\begin{aligned} P\big\{\text{Less than } b \text{ successes}\big\} &\approx P\big\{N\big(np, np(1-p)\big) \leqslant b\big\} \\ &= \Phi\Big(\frac{b-np}{\sqrt{np(1-p)}}\Big), \end{aligned}$$

and

$$\begin{aligned} P\big\{\text{More than } a \text{ successes}\big\} &\approx P\big\{N\big(np, np(1-p)\big) \geqslant a\big\} \\ &= 1 - \Phi\Big(\frac{a-np}{\sqrt{np(1-p)}}\Big). \end{aligned}$$

Here, as before, $\Phi(z)$ is the probability a standard normal is below z:

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

Two remarks are in order, before we move to examples.

Remark 14.3. Both the central limit theorem and Poisson's law of rare events give approximations of the Binomial(n,p) when n is large. So what is the difference between the two?

Poisson's approximation kicks in when pn is fixed. In this case, as n grows p decays and we are looking at the number of occurrences of a rare event in a large number of repetitions. The central limit theorem, on the other hand, kicks in when p is fixed. So we are looking at the number of occurrences of a regular (not rare) event in a large number of repetitions.

One of the uses of the central limit theorem is to help us predict the value of p by producing an interval that contains the true (unknown to us) value of p with high probability. Here is a concrete example based on exit polls.

Example 14.4 (Confidence intervals)**.** In an exit poll the evening of a presidential election a random sample of 1963 voters were queried regarding who they voted for. 1021 of them said they voted for Mr. Niceguy. The law of large numbers then tells us that the proportion in the sample should be close to the proportion in the population. So based on the results of the exit poll we predict that after all the ballots are counted we will have close to $1021/1963 \approx 0.52$ or 52% of all votes being for Mr. Niceguy. But can we quantify the accuracy of this estimate? Here is how the central limit theorem offers an answer.

First, let us set up the probability model. We have some large number N of people who plan on voting (in 2016 the turnout was about $N \approx 139$ million voters). Suppose among them there are K that will vote for Mr. Niceguy. So the proportion of the population that will vote for this candidate is $p = K/N$. We are trying to predict p, since we do not know K nor N before the ballots are all counted. In our exit poll, we pick $n = 1963$ of the N individuals at random, equally likely, but without replacement. As we have seen before in Example 10.1, since the sample size n is much smaller than the size N of the population it is in fact not a bad approximation to say that the sampling was done with replacement, in which case the number of people in the sample who vote for Mr. Niceguy would be a Binomial random variable with parameters n and p.

To mathematically justify this approximation (of pretending we are sampling with replacement instead of without replacement) we compute the mass function of the number of Mr. Niceguy voters when the sampling is done with replacement. For a given integer k between 0 and n the probability we have exactly k people in our sample that will vote for Mr. Niceguy (and hence $n-k$ that will not vote for him) is equal to

$$\frac{\overbrace{\binom{n}{k}}^{\text{which } k \text{ of the } n \text{ voted for Mr. Niceguy}} \overbrace{K(K-1)\cdots(K-k+1)}^{k \text{ terms}} \overbrace{(N-K)(N-K-1)\cdots(N-K-n+k+1)}^{n-k \text{ terms}}}{\underbrace{N(N-1)\cdots(N-n+1)}_{n \text{ terms}}}.$$

Since $K = Np$ the above becomes

$$\begin{aligned}
&\binom{n}{k}\frac{K}{N}\cdot\frac{K-1}{N-1}\cdots\frac{K-k+1}{N-k+1}\cdot\frac{N-K}{N-k}\cdot\frac{N-K-1}{N-k-1}\cdots\frac{N-K-n+k+1}{N-n+1}\\
&\quad=\binom{n}{k}p\cdot\frac{Np-1}{N-1}\cdots\frac{Np-k+1}{Np-k+1}\cdot\frac{N(1-p)}{N-k}\cdot\frac{N(1-p)-1}{N-k-1}\cdots\frac{N(1-p)-n+k+1}{N-n+1}.
\end{aligned}$$

Since N is large compared to k we can consider the limit as $N\to\infty$ and then the above becomes:

$$\binom{n}{k}p^k(1-p)^{n-k}.$$

This is indeed the mass function of a Binomial(n, p) and so sampling with replacement is not much different from sampling without replacement (when the size of the sample is much smaller than the size of the population).

OK. So as a first approximation we will pretend that we have the simpler model where the number of votes Mr. Niceguy collects among our n polled voters is a Binomial(n, p). We know $n = 1963$ and we are trying to predict p. Let us denote the number of votes for Mr. Niceguy, in the exit poll, by X. As mentioned above, the law of large numbers says that $X/n \approx p$. If we now allow ourselves a margin of error $\varepsilon > 0$, then the probability that our estimate X/n is within this margin of error is equal to

$$\begin{aligned} P\Big\{-\varepsilon \leqslant \frac{X}{n} - p \leqslant \varepsilon\Big\} &= P\Big\{-\varepsilon n \leqslant X - np \leqslant \varepsilon n\Big\} \\ &= P\Big\{-\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}} \leqslant \frac{X - np}{\sqrt{np(1-p)}} \leqslant \varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big\}. \end{aligned} \tag{14.1}$$

Now, our second approximation: The central limit theorem says that this is approximately equal to

$$\begin{aligned} &\Phi\Big(\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big) - \Phi\Big(-\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big) \\ &\quad = \Phi\Big(\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big) - \Big(1 - \Phi\Big(\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big)\Big) \\ &\quad = 2\Phi\Big(\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}}\Big) - 1. \end{aligned}$$

Suppose we want to be 95% confident that our estimate X/n is within ε of the actual value of p. Then we want to set the above to 0.95. So we look in the standard normal table for the value of z that would make $2\Phi(z) - 1 = 0.95$, or $\Phi(z) = 1.95/2 = 0.975$. We find that $z = 1.96$. So now we know that

$$\varepsilon \cdot \sqrt{\frac{n}{p(1-p)}} = 1.96$$

and hence

$$\varepsilon = \frac{1.96\sqrt{p(1-p)}}{\sqrt{n}}.$$

The problem is that we do not know what p is. But we have already estimated it to be about $1021/1963 \approx 0.5201$. Plugging this in the above (and $n = 1963$) we get that $\varepsilon \approx 0.0221$.

The upshot is that we are 95% confident that when the ballots are all counted the proportion of votes that Mr. Niceguy will collect is between $0.5201 - 0.0221 = 0.498$ (49.8%) and $0.5201 + 0.0221 = 0.5422$ (54.22%). The news would report this using the statement:

> *"The latest poll shows that 52% of voters favor candidate Niceguy, with a margin of error of 2 percentage points."*

The level of confidence used to produce the estimate is usually omitted from news reports. This is no doubt partly due to a desire to avoid confusing technicalities. It is also a fairly common convention to set the confidence level at 95%, so it does not need to be stated explicitly.

One can also use the central limit theorem to test a given hypothesis about the model, by quantifying the extent to which the data confirms the hypothesis. Here is an example.

Example 14.5 (Hypothesis testing)**.** Someone makes the claim that a certain population has more women than men. To test this hypothesis we sampled 10,000 random people from the population and observed that there were $5,050$ men in the sample. This does not contradict the hypothesis as of yet. Even if the claim were true, we could still get more men than women in a random sample due to the randomness in the sampling procedure.

What we should ask ourselves is this: suppose that the hypothesis were true. Then what are the odds that a sample of 10,000 would give as many as 5,050 men or more?

In fact, since the claim being made does not specify the actual proportion of women and since we need that number to be able to do the math, we will ask the even more extreme question: suppose that there are in fact exactly 50% men and 50% women in the population. Then what are the odds that a sample of 10,000 would give as many as 5,050 men or more? If the odds for this are low, then the would be even lower under the original claim that the percentage of women is higher than 50%. So let us now do the math.

Again, if the population is much larger than our sample of 10,000, then we can consider that sampling without replacement is very close to sampling with replacement, in which case the number of men in the sample would be a Binomial$(10,000, 1/2)$. (Sample of $10,000$ and we are assuming the probability of choosing a male is 50%.) We are wondering about the probability this binomial random variable is 5,050 or larger. The central limit theorem says that this is approximately equal to

$$1-\Phi\Big(\frac{5050-10000\times 0.5}{\sqrt{10,000\times 0.5\times(1-0.5)}}\Big)=1-\Phi(1)\approx 1-0.8413=0.1587. \tag{14.2}$$

So if we had 50% men and 50% women in the population, then there is about 15.87% chance that a random sample of 10,000 people would have 5,050 men or more. This is not small enough to say that getting a number as large as 5,050 men in a sample of 10,000 is not probable. It may well be that the population has half men and half women and we got 5,050 men in our sample just due to randomness. So in this case, we fail to reject the hypothesis.

However, had our sample had say 5,157 men, then the above probability would have been

$$1-\Phi\Big(\frac{5157-10000\times 0.5}{\sqrt{10,000\times 0.5\times(1-0.5)}}\Big)=1-\Phi(3.2)\approx 1-0.9993=0.0007.$$

Now, this probability is quite small and so we would deduce that if the population had half men and half women then it would be quite unlikely to get a sample with as many as 5,157 men. Hence, we would go ahead and reject the hypothesis that the population has more women than men in favor of the alternative (that there are more men than women), with quite some confidence (in fact, with 99.93% confidence!).

Note how maybe one would have said that 5,050 is not significantly far from 5,000 to say that there are more men than women in the population, and yet with the same reasoning it would be hard to eyeball that with only about 100 more men, 5,157 there becomes overwhelming statistical evidence that there are in fact more men than women in the population! This is the whole point of using probability (the central limit theorem in this particular example) to quantify things rather than just making vague statements and unjustified conclusions (like a casual *"I think there is a 60% chance team A will win tonight's football game.."*).

Here is one way to visualize the central limit theorem.

Example 14.6 (Galton board)**.** Suppose we place a bunch of pegs on a board as in Figure 1 (the red dots) and then drop a small ball (blue dot) above the very top peg. Then at each row the ball

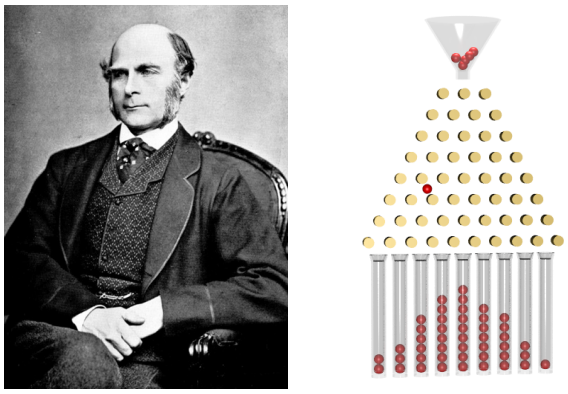


Figure 14.2. Left: Sir Francis Galton (February 16, 1822 – January 17, 1911, England). Right: The Galton board: A ball is dropped over the top peg. When the balls reach the bottom level they are stacked into columns. This basically plots the histogram of the random position of the ball when it reaches that level.

will either bounce one slot to the left or one slot to the right, with probability $1/2$ each. Fix a row n and let X be the number of times the ball went to the right between when it was first released (row 0) and the current row n. Then the ball went $n - X$ times to the left and so the position of the ball, relative to the middle, when it reaches level n is $X - (n - X) = 2X - n$.

The central limit theorem says that if we look at a large n (i.e. let the ball drop for quite a few rows), then X is approximately an $N(n/2, n/4)$ random variable. ($p = 1/2$, so $np = n/2$ and $np(1-p) = n/4$.) But this means that $2X - n$ has mean $2 \times (n/2) - n = 0$ and variance $4 \times (n/4) = n$. (Remember properties (3) and (4) in Theorem 13.14.) So the position of the ball relative to where it was dropped is approximately a normal random variable with mean 0 and variance n.

This can be visualized by dropping a lot of balls and stacking the balls in long vertical columns as they arrive at level n. The stacks of balls represent the histogram $2X - n$ (the position where the ball arrives at level n) and we can then see how the bell curve starts forming! See the left picture in Figure 1. See also the video on the course webpage:

http://www.math.utah.edu/~firas/5010/Galton_Board.mp4

Read sections 4.1, 4.2, and 4.3 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 14.1. Let X be the number of successes in a sequence of $10,000$ Bernoulli trials with probability of success 0.8. Estimate $P\{7940 \leqslant X \leqslant 8080\}$.

Exercises 4.1, 4.2, 4.4, 4.5, 4.6, 4.7 on pages 171 and 172 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 15

1. Joint distributions

Now we move from studying one random variable to studying multiple random variables simultaneously and how to quantify their dependence/independence structure. As usual, we start with discrete random variables. Also, we will focus on the study of two random variables, but one should keep in mind that the discussion is in fact more general and works for any number of random variables.

Say X and Y are two discrete random variables. Then their *joint mass function* is simply given by

$$f(x, y) = P\{X = x, Y = y\}.$$

We sometimes write $f_{X,Y}$ in place of f in order to emphasize that this is the mass function of the pair X and Y.

Just like the univariate mass function, the joint mass function has the properties:

$$f(x, y) \geqslant 0 \text{ for all } x, y \quad \text{and} \quad \sum_x \sum_y f(x, y) = 1.$$

In order to calculate the probability that the pair (X, Y) belongs to a set C (a subset of $\mathbb{R}^2$) we simply add the values of the mass function over all pairs (x, y) in C:

$$P\{(X, Y) \in C\} = \sum_{(x,y)\in C} f(x, y).$$

[0] Last modified on March 13, 2020 at 08:43:45 -06'00'

Example 15.1. 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

$$f(x,y) = \begin{cases} \frac{1}{36} & \text{if } x = 2 \text{ and } y = 0, \\ \frac{1}{36} & \text{if } x = 0 \text{ and } y = 2, \\ \frac{2}{36} & \text{if } x = y = 1, \\ \frac{8}{36} & \text{if } x = 0 \text{ and } y = 1, \\ \frac{8}{36} & \text{if } x = 1 \text{ and } y = 0, \\ \frac{16}{36} & \text{if } x = y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

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

If we want to compute the probability that $|Y - X| \geqslant 1$ then this is the same as computing the probability that the pair (X, Y) is in the set

$$C = \{(x,y) : |y - x| \geqslant 1\} = \{(0,1), (1,0), (2,0), (0,2)\}$$

and the probability is

$$\begin{aligned} P\{|Y - X| \geqslant 1\} &= P\{(X,Y) \in C\} = f(0,1) + f(1,0) + f(2,0) + f(0,2) \\ &= \frac{8}{36} + \frac{8}{36} + \frac{1}{36} + \frac{1}{36} = \frac{18}{36} = \frac{1}{2}\,. \end{aligned}$$

Once we know the joint mass function of X and Y we can describe random variables that are made out of X and Y, as in the next example.

Example 15.2. Given the pair of random variables in the previous example, when is the mass function of the random variable $Z = \min(X,Y)$? To answer this question we first identify the possible values of Z. Inspecting the above table we see that Z can only take the values 0 or 1. And we have

$$\begin{aligned} P(Z = 0) &= f(2,0) + f(0,2) + f(0,1) + f(1,0) + f(0,0) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{8}{36} + \frac{8}{36} + \frac{16}{36} = \frac{34}{36} \end{aligned}$$

and

$$P(Z = 1) = f(1,1) = \frac{2}{36}\,.$$

So Z is a Bernoulli(2/36).

In particular, if we know the joint mass function of X and Y we can recover the individual mass functions f_X and f_Y of the random variables X and Y. For instance, in the above example

$$f_X(1) = P\{X = 1\} = P\{X = 1, Y = 0\} + P\{X = 1, Y = 1\} = f(1,0) + f(1,1) = \frac{10}{36}.$$

If we want to use the joint mass function table, then what we do is compute the row sums to get the mass function for one of the variables (f_X in our example) and do the same with the column sums to get the mass function of the other variable (f_Y here). So in the above example we have:

$x \setminus y$	**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**" in the bottom right designates three things: the right-most column entries add up to 1 (being the entries of the mass function of X), the bottom-row entries add up to 1 (being the entries of the mass function of Y), and the entries of the table itself add up to 1 (being the entries of the joint mass function of X and Y).

This is of course more general than just the above example.

Theorem 15.3. *The individual mass functions can be obtained from the joint mass function as follows:*

(1) $f_X(x) = \sum_y f(x, y)$ *(sum is over the possible values* y *that* Y *can take),*

(2) $f_Y(y) = \sum_x f(x, y)$ *(sum is over the possible values* x *that* X *can take).*

These individual mass functions are called the *marginal* mass functions or *marginal* distributions.

Just like in the univariate case, once we are given the joint mass function we can compute the expected values of functions of the two random variables.

Theorem 15.4. *Suppose* $h(x, y)$ *is some given function. Then*

$$E[h(X, Y)] = \sum_{x,y} h(x, y) f_{X,Y}(x, y). \tag{15.1}$$

Proof. Let us denote $h(X, Y)$ by Z. Then the mass function of Z is given by

$$f_Z(z) = P(Z = z) = \sum_{x,y:h(x,y)=z} P(X = x, Y = y) = \sum_{x,y:h(x,y)=z} f_{X,Y}(x, y).$$

And so

$$\begin{aligned} E[h(X, Y)] = E[Z] &= \sum_z z f_Z(z) = \sum_z \sum_{x,y:h(x,y)=z} z f_{X,Y}(x, y) \\ &= \sum_z \sum_{x,y:h(x,y)=z} h(x, y) f_{X,Y}(x, y). \end{aligned}$$

In the last equality we replaced z by $h(x, y)$ because the second double sum is only over x, y that satisfy $h(x, y) = z$. So we can replace z by $h(x, y)$ since they are equal inside that sum. Now to finish notice that on the second line of the above display we are adding over all x, y such that $h(x, y) = z$ but then adding over all the values of z. So the restriction that $h(x, y) = z$ is lifted and we are actually adding over all the possible values of x and y, without the need to add over z. This leads to

$$E[h(X, Y)] = \sum_{x,y} h(x, y) f_{X,Y}(x, y),$$

as claimed. □

Example 15.5. Back to Example 15.1. We can compute

$$\begin{aligned}\mathrm{E}[|X-Y|] &= \sum_{x,y} |x-y|\, f(x,y) \\ &= |2-0|\cdot\frac{1}{36} + |0-2|\cdot\frac{1}{36} + |1-1|\cdot\frac{2}{36} + |0-1|\cdot\frac{8}{36} + |1-0|\cdot\frac{8}{36} + |0-0|\cdot\frac{16}{36} \\ &= \frac{5}{9}.\end{aligned}$$

Just as in the univariate case, the expected value is a linear operation.

Theorem 15.6. *Let X, Y be two random variables and let $h(x,y)$ and $g(x,y)$ be two functions. Then*

$$\mathrm{E}[h(X,Y) + g(X,Y)] = \mathrm{E}[h(X,Y)] + \mathrm{E}[g(X,Y)].$$

The proof is straightforward from Theorem 15.4. (Work it out!) In particular,

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

2. Independence

Two random variables X and Y are independent if any statement involving only X is independent of any statement involving only Y. In particular, for any x and y, possible values of X and Y, respectively, the events $\{X = x\}$ and $\{Y = y\}$ should be independent. But this implies that

$$f_{X,Y}(x,y) = P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\} = f_X(x)f_Y(y). \tag{15.2}$$

In words: the joint mass function happens to factor into a product of two functions, one in each of the variables. In fact, this turns out to not only follow from the independence of X and Y but to also imply it. Namely, suppose A and B are two sets, and the joint mass function of the pair (X,Y) factors out as above. Then

$$\begin{aligned}P\{X \in A\,, Y \in B\} &= \sum_{x\in A}\sum_{y\in B} f_{X,Y}(x\,,y) = \sum_{x\in A}\sum_{y\in B} f_X(x)f_Y(y) \\ &= \sum_{y\in B}\sum_{x\in A} f_X(x)f_Y(y) \qquad \text{(addition is commutative)} \\ &= \sum_{y\in B} f_Y(y)\Big(\sum_{x\in A} f_X(x)\Big) \qquad \text{(factored out the } f_Y(y) \text{ from the sum over } x) \\ &= \Big(\sum_{x\in A} f_X(x)\Big)\sum_{y\in B} f_Y(y) \qquad \text{(the sum over } x \text{ does not depend on } y \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{and was factored out of the sum over } y) \\ &= P\{X \in A\}P\{Y \in B\}.\end{aligned}$$

Since A and B are arbitrary, this does say that if the mass function of (X,Y) factors as in (15.2), then any statement only involving X is independent of any statement only involving Y. So this factoring property of the joint mass function is in fact equivalent to saying that X and Y are two independent random variables.

In fact, one can push the above discussion further: if h and g are any two functions and X and Y are independent random variables, then $h(X)$ and $g(Y)$ are also independent. We omit the (not too difficult) proof of this fact, but hope that the fact itself feels natural to the student.

Example 15.7 (Example 15.1, continued)**.** Note that in this example, X and Y are not independent. For instance,

$$f(1,2) = 0 \neq f_X(1)f_Y(2) = \frac{10}{36} \times \frac{1}{36}.$$

As formula (15.1) says (and as we saw in Example 15.5), one needs the joint mass function to answer questions that involve several random variables. Without independence one cannot construct this joint mass function from only knowing the marginal mass functions. However, this can be done when the random variables are independent: the joint mass function can be obtained from the marginal mass functions using (15.2).

Example 15.8. *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 are running repeated independent trials and waiting for the first success, where failure means both coins landed tails and success means one of the two coins landed heads. So the result should be a geometric random variable with probability of failure being $q_1 q_2$ (both coins landed tails) and thus the probability of success is $1 - q_1 q_2$.

Example 15.9. Suppose we repeatedly roll a fair die. Let X denote the number of rolls we need until we get a 1, 2, 3, or 4. This is of course a geometric random variable with success probability $4/6 = 2/3$. Its mass function is therefore given by $f_X(k) = (1/3)^{j-1} \cdot (2/3)$, for $j = 1, 2, 3, \ldots$. Let Y denote the outcome of the die at the last roll. So Y takes values 1, 2, 3, or 4. Since the die is fair it should be clear that all four outcomes should have the same probability. (1 is no different from 2, 3, or 4 and so we should have the same probability to stop on a 1 as on a 2 and so on.) So the mass function for Y is $f_Y(k) = 1/4$ for $k = 1, 2, 3, 4$. Are X and Y independent random variables? In principle, they may be dependent. So let us check out the definition. For this we need to compute the joint mass function of the pair (X, Y). That is, for an integer $j \geqslant 1$ and a $k \in \{1, 2, 3, 4\}$ we want to compute

$$f_{X,Y}(j, k) = P\{X = j, Y = k\}.$$

Saying that $X = j$ and $Y = k$ is the same as saying that for the first $j - 1$ rolls the die landed on 5s or 6s (and so we kept rolling the die) and then the j-th roll was a k (so we stopped on a k). The probability of this happening is

$$P\{X = j, Y = k\} = \frac{2^{j-1} \cdot 1}{6^j}.$$

Let us check now if we have independence:

$$f_X(j)f_Y(k) = \left(\frac{1}{3}\right)^{j-1} \cdot \frac{2}{3} \times \frac{1}{4} = \frac{1}{3^j} \cdot \frac{1}{2}$$

while

$$f_{X,Y}(j,k) = \frac{2^{j-1}\cdot 1}{6^j} = \frac{1}{2}\cdot\left(\frac{2}{6}\right)^j = \frac{1}{2}\cdot\frac{1}{3^j}\,.$$

Since $f_{X,Y}(j,k) = f_X(j)f_Y(k)$ for all integers $j \geqslant 1$ and all $k \in \{1,2,3,4\}$ we deduce that X and Y are in fact independent.

One last remark to close this lecture: Recall what we said at the beginning of the lecture: this whole discussion holds for more than just two random variables. In particular, random variables $X_1, X_2, \ldots, X_n$ are independent if their joint mass function

$$f(x_1, x_2, \ldots, x_n) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

factors into a product of n functions, one in each variable:

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

Read section 6.1, section 6.3 up to (but not including) Fact 6.25, and then Examples 6.31 and 6.32 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 15.1. Let X and Y be two discrete random variables with joint mass function $f(x, y)$ given by

$x \backslash y$	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 15.2. 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 15.3. 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?

Exercises 6.1, 6.2, 6.4, 6.8(a), 6.9 on pages 236, 237, and 238 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 16

1. Jointly distributed continuous random variables

Similarly to the discrete case a pair of random variables are said to be *jointly continuous* if they have a joint probability distribution function (pdf). Then one can compute the probability the pair (X, Y) is in some given set A by performing a double integral:

$$P\{(X, Y) \in A\} = \iint_A f(x, y)\, dx\, dy.$$

Joint pdfs have the properties:

$$f(x, y) \geqslant 0 \text{ for all } x \text{ and } y \quad \text{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

Example 16.1 (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 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 16.1. 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 16.2. Let (X, Y) be uniformly distributed on $[-1, 1]^2$. That is,

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{1}{4} & \text{if } -1 \leqslant x \leqslant 1 \text{ and } -1 \leqslant y \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

[0] Last modified on May 12, 2020 at 13:09:25 -06'00'

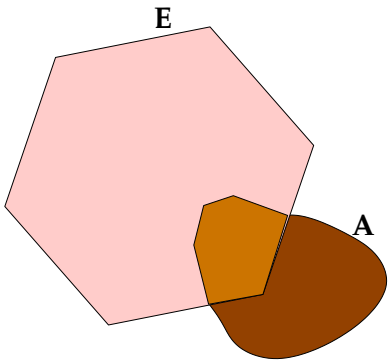


Figure 16.1. Region of integration in Example 16.1

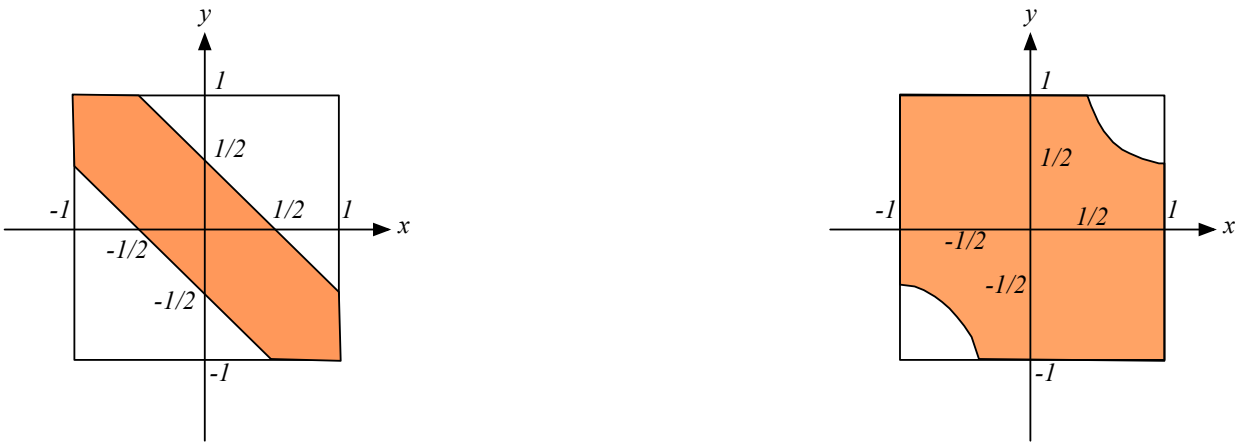


Figure 16.2. Regions of integration for Example 16.2. Left: $|x+y| \leqslant 1/2$. Right: $xy \leqslant 1/2$.

We want to find $P\{|X+Y| \leqslant 1/2\}$. In this case, the areas are easy to compute geometrically; see Figure 16.2. 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

$$\begin{aligned} P\{|X+Y| \leqslant 1/2\} &= \iint_{|x+y|\leqslant 1/2} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-1}^{-1/2}\int_{-x-1/2}^{1} \frac{1}{4}\,dy\,dx + \int_{-1/2}^{1/2}\int_{-x-1/2}^{-x+1/2} \frac{1}{4}\,dy\,dx + \int_{1/2}^{1}\int_{-1}^{-x+1/2} \frac{1}{4}\,dy\,dx \\ &= \frac{7}{16}. \end{aligned}$$

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 16.2. Thus, we need to compute it using the definition of joint

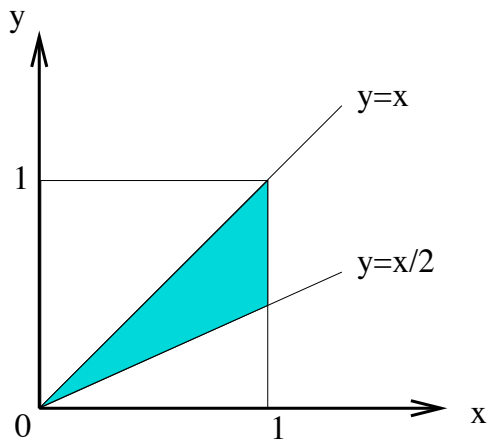


Figure 16.3. Region of integration in Example 16.3.

density functions.

$$\begin{aligned}
\mathrm{P}\{XY \leqslant 1/2\} &= \iint\limits_{xy\leqslant 1/2} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-1}^{-1/2} \underbrace{\int_{1/2x}^{1} \frac{1}{4}\,dy}_{(1/4-1/8x)}\,dx + \int_{-1/2}^{1/2} \underbrace{\int_{-1}^{1} \frac{1}{4}\,dy}_{2/4}\,dx + \int_{1/2}^{1} \underbrace{\int_{-1}^{1/2x} \frac{1}{4}\,dy}_{(1/8x+1/4)}\,dx \\
&= \Big(\frac{x}{4} - \frac{\ln|x|}{8}\Big)\Big|_{-1}^{-1/2} + \frac{1}{2} + \Big(\frac{\ln|x|}{8} + \frac{x}{4}\Big)\Big|_{1/2}^{1} = \frac{3}{4} + \frac{\ln 2}{4}.
\end{aligned}$$

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 16.2. 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

$$\mathrm{P}\{XY \leqslant 1/2\} = 1 - 2\int_{1/2}^{1} \Big(\int_{1/2x}^{1} \frac{1}{4}\,dy\Big)\,dx = 1 - 2\int_{1/2}^{1} (\frac{1}{4} - \frac{1}{8x})\,dx = \frac{3}{4} + \frac{\ln 2}{4}.$$

Example 16.3. 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 $\mathrm{P}\{X \leqslant 2Y\}$. To find C:

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^x Cxy\,dy\,dx \\
&= C\int_0^1 x\,\underbrace{\Big(\int_0^x y\,dy\Big)}_{\frac{1}{2}x^2}\,dx = \frac{C}{2}\int_0^1 x^3\,dx = \frac{C}{8}.
\end{aligned}$$

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 16.3. (Graphing a figure always helps!)

2. Marginals distributions.

If (X,Y) has joint density f, then we can recover the individual pdfs by observing that

$$P\{X \leqslant a\} = P\{(X,Y) \in A\},$$

where $A = \{(x,\, y) : x \leqslant a\}$. Thus,

$$P\{X \leqslant a\} = \int_{-\infty}^a \left(\int_{-\infty}^\infty f(x,y)\,dy \right) dx,$$

But recall that if X has pdf f_X, then

$$P\{X \leqslant a\} = \int_{-\infty}^a f_X(x)\,dx\,.$$

This means that $P\{X \leqslant x\}$ is an antiderivative to both $f_X(x)$ and to $\int_{-\infty}^\infty f(x,y)\,dy$. So these two functions of x differ only by a constant. But since they both go to 0 as $x \to -\infty$ (because both have to integrate to 1), the constant is zero and the two functions are actually equal. We have thus shown that

$$f_X(x) = \int_{-\infty}^\infty f(x,y)\,dy.$$

Similarly,

$$f_Y(y) = \int_{-\infty}^\infty f(x,y)\,dx.$$

Compare this with Theorem 15.3.

Example 16.4 (Example 16.3, continued)**.** Let

$$f(x,y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} f_X(x) &= \begin{cases} \int_0^x 8xy\,dy & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 4x^3 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note how we had to be careful when integrating out the y variable. The joint mass function is not zero only on the interval $[0, x]$, which depends on x. So the boundary of integration had to depend on x. A similar story happens when computing the pdf of Y:

$$f_Y(y) = \begin{cases} \int_y^1 8xy\,dx & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$
$$= \begin{cases} 4y(1-y^2) & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 16.5. Suppose (X, Y) is distributed uniformly in the circle of radius one about $(0, 0)$. That is,

$$f(x, y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$f_X(x) = \begin{cases} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$
$$= \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

By symmetry, f_Y is the same function:

$$f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1-y^2} & \text{if } -1 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

3. 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 $g(X)$ and $h(Y)$ are also independent, for any functions g and h. We omit the short proof.

Example 16.6. *Let $X \sim$ Exponential(λ_1) and $Y \sim$ Exponential(λ_2). What kind of variable is $Z = \min(X, Y)$?*

Let us compute Z's pdf. For this we start by computing for $z > 0$:

$$P\{Z \leqslant z\} = P\{\min(X, Y) \leqslant z\} = 1 - P\{X > z, Y > z\}$$
$$= 1 - P\{X > z\}P\{Y > z\} = 1 - (1 - F_X(z))(1 - F_Y(z))$$
$$= 1 - e^{-\lambda_1 z} e^{-\lambda_2 z} = 1 - e^{-(\lambda_1+\lambda_2)z}.$$

Recall that $P\{Z \leqslant a\} = \int_{-\infty}^{a} f_Z(z)\,dz$, which says that $P\{Z \leqslant a\}$ is the antiderivative of f_Z. So we can differentiate the above to find that

$$f_Z(z) = (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)z},$$

when $z > 0$. Since Z is the minimum of two positive random variables, it has to be positive. So the pdf is zero when $z < 0$. We have thus shown that $Z \sim \text{Exponential}(\lambda_1 + \lambda_2)$.

This makes sense: you are first in line and there are two cashiers each currently serving a customer. The first booth will take time X, an Exponential(λ_1), to clear and the second takes time Y, an Exponential(λ_2). You will go to whoever clears first. Thus, $Z = \min(X, Y)$ is your waiting time and it stands reason that this is again an exponential random variable. To compute its rate with think this way: the first employee will serve λ_1 people per unit time and the second will serve λ_2 employees per unit time. If customers come and go always to the first available booth, then that is like pooling the two cashiers together. They will thus serve $\lambda_1 + \lambda_2$ customers per unit time and this is the rate of Z. Be careful though: this argument is only for intuition. The proof that Z is Exponential($\lambda_1 + \lambda_2$) is the mathematical computation we did above.

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 f_X and f_Y integrate to one.

Example 16.7. 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 16.8. 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.

Another way to see the dependence comes by computing the probability $P\{X \geqslant 0 \text{ and } Y \geqslant 0\}$ and compare it to $P\{X \geqslant 0\}P\{Y \geqslant 0\}$. 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 16.9 (Example 16.4, 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 16.4, 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 16.5. There, X and Y are not independent either and this should be clear not only from the fact that the joint pdf is not equal to the product of the two marginals, but also from the fact that the two random variables have the restriction $X^2 + Y^2 \leqslant 1$. So if we know X is very close to 1, this forces Y to be small.

Example 16.10. Let (X,Y) be two independent standard normal random variables. Let $U = X+Y$ and $V = X-Y$. What is the joint pdf of U and V?

To answer this question start by observing that the joint pdf of (X,Y) is given by

$$f(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Let us compute

$$P\{U \leqslant a, V \leqslant b\} = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2}\,dx\,dy,$$

where $A = \{(x,y) : x+y \leqslant a,\ x-y \leqslant b\}$. Changing variables from x and y to $u = x+y$ and $v = x-y$ we note that the region A becomes simply $A = \{(u,v) : u \leqslant a, v \leqslant b\}$. Also, we have $x = (u+v)/2$ and $y = (u-v)/2$ and so we can substitute

$$\frac{x^2+y^2}{2} = \frac{u^2+v^2+2uv+u^2+v^2-2uv}{8} = \frac{u^2+v^2}{4}.$$

Lastly, recall from calculus that to perform the change of variables in the integration we need to compute the determinant of the Jacobian matrix:

$$\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}.$$

The determinant of this matrix is $-1/4 - 1/4 = -1/2$ and therefore $dx\,dy$ needs to be replaced by $\left|-\frac{1}{2}\right| du\,dv$. So the above integral becomes

$$P\{U \leqslant a, V \leqslant b\} = \frac{1}{4\pi}\int_{-\infty}^{a}\int_{-\infty}^{b} e^{-(u^2+v^2)/4}\,du\,dv.$$

On the other hand, we can also compute this in terms of the joint pdf of U and V:

$$P\{U \leqslant a, V \leqslant b\} = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{U,V}(u,v)\,du\,dv.$$

Therefore, we see that

$$f_{U,V}(u,v) = \frac{1}{4\pi}e^{-(u^2+v^2)/4}.$$

Note that this can be written as a product of a function in u and a function in v. Namely,

$$f_{U,V}(u,v) = \frac{1}{4\pi}e^{-u^2/4}\cdot e^{-v^2/4}.$$

Therefore, U and V are independent! This is not clear from the outset, since both U and V involve X and Y! But this particular combination (U is the sum and V is the difference) turns out to give two independent random variables. This goes to show again that dependence and independence are subtle issues.

Note also that the two functions into which $f_{U,V}$ factored are of the form constant times $e^{-u^2/4}$ and constant times $e^{-v^2/4}$. This means that U and V are normal random variables. We can also read off the mean and variance: the mean is 0 and the variance is 2, because $e^{-u^2/4} = e^{-\frac{(u-\mu)^2}{2\sigma^2}}$ says that $\mu = 0$ and $\sigma^2 = 2$. So in fact,

$$f_U(u) = \frac{1}{\sqrt{4\pi}}e^{-u^2/4} \quad \text{and} \quad f_V(v) = \frac{1}{\sqrt{4\pi}}e^{-v^2/4}$$

and indeed we see that $f_U(u)f_V(v) = f_{U,V}(u,v)$.

4. Expected values

The following is the analogue of Theorem 15.4 for jointly continuous random variables.

Theorem 16.11. *Suppose $h(x,y)$ is some given function. Then*

$$E[h(X,Y)] = \int h(x,y)f_{X,Y}(x,y)\,dx\,dy.$$

Example 16.12. Let X and Y be independent Exponential(λ) random variables. Can you see real quick why $E[X-Y]=0$? We now want to compute $E[|X-Y|]$. Then,

$$\begin{aligned}
E[|X-Y|] &= \iint |x-y|\, f(x,y)\,dx\,dy \\
&= \iint_{x>y} (x-y) f(x,y)\,dx\,dy + \iint_{x<y} (y-x) f(x,y)\,dx\,dy \\
&= 2\lambda^2 \int_0^\infty \Big(\int_0^x (x-y) e^{-\lambda x} e^{-\lambda y}\,dy \Big)\,dx \\
&= 2\lambda \int_0^\infty x e^{-\lambda x}(1-e^{-\lambda x})\,dx - 2\lambda^2 \int_0^\infty e^{-\lambda x} \Big(\int_0^x y e^{-\lambda y}\,dy \Big)\,dx.
\end{aligned}$$

We integrate by parts to compute

$$\begin{aligned}
\int_0^x y e^{-\lambda y}\,dy &= -\frac{1}{\lambda} \int_0^x y (e^{-\lambda y})'\,dy \\
&= -\frac{1}{\lambda} y e^{-\lambda y}\Big|_0^x + \frac{1}{\lambda} \int_0^x e^{-\lambda y}\,dy \\
&= -\frac{1}{\lambda} x e^{-\lambda x} + \frac{1}{\lambda^2}(1-e^{-\lambda x}).
\end{aligned}$$

Now observe that $\lambda \int_0^\infty x e^{-\lambda x}\,dx = E[\text{Exponential}(\lambda)] = 1/\lambda$. The same way we have

$$2\lambda \int_0^\infty x e^{-2\lambda x}\,dx = 1/(2\lambda).$$

Also, we already know that $\lambda \int_0^\infty e^{-\lambda x}\,dx = 1$ and $2\lambda \int_0^\infty e^{-2\lambda x}\,dx = 1$. Putting all this together we get

$$E[|X-Y|] = \frac{2}{\lambda} - \frac{1}{2\lambda} + \frac{1}{2\lambda} - \frac{2}{\lambda} + \frac{2}{2\lambda} = \frac{1}{\lambda}.$$

When random variables are independent, something special happens regarding expected values of products.

Theorem 16.13. *Suppose X and Y are two independent random variables and suppose g and h are two functions from $\mathbb{R}$ to $\mathbb{R}$. Then*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

In particular,

$$E[XY] = E[X]E[Y]. \tag{16.1}$$

Proof. We prove this in the continuous case, the discrete case being similar (with integrals replaced by sums). Recall that since X and Y are independent, $f(x,y) = f_X(x) f_Y(y)$. Therefore,

$$\begin{aligned}
E[g(X)h(Y)] &= \int g(x)h(y) f(x,y)\,dx\,dy = \int g(x)h(y) f_X(x) f_Y(y)\,dx\,dy \\
&= \Big(\int g(x) f_X(x)\,dx \Big) \Big(\int h(y) f_Y(y)\,dy \Big) = E[g(X)]E[h(Y)],
\end{aligned}$$

which is the claim. □

Example 16.14. Suppose X is Exponential(λ_1), Y is Exponential(λ_2), and the two are independent. What is $E[XY]$?

Since the two random variables are independent we have $E[XY] = E[X]E[Y] = \frac{1}{\lambda_1} \cdot \frac{1}{\lambda_2}$. It is as simple as that. Of course, one could instead compute

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dx\,dy = \int_{0}^{\infty}\int_{0}^{\infty} xy \cdot \lambda_1 e^{-\lambda_1 x} \cdot \lambda_2 e^{-\lambda_2 y}\,dx\,dy = \cdots$$

But then one would be redoing the proof of the above theorem, only to discover that the two integrals split and the result is the product of the answers from each single integral. You should do it, though, and see for yourself!

Of course the whole discussion in this lecture works for more than two random variables. One can define a joint pdf $f(x_1,\ldots,x_n)$ of random variables $X_1,\ldots,X_n$ and from knowing the joint pdf we can find the individual pdfs by integrating out the rest of the variables. For example

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,x_2,\ldots,x_n)\,dx_2\ldots dx_n.$$

And random variables $X_1,\ldots,X_n$ are independent if and only if the joint pdf is the product of the individual ones:

$$f(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n).$$

Read sections 6.2, the rest of section 6.3, and sections 8.1 and 8.2 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 16.1. 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) $P\{|Y| \leqslant e^X\}$.

Exercise 16.2. 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 16.3. Let X and Y be two independent random variables, each uniformly distributed on $[-1, 1]$. Compute $E[\max(X, Y)]$.

Exercise 16.4. 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 16.5. 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 16.6. 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 16.7. 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.

Exercises 6.5, 6.8(b), 6.11, 6.12, 8.1, 8.5, 8.6, and 8.8 on pages 237, 238, 297, and 298 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 17

1. Covariance

Just like the variance of a random variable gave a way to measure the amount of randomness in the variable, there is a way to quantify how much a pair of random variables are "related" to each other. The covariance between X and Y is defined to be

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right], \tag{17.1}$$

where μ_X is the mean of X and μ_Y is the mean of Y. Also, just like the variance had an alternate formula, we can expand

$$(X-\mu_X)(Y-\mu_Y) = XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y,$$

and take the mean of both sides (and use the fact that e.g. $\mathrm{E}[X\mu_Y] = \mathrm{E}[X]\mu_Y = \mu_X\mu_Y$) to obtain the computationally useful formula for covariance:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mu_X\mu_Y = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{17.2}$$

Here are some properties of the covariance.

Theorem 17.1. *The following all hold true.*

(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* $\mathrm{Cov}(X,bY) = b\,\mathrm{Cov}(X,Y)$)*;*

(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 to prove (1)-(4) as an exercise! Note though that (1) says that variance is a special case of a covariance. It is simply the covariance of the random variable with itself. (2) says covariance is symmetric. (3) says that if one of the two variables is in fact not random, then there is no covariance between the two random variables (i.e. one of the two does not vary in relation to the other). (4) says that if we rescale one

[0] Last modified on March 20, 2020 at 14:46:05 -06'00'

of the variables the covariance scales accordingly. Note that if we rescale both variables then the covariance will pick up both scalings: $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$. You can see this by applying property (4) twice, once to the pull a out and then once to pull the b out.

Putting properties (4) and (5) together we see that covariance is a biliniear opertation: If a, b, c, and d are nonrandom numbers, and X, Y, Z, U are random variables, then

$$\mathrm{Cov}(aX + bY, cZ + dU) = ac\,\mathrm{Cov}(X, Z) + ad\,\mathrm{Cov}(X, U) + bc\,\mathrm{Cov}(Y, Z) + bd\,\mathrm{Cov}(Y, U).$$

Note how this is just like when you distribute addition over multiplication to write $(aX+bY)(cZ+dU) = acXZ + adXU + bcYZ + bdYU$.

Let us see why property (6) holds. For this, apply property (1), then the bilinearity, then again property (1) and the symmetry in property (2) to get

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathrm{Cov}(X + Y, X + Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y)\\ &= \mathrm{Var}(X) + 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y),\end{aligned}$$

as claimed.

Example 17.2 (Example 15.1, 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,

$$\mathrm{E}[XY] = 1 \times 1 \times \frac{2}{36} = \frac{2}{36}.$$

Also,

$$\mathrm{E}[X] = \mathrm{E}[Y] = 0 \times \frac{25}{36} + 1 \times \frac{10}{36} + 2 \times \frac{1}{36} = \frac{12}{36}.$$

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,

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\ \mathrm{Var}(Y)}}. \tag{17.3}$$

The advantage in using correlation instead of covariance is that it does not change when we rescale the variables X and Y. This means that if we change the units in which we are recording our data, the correlation will remain unchanged. Mathematically, this means that if a and b are two positive numbers, then $\rho(aX, bY) = \rho(X, Y)$. This follows from the bilinearity of the covariance in Theorem 17.1(4) and the quadratic scaling of the variance in Theorem 13.14(3). (Do the computation!)

Example 17.3 (Example 15.1, continued)**.** Note that

$$\mathrm{E}[X^2] = \mathrm{E}[Y^2] = 0^2 \times \frac{25}{36} + 1^2 \times \frac{10}{36} + 2^2 \times \frac{1}{36} = \frac{14}{36}.$$

We already calculated the mean to be $\mathrm{E}[X] = \mathrm{E}[Y] = 14/36$ and thus

$$\mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{14}{36} - \left(\frac{12}{36}\right)^2 = \frac{5}{18}.$$


Figure 17.1. 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)

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 rest of the lecture will help explain this.

First let us note that correlation is always a number between -1 and 1.

Theorem 17.4. *If* $E[X^2]$ *and* $E[Y^2]$ *are positive and finite, then* $-1 \leqslant \rho(X,Y) \leqslant 1$.

This is a straightforward variant of the following inequality.

Theorem 17.5 (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 $0 \leqslant 0$. □

To see now why Theorem 17.4 is true apply the above inequality but to the random variables $\overline{X} = X - \mu_X$ and $\overline{Y} = Y - \mu_Y$. This gives

$$E[(X-\mu_X)(Y-\mu_Y)] \leqslant \sqrt{E[(X-\mu_X)^2]\,E[(Y-\mu_Y)^2]}\,.$$

Recognizing the quantities under the square root as being the variances of X and Y and the quantity on the left as being the covariance, we get that this is saying exactly that $|\rho(X,Y)| \leqslant 1$.

3. Correlation and independence

We say that X and Y are *uncorrelated* if $\rho(X,Y) = 0$; equivalently, if $\mathrm{Cov}(X,Y) = 0$, which in turn means that

$$E[XY] = E[X]E[Y].$$

Recall now how we saw already in (16.1) that this property holds for independent random variables. So we have the following result.

Theorem 17.6. *If X and Y are independent then X and Y are uncorrelated.*

Example 17.7 (A counter example)**.** Sadly, it is only too common that people some times think that the converse to Theorem 17.6 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, $XY = Y^2Z = \pm 1$ 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. This is clear because $|X| = Y$. Here is one way to justify this mathematically:

$$P\{X = 1\,, Y = 2\} = 0 \neq \frac{1}{8} = P\{X = 1\}P\{Y = 2\}.$$

A striking property of uncorrelated random variables that follows from Theorem 17.1(8) is the following.

Theorem 17.8. *If X and Y are uncorrelated, then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

There is a sense in which the above formula can be thought of as the probabilistic version of the Pythagorean theorem! Random variables that are uncorrelated play the role of orthogonal vectors and the variance plays the role of the square of the distance. Then the above says exactly that the square of the norm of a sum of two orthogonal vectors is the sum of their norms. Exactly what the Pythagorean theorem says. This is the reason why we define the variance the way we define it and not, for example, as $E[|X - \mu_X|]$ (without the square). Even though this would be a perfectly fine way to measure the amount of random variation in X (since it measures how far X is from its mean μ_X), we would not have a nice formula like the one above. (Similarly to how in the two-dimensional plane, if we define the distance between points (x_1, y_1) and (x_2, y_2) as $|x_2 - x_1| + |y_2 - y_1|$, then we would not have the Pythagorian theorem.)

The formula in the above theorem says that if the random variables are uncorrelated, then the random variation in their sum comes simply from adding the random variation in each of the two variables. This is of course not true if there is some correlation between the random variables, because Theorem 17.1(8) tells us that the discrepancy between the variance of the sum and the sum of the variances is exactly twice the covariance.

It follows from the above that if random variables are independent, then we have just seen that they are uncorrelated, and hence, according to Theorem 17.8 the variance of the sum is equal to the sum of the variances.

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 also true.

Theorem 17.9. *Assume none of X and Y is constant (i.e. $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$). 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 13.13 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 $aX + b$. Taking derivatives in a and in b and setting them to 0 gives the equations

$$E[X^2]\, a + E[X]\, b - E[XY] = 0 \quad \text{and} \quad b + E[X]\, a - E[Y] = 0. \tag{17.4}$$

With a bit more elementary calculus we see that the a and b that solve the above equations minimize the function f (and thus make Y "as close as possible" to $aX + b$).

Solving the second equation in (17.4) for b in terms of a and plugging back into the first equation we get

$$Var(X)\, a + E[X]E[Y] - E[XY] = 0 \quad \text{and} \quad b = E[Y] - E[X]\, a.$$

Solving the first of these equations for a we get

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

Plugging back into f one gets the smallest value f can reach:

$$\begin{aligned} f(a,b) &= E\big[\big((Y-aX)-b\big)^2\big] \\ &= E\big[\big((Y-aX)-E[Y-aX]\big)^2\big] = \mathrm{Var}(Y-aX) \\ &= \mathrm{Var}(Y)+a^2\,\mathrm{Var}(X)-2a\,\mathrm{Cov}(X,Y) \\ &= \mathrm{Var}(Y)(1-\rho(X,Y)^2). \end{aligned}$$

So the closer $\rho(X,Y)$ is to 1 or -1, the closer Y is to being a linear function of X. This is why $\rho(X,Y)$ is called the *linear* correlation coefficient. It measures how close X and Y are being to be in a linear relationship.

We conclude this lecture with a couple of remarks.

Remark 17.10. As was just explained, the correlation that we studied in this lecture detects linear relationships. That is, if the correlation is very close to 1 or -1 then the two random variables are very close to being in a linear dependence on each other. However, as we also pointed out above, if the correlation is not very close to 1 nor to -1, then this does not mean that there is no relationship between X and Y (which is to say that X and Y are independent). All it says is that the two are not in a linear relationship. For example, if X is a Uniform$(-1,1)$ random variable and $Y = X^2$, then $E[X] = 0$ and $E[XY] = E[X^3] = 0$ (due to symmetry; do the integral calculation and see for yourself). Therefore, $\mathrm{Cor}(X,Y) = 0$ and yet Y is fully dependent on X, just not in a linear way.

Remark 17.11. Another common misconception is to confuse correlation with causation. Even if the correlation between X and Y is exactly 1, all this says is that the two depend linearly on each other. So knowing the value of one gives the value of the other by plugging into the equation of a line. And if one say increases then so does the other. But this does not say that increasing say X *causes* the increase in Y. An almost comical example is the following. A study on elementary school students revealed that the size the student's shoe and the amount of vocabulary they know are correlated with a correlation coefficient very close to 1. Does this mean that having a bigger foot makes you read better? Or does it mean that if you start reading lots of books your foot will grow in size? Both statements are nonsense, of course. What is going on is that there is what is called a *lurking variable* (age in this case) driving both variables and thus causing them to correlate. Having a bigger foot most probably means you are older and so you read more sophisticated books and vice versa.

Read section 8.4 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 17.1. 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 17.2. Let $X_1, \ldots, X_n$ be a sequence of random variables with $E[X_i^2] < \infty$ for all $i = 1, \ldots, n$. Prove that

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathrm{Cov}(X_i, X_j).$$

Conclude that if X_i are pairwise uncorrelated, then the variance of their sum is the sum of their variances.

Exercises 8.14, 8.15, and 8.16 on page 299 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 18

1. The moment generating function

Now we introduce a new tool that will be useful for studying sums of independent random variables. (One situation where one encounters a sum of independent random variables is when one studies the sample mean: if $X_1, \dots, X_n$ are independent samples of some quantity, then the sample mean is $\frac{X_1+\cdots+X_n}{n}$.)

The *moment generating function* (mgf) of a random variable X is the function of t given by

$$M(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} f(x), & \text{in the discrete setting,} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, & \text{in the continuous setting.} \end{cases}$$

Sometimes we write $M_X(t)$ to make it clear that we are talking about the mgf of the random variable X.

Note that $M(0)$ always equals 1 (because it is the mean of $e^{0\cdot X} = 1$) and $M(t)$ is always nonnegative (since it is the mean of the nonnegative quantity e^{tX}).

Example 18.1. Suppose X takes the finitely many values $\{0, 1, 2, 3\}$ and has the mass function $f(0) = 1/2$, $f(1) = 1/4$, $f(2) = f(3) = 1/8$. Then its mgf is the function

$$M(t) = E[e^{tX}] = e^{t\cdot 0} f(0) + e^{t\cdot 1} f(1) + e^{t\cdot 2} f(2) + e^{t\cdot 3} f(3) = \frac{1}{2} + \frac{e^t}{4} + \frac{e^{2t} + e^{3t}}{8}.$$

Example 18.2. Suppose X has pdf $f(x) = 2x$ for $0 \leqslant x \leqslant 1$ and 0 otherwise. Then its mgf is the function

$$M(t) = E[e^{tX}] = \int_0^1 e^{tx} \cdot (2x)\, dx.$$

To compute the integral we use integration by parts. Set $u = 2x$ and $v' = e^{tx}$. Then $u' = 2$ and $v = \frac{1}{t} e^{tx}$. So

$$M(t) = \frac{2x}{t} e^{tx} \Big|_0^1 - \frac{2}{t} \int_0^1 e^{tx}\, dx = \frac{2}{t} e^t - \frac{2}{t^2}(e^t - 1).$$

[0] Last modified on May 12, 2020 at 13:10:55 -06'00'

(You may want to check that the formula does give $M(0) = 1$ and then notice that we cannot plug in $t = 0$. But we can take a limit as $t \to 0$ and use de l'Hôpital's rule to find that

$$\lim_{t\to 0} \frac{2(te^t - e^t + 1)}{t^2} = \lim_{t\to 0} \frac{e^t + te^t - e^t + 0}{t} = \lim_{t\to 0} e^t = 1,$$

as it should be.)

Now let us compute the mgf of the familiar random variables.

Example 18.3 (Binomial)**.** If $X \sim \text{Bernoulli}(p)$, then its mgf is

$$M(t) = (1-p)e^{0\cdot t} + pe^{1\cdot t} = 1 - p + pe^t.$$

More generally, if $X \sim \text{Binomial}(n, p)$, then its mgf is

$$M(t) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k},$$

which using the Binomial theorem gives

$$M(t) = (pe^t + 1 - p)^n.$$

Example 18.4 (Negative Binomial)**.** If $X \sim \text{Negative Binomial}(r, p)$, then its mgf is

$$\begin{aligned} M(t) &= \sum_{k=r}^{\infty} e^{tk} \binom{k-1}{r-1} p^r (1-p)^{k-r} = p^r e^{tr} \sum_{k=r}^{\infty} \binom{k-1}{r-1} e^{t(k-r)} (1-p)^{k-r} \\ &= p^r e^{tr} \sum_{k=r}^{\infty} \frac{(k-1)(k-2)\cdots(k-r+1)}{(r-1)!} \big((1-p)e^t\big)^{k-r} \\ &= \frac{p^r e^{tr}}{(r-1)!} \sum_{k=r}^{\infty} (k-1)(k-2)\cdots(k-r+1)\big((1-p)e^t\big)^{k-r} \end{aligned} \tag{18.1}$$

Abbreviate $x = (1-p)e^t$. We will next find a formula for the sum

$$\sum_{k=r}^{\infty} (k-1)(k-2)\cdots(k-r+1)x^{k-r}. \tag{18.2}$$

Note that if k is any integer between 1 and $r-1$, then $(k-1)(k-2)\cdots(k-r+1) = 0$. So in fact, the sum can be started at $k = 1$ instead of at $k = r$.

Next, observe that the derivative of x^{k-1} is $(k-1)x^{k-2}$ and the second derivative is $(k-1)(k-2)x^{k-3}$ and so on. So

$$(k-1)(k-2)\cdots(k-r+1)x^{k-r}$$

is derivative number $r-1$ of x^{k-1}. Consequently, the sum (18.2) is derivative number $r-1$ of $\sum_{k=1}^{\infty} x^{k-1}$ which by the geometric series formula 6.1 equals $\frac{1}{1-x}$. Write this as $(1-x)^{-1}$ and notice that taking one derivative gives $(1-x)^{-2}$, then the second derivative is $2(1-x)^{-3}$, and so on. So taking $r-1$ derivatives gives $(r-1)!(1-x)^{-r}$. Plugging this back into (18.1) (and recalling that $x = (1-p)e^t$) gives

$$M(t) = \frac{p^r e^{tr}}{(r-1)!} \times (r-1)!\big(1-(1-p)e^t\big)^{-r} = \Big(\frac{pe^t}{1-(1-p)e^t}\Big)^r.$$

In the above calculation we used the fact that the derivative of a sum is the sum of derivatives. This fact is obvious when we have finitely many summands. But it is not such a trivial fact when we have infinitely many summands, as above. Nevertheless, this fact is true for geometric series

as long as we are in the converging regime. That is, when $|x| < 1$. Otherwise, when $|x| \geqslant 1$, the answer is infinite. So the above formula we found for the mgf is only valid when $(1-p)e^t < 1$, i.e. when $t < -\log(1-p)$. To summarize: the mgf of a Negative Binomial(r,p) is

$$M(t) = \Big(\frac{pe^t}{1-(1-p)e^t}\Big)^r \quad \text{for } t < -\log(1-p) \quad \text{and } \infty \text{ for } t \geqslant -\log(1-p).$$

As a bonus, plugging in $t = 0$ gives $M(0) = 1$. This proves that the mass function we derived for the negative binomial (in an earlier lecture) does indeed add up to 1. Back when we introduced the negative binomial we only argued for this fact intuitively. Now we have a proof.

As a special case, if X is Geometric(p) then we can apply the above with $r = 1$ to get

$$M(t) = \frac{pe^t}{1-(1-p)e^t} \quad \text{for } t < -\log(1-p) \quad \text{and } \infty \text{ for } t \geqslant -\log(1-p).$$

Example 18.5 (Poisson)**.** If $X \sim \text{Poisson}(\lambda)$, then

$$M(t) = E\left[e^{tX}\right] = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!}.$$

The sum gives the Taylor expansion of $e^{\lambda e^t}$. Therefore,

$$M(t) = e^{\lambda(e^t-1)}.$$

This completes the computation of the moment generating functions for the familiar discrete random variables. Next, we look into the continuous random variables.

Example 18.6 (Uniform)**.** If $X \sim \text{Uniform}(a,b)$, then

$$M(t) = E[e^{tX}] = \frac{1}{b-a}\int_a^b e^{tx}\,dx = \frac{e^{bt}-e^{at}}{(b-a)t}.$$

(Note again that although we cannot plug in $t = 0$ on the right-hand side we can take a limit $t \to 0$ and use de l'Hôpital's rule to get that $M(t) \to 1$ as $t \to 0$.)

Example 18.7 (Exponential)**.** If $X \sim \text{Exponential}(\lambda)$, 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

$$M(t) = \frac{\lambda}{\lambda - t}, \text{ for } t < \lambda.$$

Example 18.8 (Normal)**.** If $X = N(\mu, \sigma^2)$, then

$$\begin{aligned}
M(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/(2\sigma^2)}\,dx \\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2-2(\sigma^2t+\mu)x+(\sigma^2t+\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\sigma^2t-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{e^{\mu t+\sigma^2t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\,du \qquad (u = (x-\sigma^2t-\mu)/\sigma) \\
&= e^{\mu t+\sigma^2t^2/2}. \qquad \text{(The integral is that of the pdf of a standard normal and thus } = 1.)
\end{aligned}$$

In particular, the mgf of a standard normal N(0,1) is

$$M(t) = e^{t^2/2}.$$

2. Relation of the mgf to moments

The quantity $E[X^n]$ is called the n-th moment of a random variable X. These quantities are useful for various things. For example, $E[X]$ is the mean and $E[X^2]$ is related to the variance. The mgf can help compute these moments, hence the name "moment generating function".

Indeed, suppose X is a continuous random variable with moment generating function M and density function f. Then

$$M'(t) = \frac{d}{dt}\left(E[e^{tX}]\right) = \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

Now, using tools from mathematical analysis (that are beyond the scope of this course) one can show that under certain conditions (that happen to be satisfied here) we can move the derivative under the integral sign. That is, we can continue the above computation by:

$$M'(t) = \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} x e^{tx} f(x)\, dx = E\left[Xe^{tX}\right].$$

The same end-result holds if X is discrete with mass function f, but now

$$M'(t) = \sum_x x e^{tx} f(x) = E\left[Xe^{tX}\right].$$

Therefore, eitherway:

$$M'(0) = E[X].$$

Taking more derivatives gives

$$M^{(n)}(t) = E\left[X^n e^{tX}\right]$$

and therefore

$$M^{(n)}(0) = E\left[X^n\right].$$

(Here, $M^{(n)}$ denotes the n-th derivative of M.)

This relation of the mgf to moments is useful for example when we know the formula for the mgf and would like to compute the variance. Then we can write

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = M''(0) - (M'(0))^2.$$

See Example 19.2 for a situation where we compute the mean and the variance of a random variable this way.

Example 18.9 (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,

$$M'(t) = \frac{te^t - e^t + 1}{t^2} \quad \text{and} \quad M''(t) = \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3},$$

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,

$$\mathrm{E}\left[X^2\right] = \lim_{t\searrow 0} \frac{t^2e^t - 2te^t + 2e^t - 2}{t^3} = \lim_{t\searrow 0} \frac{t^2e^t}{3t^2} = \frac{1}{3}.$$

Alternatively, we could have done a direct computation:

$$\mathrm{E}[X^n] = \int_0^1 x^n \, dx = \frac{1}{n+1}\,.$$

Example 18.10 (Standard Normal)**.** If $X \sim \mathrm{N}(0,1)$, then we have seen that $M(t) = e^{t^2/2}$. Thus, $M'(t) = te^{t^2/2}$ and $\mathrm{E}[X] = M'(0) = 0$. Also, $M''(t) = (t^2+1)e^{t^2/2}$ and $\mathrm{E}[X^2] = M''(0) = 1$. A homework exercise asks to continue this procedure and compute $\mathrm{E}[X^n]$ for all positive integers n.

3. Identifying random variables using mgfs

The following theorem gives one use of moment generating functions.

Theorem 18.11. *If X and Y are two random variables—discrete* or *continuous—with moment generating functions M_X and M_Y that are $< \infty$ on a neighborhood containing 0, and if $M_X(t) = M_Y(t)$ for all t, then 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 probability density functions are the same.*

We omit the proof as it requires knowing quite a bit of analysis, beyond the level of this course. 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 18.12. 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

$$f(x) = \begin{cases} 1/2 & \text{if } x = 1, \\ 1/4 & \text{if } x = -\pi \text{ or } x = \sqrt{2}, \\ 0 & \text{otherwise.} \end{cases}$$

4. Sums of independent random variables

Here is the reason for which we introduced moment generating functions in this course.

Theorem 18.13. *If $X_1, \ldots, X_n$ are independent, with moment generating functions $M_{X_1}, \ldots, M_{X_n}$, respectively, 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

$$M_{X_1+X_2}(t) = \mathrm{E}\left[e^{t(X_1+X_2)}\right] = \mathrm{E}\left[e^{tX_1} \times e^{tX_2}\right].$$

Since X_1 and X_2 are independent and the above is computing the expected value of a function of X_1 times a function of X_2, the expected value splits into a product of the two expected values $E[e^{tX_1}]$ and $E[e^{tX_2}]$ and we get the desired result:

$$M_{X_1+X_2}(t) = E\left[e^{t(X_1+X_2)}\right] = E\left[e^{tX_1}\right] \times E\left[e^{tX_2}\right] = M_{X_1}(t)M_{X_2}(t). \qquad \square$$

One way the above theorem is used is in conjunction with Theorem 18.11. Namely, suppose we have some independent random variables $X_1, \ldots, X_n$ and we know the individual distributions of each of them. Then this completely determines the joint distribution of $(X_1, \ldots, X_n)$ and we should be able to describe the distribution of any combination of them. This may be a complicated task for some general combination of the random variables. But if we are interested in the sum $X_1 + \cdots + X_n$, then we may be able to identify the distribution of this sum by computing its mgf using Theorem 18.13 and then looking at the result and trying to identify the distribution that has that mgf, appealing to Theorem 18.11. Here are a few instances of this situation.

Example 18.14 (Sum of Binomials)**.** Suppose $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent. Then using the formulas we already derived for the mgf of a Binomial, we get

$$M_{X+Y}(t) = M_X(t)M_Y(t) = (1 - p + pe^t)^n(1 - p + pe^t)^m = (1 - p + pe^t)^{n+m},$$

which we identify as the mgf of yet another Binomial$(n + m, p)$. So the upshot is that adding two independent binomial random variables with the same success probability p gives a binomial random variable with that same success probability p, but with the parameter that indicates the number of trials being the sum of the two parameters of the two random variables being added. This is logical: X models the number of heads when tossing n p-coins and Y models the number of heads when tossing m p-coins. If X and Y are independent then $X + Y$ is like tossing $n + m$ p-coins and looking at the total number of successes. So this is a Binomial$(n + m, p)$.

Of course, by induction, we can add more than two independent binomials, as long as they share the same parameter p, and then we would get a binomial with that parameter p and the parameter that indicates the number of trials would be the sum of the corresponding parameters for the binomials being added. In particular, we can think of a Binomial(n, p) as a sum of n independent Bernoulli(p) random variables. Again, this makes sense: the experiment where we toss n independent p-coins and count the number of heads is the same as the experiment where we toss one p-coin n times and add the total number of heads!

One consequence of the above observation is that the mean of a Binomial(n, p) is the sum of the means of n Bernoulli(p) random variables. The mean of a Bernoulli(p) is easily computed to be $0 \times (1 - p) + 1 \times p = p$ and so the mean of a Binomial(n, p) is p added n times, which is np. This is a much faster way to compute the mean of a Binomial than we had when we introduced Binomials! Similarly, we know that the variance of the sum of independent random variables is the sum of the variances. So the variance of a Binomial(n, p) is the sum of the variances of n Bernoulli(p) random variables. The variance of a Bernoulli(p) is easily computed to be $p - p^2 = p(1 - p)$ and therefore the variance of a Binomial(n, p) is $np(1 - p)$. Again, this is a much faster way to compute the variance of a Binomial than we did before!

Example 18.15 (Sum of Negative Binomials)**.** Suppose $X \sim \text{Negative Binomial}(r, p)$ and $Y \sim \text{Negative Binomial}(s, p)$ are independent. Then using the formulas we already derived for the mgf of a Negative Binomial, we get

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \left(\frac{pe^t}{1 - (1 - p)e^t}\right)^r \left(\frac{pe^t}{1 - (1 - p)e^t}\right)^s = \left(\frac{pe^t}{1 - (1 - p)e^t}\right)^{r+s},$$

when $t < -\log(1-p)$ and $M_{X+Y}(t) = M_X(t)M_Y(t) = \infty \cdot \infty = \infty$ when $t \geqslant -\log(1-p)$. We see that this formula identifies the mgf of yet another Negative Binomial$(r+s, p)$. So the upshot is that adding two independent negative binomial random variables with the same success probability p gives a negative binomial random variable with that same success probability p, but with the parameter that indicates the number of successes we are waiting for being the sum of the two parameters of the two random variables being added. This is logical: X models the number of times we need to toss a p-coin to get r successes and Y models the number of times we need to toss a p-coin to get s successes. If X and Y are independent then $X + Y$ tells us the number of times we need to toss a p-coin until we get $r+s$ successes. So this is a Negative Binomial$(r+s, p)$.

Of course, by induction, we can add more than two independent negative binomials, as long as they share the same parameter p, and then we would get a negative binomial with that parameter p and the parameter that indicates the number of successes we are waiting for would be the sum of the corresponding parameters for the negative binomials being added. In particular, we can think of a Negative Binomial(r, p) as a sum of r independent Geometric(p) random variables. Again, this makes sense: we can think of the experiment where we toss a p-coin until we get r heads as repeating r independent experiments where in each we toss the p-coin until we get one success.

Similarly to the previous example, we see that it is not a coincidence that the mean of a Negative Binomial(r, p) is r/p which is r times the mean of a Geometric(p) (which is $1/p$). More importantly, when we computed variances of the various random variables, we did not compute the variance of a Negative Binomial(r, p). But we now know that it equals r times the variance of a Geometric(p), which we computed to be $(1-p)/p^2$. Therefore, we now know that

$$\operatorname{Var}\big(\text{Negative Binomial}(r, p)\big) = \frac{(1-p)r}{p^2}.$$

Example 18.16 (Sum of Poissons)**.** Now suppose $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\gamma)$ are independent. Then

$$M_{X+Y}(t) = e^{\lambda(e^t-1)}e^{\gamma(e^t-1)} = e^{(\lambda+\gamma)(e^t-1)},$$

which we identify as the mgf of a Poisson with parameter $\lambda + \gamma$. Thus, we deduce that $X + Y \sim \text{Poisson}(\lambda + \gamma)$.

Again, by induction we can do this for more than two variables. The result, in words, says that the sum of independent Poisson random variables is a Poisson random variable with parameter being the sum of the parameters of the random variables being added.

This is again intuitively clear: if we have multiple lines in a store being formed independently then adding the lengths of the lines together is like asking everyone to move to one single line. This should thus be modeled by a Poisson random variable. To figure out the parameter we recall that the parameter is precisely the mean of the random variable. And the mean of the new single queue is clearly the sum of the means of the original queues. Hence, the parameter of the sum is the sum of the parameters.

Example 18.17 (Sum of Normals)**.** Suppose we have n normal random variables: $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. Then

$$\begin{aligned} M_{X_1+\cdots+X_n}(t) &= M_{X_1}(t)\cdots M_{X_n}(t) = \exp\{\mu_1 t + \sigma_1^2 t^2/2\} \times \cdots \times \exp\{\mu_n t + \sigma_n^2 t^2/2\} \\ &= \exp\{(\mu_1 + \cdots + \mu_n)t + (\sigma_1^2 + \cdots + \sigma_n^2)t^2/2\}. \end{aligned}$$

This is the mgf of a normal with mean $\mu = \mu_1 + \cdots + \mu_n$ and variance $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$. So the sum of independent normals is another normal. The fact that the mean of the sum is the sum

of the means is not surprising. Also, the fact that the variance of the sum (of the independent normals) is the sum of the variances should not be surprising.

In the above examples we saw that the sum of independent Binomials (with the same success probability) is a Binomial, the sum of independent Negative Binomials (with the same success probability) is a Negative Binomial, the sum of independent Poissons is a Poisson, and the sum of independent Normals is a Normal. What about the sum of independent Exponential random variables?

Example 18.18 (Sum of Exponentials)**.** Let X and Y be two independent Exponential(λ) random variables. Then

$$M_{X+Y}(t) = \frac{t}{\lambda - t} \times \frac{t}{\lambda - t} = \left(\frac{t}{\lambda - t}\right)^2$$

for $t < \lambda$ and $M_{X+Y}(t) = \infty \times \infty = \infty$ when $t \geqslant \lambda$. This is not the mgf of an exponential. So the sum of two independent exponentials (even with the same parameter) is not an exponential random variable. But what is then that random variable? This will be resolved in the next lecture.

Read sections 5.1 and 8.3 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 18.1. 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) $M(t) = 1 - t$ for all t.

(b) $M(t) = 2e^{-t}$ for all t.

(c) $M(t) = 1/(1-t)$ for $t < 1$ and $M(t) = \infty$ for $t \geqslant 1$.

(d) $M(t) = \frac{1}{3} + \frac{1}{2}e^{2t} + \frac{1}{12}e^{-2t} + \frac{1}{12}e^{13t}$ for all t.

Exercise 18.2. 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 18.3. Show that if $Y = aX + b$, with nonrandom constants a and b, then

$$M_Y(t) = e^{bt} M_X(at).$$

Exercise 18.4. Let X and Y take only the values 0, 1, or 2. Explain why 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 18.11.

Exercises 5.1, 5.2, 5.4, 8.10, and 8.13 on pages 197, 198, 298, and 299 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 19

1. Gamma random variables

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

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

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 ∞). Performing the change of variables $y = \lambda x$ we get

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

Just like it was the case for $\Phi(a) = \int_{-\infty}^a e^{-z^2/2}\,dz$, one can prove that there is "no nice formula" that "describes" $\int_0^\infty x^{\alpha-1}e^{-\lambda x}\,dx$ for all α in terms of other known functions (a theorem due to Liouville). Therefore, the best we can do is to define a new function, called the *gamma function*, as

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

Then, the gamma density with parameters $\alpha, \lambda > 0$ is the pdf:

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

A random variable with this pdf is called a Gamma(α, λ) random variable.

It turns out that $\Gamma(\alpha)$ is computable for some reasonable values of $\alpha > 0$. The key to unraveling this remark is the following "recursive property":

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \qquad \text{for all } \alpha > 0. \tag{19.1}$$

[0] Last modified on May 12, 2020 at 13:12:03 -06'00'

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 $v'(x) = e^{-x}$. 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.

Next, it easy to see that $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$. Therefore, $\Gamma(2) = 1\times\Gamma(1) = 1$, $\Gamma(3) = 2\times\Gamma(2) = 2$, ..., and in general,

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

(This is why sometimes it is said that the Γ function extends the factorial to non-integer values.) It is also not too hard to see that

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty x^{-1/2}e^{-x}\,dx = \sqrt{2}\int_0^\infty e^{-y^2/2}\,dy \\ &= \frac{\sqrt{2}}{2}\cdot\sqrt{2\pi}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-y^2/2}\,dy \\ &= \frac{\sqrt{2}}{2}\cdot\sqrt{2\pi}\cdot 1 = \sqrt{\pi}.\end{aligned}$$

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 $\Gamma(\alpha)$ cannot be computed explicitly for a general α, it is quite easy to compute for α's that are are half nonnegative integers.

Notice that if $\alpha = 1$ then because $\Gamma(1) = 1$ we have that a Gamma$(1,\lambda)$ random variable has the pdf $\lambda e^{-\lambda x}$ when $x > 0$ and 0 when $x \leqslant 0$. So a Gamma$(1,\lambda)$ random variable is actually an Exponential(λ) random variable! In other words, the Gamma distribution is a generalization of the Exponential distribution.

One may recall that exponential random variables can be derived as limits of geometric random variables. Is there something similar for Gamma random variables? The answer is yes: Gamma random variables can be derived as limits of Negative Binomial random variables! We will see how this works in an exercise in the next lecture.

Now that we have a new distribution, let us compute its mean, variance, and moment generating function. We could of course compute the mean and the variance by computing the appropriate integrals directly. But we will instead start with the moment generating function and then compute the mean and the variance by following the idea in Section 2.

Example 19.1 (mgf of a Gamma)**.** If $X \sim \text{Gamma}(\alpha,\lambda)$, 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,

$$M(t) = \Big(\frac{\lambda}{\lambda-t}\Big)^\alpha \text{ if } t < \lambda \quad \text{and} \quad M(t) = \infty \text{ if } t \geqslant \lambda.$$

In particular, if $\alpha = 1$ we recover the formula we already knew for the mgf of an Exponential(λ):

$$M(t) = \frac{\lambda}{\lambda-t} \text{ if } t < \lambda \quad \text{and} \quad M(t) = \infty \text{ if } t \geqslant \lambda.$$

Example 19.2 (Mean and variance of a Gamma)**.** Now that we computed the mgf of a Gamma(α, λ) random variable X, we can apply the ideas from Section 2 to get $E[X]$ and $E[X^2]$. For this write $M(t) = \lambda^\alpha(\lambda-t)^{-\alpha}$ and compute

$$M'(t) = \lambda^\alpha \cdot \alpha(\lambda-t)^{-\alpha-1}$$

and

$$M''(t) = \lambda^\alpha \cdot \alpha(\alpha+1)(\lambda-t)^{-\alpha-2}.$$

This gives

$$E[X] = \lambda^\alpha \cdot \alpha \cdot \lambda^{-\alpha-1} = \frac{\alpha}{\lambda}$$

and

$$E[X^2] = \lambda^\alpha \cdot \alpha(\alpha+1)\lambda^{-\alpha-2} = \frac{\alpha(\alpha+1)}{\lambda^2},$$

which leads to

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

Example 19.3 (Sum of Gammas)**.** Suppose now $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ and the two are independent. Note that the two share the parameter that goes in the exponential term in the pdf but may have different power of x parameter. Then

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \Big(\frac{\lambda}{\lambda-t}\Big)^\alpha \cdot \Big(\frac{\lambda}{\lambda-t}\Big)^\beta = \Big(\frac{\lambda}{\lambda-t}\Big)^{\alpha+\beta},$$

for $t < \lambda$, and $M_{X+Y}(t) = \infty \cdot \infty$ for $t \geqslant \lambda$. This is the mgf of another Gamma with parameters $\alpha+\beta$ and λ.

As usual, induction allows us to extend this to more than two random variables and we get that the sum of independent Gamma random variables that share the λ parameter is another Gamma random variable with this same λ parameter and the other parameter is the sum of the corresponding parameters in the Gamma random variables being added. In particular, the sum of n independent Exponential(λ) random variables is a Gamma(n, λ) random variable (because the Exponential(λ) is a Gamma$(1, \lambda)$ and we are adding n of these).

Recall how Exponential random variables are used to model waiting times. The above connection to Gamma random variables explains why such random variables could be used to model more general service times, where you have to go through several servers before you are done.

Each server keeps you waiting for an Exponential random variable. If the rate of service at all the servers is the same, then the total service time is the sum of n independent Exponential random variables (where n is the number of servers) and so the total service time is a Gamma random variable. Gamma random variables are also often used to model commute times. We will later also see that the Gamma distribution is related to another distribution that appears very often in statistical applications, namely to the Chi square distribution.

Homework Problems[1]

Exercise 19.1. The magnitude of earthquakes recorded in a certain region are modeled with an exponential distribution having mean of 2.5 on the Richter scale.

(a) What is the probability that the next earthquake in the region will exceed 3.0 on the Richter scale?

(b) Assuming successive earthquakes have independent sizes, what is the probability that out of the next five earthquakes to strike this region, at least one will exceed 6.0 on the Richter scale?

Exercise 19.2. Four-week Summer rainfall totals in a section of the Midwest can be modeled by a Gamma distribution with parameters $\alpha = 1.6$ and $\lambda = 0.5$. Find the mean and variance of the four-week totals.

Exercise 19.3. The response times on an online computer terminal are modeled using a Gamma distribution with mean 4 seconds and variance 8 seconds. Write the probability density function for the response times.

Exercise 19.4. Annual incomes for household heads in a certain section of a city are modeled by a Gamma distribution with parameters $\alpha = 1000$ and $\lambda = 0.05$. Find the mean and variance of these incomes.

Exercise 19.5. The weekly amount of "downtime" X (in hours) for a certain industrial machine has a Gamma distribution with parameters $\alpha = 3$ and $\lambda = 0.5$. The loss, in dollars, to the industrial operation as a result of this downtime is given by $L = 30X + 2X^2$. Find the expected value and variance of L.

[1] These exercises are due to Pat Rossi.

Lecture 20

1. Distributional convergence

One way to say that two discrete random variables are *statistically close* to each other is to say that their mass functions are close. For continuous random variables we cannot talk about mass functions, but one way to say that two such random variables are statistically close would be by saying that the quantities $P(a \leqslant X \leqslant b)$ and $P(a \leqslant Y \leqslant b)$ are close, for all $a < b$. We have already used these ideas when we discussed approximating Binomial random variables by Poisson random variables (the law of rare events) and by Normal random variables (the central limit theorem).

According to Theorem 18.11, moment generating functions determine the statistics of random variables (i.e. uniquely determine the mass function in the discrete case and the pdf in the continuous case). Hence, it stands reason that an alternate (perhaps easier to handle) way to say that two random variables are close would be by saying that their moment generating functions are close. This is exactly what Lévy's continuity theorem states. The following are two versions of this theorem. We do not give proofs as they require tools from mathematical analysis which are beyond the scope of this course.

Theorem 20.1. *Let X_n be a sequence of discrete random variables with moment generating functions M_n. Also, let X be a discrete random variable with moment generating function M. Suppose that there exists a $\delta > 0$ such that for all $t \in (-\delta, \delta)$ we have $M(t) < \infty$ and $\lim_{n\to\infty} M_n(t) = M(t)$. Then*

$$\lim_{n\to\infty} P(X_n = a) = P(X = a) \quad \textit{for all } a.$$

(*That is: the mass function of X_n converges to the mass function of X.*)

Here is an immediate application of this theorem.

Example 20.2 (Law of rare events, rebooted)**.** Suppose $X_n \sim \text{Binomial}(n, \lambda/n)$, where $\lambda > 0$ is fixed, and $n \geqslant \lambda$. Then, for all t

$$\begin{aligned} M_{X_n}(t) &= \left(1 - p + pe^t\right)^n = \left(1 - \frac{\lambda}{n} + \frac{\lambda e^t}{n}\right)^n \\ &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^{\lambda(e^t-1)}. \end{aligned}$$

[0] Last modified on April 22, 2020 at 23:24:59 -06'00'

Note that the limit equals $M_X(t)$, where $X \sim \text{Poisson}(\lambda)$. Therefore, by Theorem 20.1, the mass function of X_n converges to the mass function of X. We have thus given a second proof of the law of rare events, which says that one can approximate a Binomial, with a large number of trials n and a very small probability of success p (λ/n in the above theorem), by a Poisson with parameter np.

Theorem 20.1 covered the case of discrete random variables converging in distribution to a limit that itself is also a discrete random variable. The next theorem is the version where the random variables converge to a continuous random variable. In this case, the random variables that are converging do not have to be continuous. They can be continuous or discrete. What matters is that the limit is a continuous random variable.

Theorem 20.3. *Let X_n be a sequence of random variables—discrete* or *continuous—with moment generating functions M_n. Also, let X be a continuous random variable with moment generating function M. Suppose that there exists a $\delta > 0$ such that for all $t \in (-\delta, \delta)$ we have $M(t) < \infty$ and $\lim_{n\to\infty} M_n(t) = M(t)$. Then*

$$\lim_{n\to\infty} P(X_n \leqslant b) = \lim_{n\to\infty} P(X_n < b) = P(X \leqslant b) \quad \textit{for all } b.$$

A few remarks are in order.

Remark 20.4. Note how both Theorems 20.1 and 20.3 only ask for convergence of the moment generating functions for all t in an interval around 0. Convergence for all t is not necessary although, surprisingly, once one has convergence over an interval around 0 it turns out that one can prove the convergence for all t!

Remark 20.5. Since X is a continuous random variable, the probabilities $P(X \leqslant b)$ and $P(X < b)$ are actually the same, namely they are both equal to $\int_{-\infty}^{b} f_X(x)\,dx$. Similarly, $P(X \geqslant b)$ and $P(X > b)$ is the same number and also

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

So for a continuous random variable, it does not matter whether we include or exclude any of the endpoints of the interval.

Remark 20.6. If X_n is not a continuous random variable, then it may be in principle that $P(X_n \leqslant b)$ is not the same as $P(X_n < b)$. In fact, if b is one of the possible values that X_n can take with a positive probability, then $P(X_n \leqslant b) = P(X_n < b) + P(X_n = b) > P(X_n < b)$. And yet, the theorem says that in the limit as $n \to \infty$ both $P(X_n \leqslant b)$ and $P(X_n < b)$ will converge to the same limit $P(X \leqslant b)$.

Remark 20.7. The above theorem says that probabilities like $P(X_n \leqslant b)$ and $P(X_n < b)$ are close to the probability $P(X \leqslant b)$. But what about other types of probabilities? To this end, note that

$$P(X_n > b) = 1 - P(X_n \leqslant b).$$

Therefore, the above theorem also implies that

$$\lim_{n\to\infty} P(X_n > b) = \lim_{n\to\infty} \big(1 - P(X_n \leqslant b)\big) = 1 - P(X \leqslant b) = P(X > b).$$

Similarly,

$$P(X_n \geqslant b) = 1 - P(X_n > b)$$

and the theorem implies that

$$\lim_{n\to\infty} P(X_n \geqslant b) = \lim_{n\to\infty} \big(1 - P(X_n < b)\big) = 1 - P(X \leqslant b) = P(X > b).$$

(So again, the limit of $P(X_n > b)$ and $P(X_n \geqslant b)$ is the same number $P(X > b)$m, which actually equals $P(X \geqslant b)$ since X is a continuous random variable.)

Also

$$P(a < X_n \leqslant b) = P(X_n \leqslant b) - P(X_n \leqslant a)$$

and so we have

$$\begin{aligned}\lim_{n\to\infty} P(a < X_n \leqslant b) &= \lim_{n\to\infty} P(X_n \leqslant b) - \lim_{n\to\infty} P(X_n \leqslant a) \\ &= P(X \leqslant b) - P(X \leqslant a) = P(a < X \leqslant b).\end{aligned}$$

The same thing holds for the limits of $P(a \leqslant X_n < b)$, $P(a \leqslant X_n \leqslant b)$, and $P(a < X_n < b)$.

The upshot is that the theorem in fact implies that if the moment generating function of X_n converges to the moment generating function of X, then we can approximate the probabilities that X_n is less than a, bigger than b, or between a and b by the corresponding probabilities for X. This says that X_n becomes statistically very close to X, as n grows.

Here is an application of the above theorem. Recall that the De Moivre-Laplace central limit theorem says, in words, that for a fixed probability of success $p \in (0,1)$, as the number of trials n grows large a Binomial(n,p) gets closer and closer, statistically, to a normal with the same mean and variance. We can now make this mathematically precise and give a proof.

Recall that if Z is a standard normal random variable, then

$$P(Z \leqslant b) = \Phi(b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-z^2/2}\, dz.$$

Theorem 20.8 (The De Moivre–Laplace central limit theorem (CLT), 1733)**.** *Suppose* $0 < p < 1$ *is fixed. For each positive integer* n *let* X_n *be a Binomial*(n,p) *random variable. Then, we have for all* b

$$\lim_{n\to\infty} P\Big\{ \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b \Big\} = \Phi(b).$$

As mentioned in the above remarks, the above limit implies similar limits for the probabilities of $\big\{\frac{X_n-np}{\sqrt{np(1-p)}} < b\big\}$, $\big\{\frac{X_n-np}{\sqrt{np(1-p)}} \geqslant a\big\}$, $\big\{\frac{X_n-np}{\sqrt{np(1-p)}} > a\big\}$, and also $\big\{a \leqslant \frac{X_n-np}{\sqrt{np(1-p)}} \leqslant b\big\}$, $\big\{a < \frac{X_n-np}{\sqrt{np(1-p)}} \leqslant b\big\}$, $\big\{a \leqslant \frac{X_n-np}{\sqrt{np(1-p)}} < b\big\}$, and $\big\{a < \frac{X_n-np}{\sqrt{np(1-p)}} < b\big\}$.

Proof of the De Moivre-Laplace CLT. Let

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}.$$

Then the claim of the theorem is that $P(Z_n \leqslant b)$ converges to $P(Z \leqslant b)$, where Z is a standard normal random variable. By Theorem 20.3, this claim would follow from proving that the moment generating function of Z_n converges as $n \to \infty$ to the moment generating function of Z.

We know that for all real numbers t, $M_{X_n}(t) = \left(1-p+pe^t\right)^n$. We can use this to compute M_{Z_n} as follows:

$$\begin{aligned}
M_{Z_n}(t) &= E\Big[\exp\Big\{t\cdot\frac{X_n-np}{\sqrt{np(1-p)}}\Big\}\Big] = E\Big[e^{-npt/\sqrt{np(1-p)}}\exp\Big\{\frac{t}{\sqrt{np(1-p)}}\cdot X_n\Big\}\Big] \\
&= e^{-npt/\sqrt{np(1-p)}}M_{X_n}\Big(\frac{t}{\sqrt{np(1-p)}}\Big) \\
&= \Big(e^{-pt/\sqrt{np(1-p)}}\Big)^n\Big(1-p+pe^{t/\sqrt{np(1-p)}}\Big)^n \\
&= \Big((1-p)e^{-t\sqrt{\frac{p}{n(1-p)}}}+pe^{t\sqrt{\frac{1-p}{np}}}\Big)^n.
\end{aligned}$$

Let $h = 1/\sqrt{n}$ and rewrite the above as

$$M_{Z_n}(t) = \exp\left\{\frac{\log\Big((1-p)e^{-ht\sqrt{\frac{p}{1-p}}}+pe^{ht\sqrt{\frac{(1-p)}{p}}}\Big)}{h^2}\right\}.$$

We want to take $n\to\infty$ which means taking $h\to 0$. If we plug in $h=0$ in the above we will get $0/0$. So we use de l'Hôpital's rule to get that

$$\begin{aligned}
\lim_{n\to\infty} M_{Z_n}(t) &= \exp\left\{\lim_{h\to 0}\frac{\log\Big((1-p)e^{-ht\sqrt{\frac{p}{1-p}}}+pe^{ht\sqrt{\frac{(1-p)}{p}}}\Big)}{h^2}\right\} \\
&= \exp\left\{\lim_{h\to 0}\frac{\Big[\log\Big((1-p)e^{-ht\sqrt{\frac{p}{1-p}}}+pe^{ht\sqrt{\frac{(1-p)}{p}}}\Big)\Big]'}{2h}\right\} \\
&= \exp\left\{\lim_{h\to 0}\frac{-t\sqrt{p(1-p)}e^{-ht\sqrt{\frac{p}{1-p}}}+t\sqrt{p(1-p)}e^{ht\sqrt{\frac{(1-p)}{p}}}}{2h\Big[(1-p)e^{-t\sqrt{\frac{ph}{1-p}}}+pe^{t\sqrt{\frac{(1-p)h}{p}}}\Big]}\right\}.
\end{aligned}$$

Note that plugging in $h=0$ in the term in square brackets in the denominator gives $p+1-p=1$. So we can ignore this term when computing the limit. For the rest, if we plug in $h=0$ we again get $0/0$ and so we can use de l'Hôpital's rule again (after ignoring the term in brackets in the denominator; this simplifies the derivatives!) to get

$$\begin{aligned}
\lim_{n\to\infty} M_{Z_n}(t) &= \exp\left\{\lim_{h\to 0}\frac{\Big[-t\sqrt{p(1-p)}e^{-ht\sqrt{\frac{p}{1-p}}}+t\sqrt{p(1-p)}e^{ht\sqrt{\frac{(1-p)}{p}}}\Big]'}{2}\right\} \\
&= \exp\left\{\lim_{h\to 0}\frac{pt^2e^{-ht\sqrt{\frac{p}{1-p}}}+(1-p)t^2e^{ht\sqrt{\frac{(1-p)}{p}}}}{2}\right\} = e^{t^2/2}.
\end{aligned}$$

We recognize the right-hand side as the moment generating function of a standard normal. We have therefore shown that the moment generating function of Z_n converges to that of a standard normal, which is what we were after. □

Homework Problems

Exercise 20.1. Let $X_n \sim$ Negative Binomial$(r, \lambda/n)$. Show that X_n/n converges in distribution to a Gamma(r, λ). Note that when $r = 1$ this gives a new proof of an old fact: a Geometric(λ/n) converges in distribution to an Exponential(λ).

Hint: start by showing that $M_{X_n/n}(t) = M_{X_n}(t/n)$.

Exercise 20.2. Let $X_1, \ldots, X_n$ 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 of some random variable?

Exercise 20.3. Let $X_1, \ldots, X_n$ be independent Gamma(r, λ) random variables. What is the distribution of $X_1 + \cdots + X_n$? What is the moment generating function of $(X_1 + \cdots + X_n - nr/\lambda)/\sqrt{nr/\lambda^2}$? Find the limit of this function as $n \to \infty$. Can you recognize the outcome as a moment generating function of some random variable?

Lecture 21

1. The Central Limit Theorem

Now, we are ready to give the general version of the central limit theorem, one of the most important and used theorems in probability and statistics. In fact, the name of the theorem is to suggest that it is a theorem that is central in probability theory, like the fundamental theorem of calculus and the fundamental theorem of algebra. (It is a limit theorem because in its statement the size n of the sample is taken to infinity.) In particular, we will see why the normal distribution so important.

The point of the central limit theorem (CLT) is to estimate the distribution of a sum of a large number of independent random variables that all have the same distribution. More precisely, we have a sequence (also called a time series) $X_1, \dots, X_n$ of independent random variables that all have the same distribution. This means that they are all discrete and have the same mass function or all continuous and have the same pdf. For example, these could be measurements from a random sample of size n. Such a sequence is called a sequence of independent identically distributed random variables and abbreviated using the acronym i.i.d. We are interested in approximating probabilities like $P(a \leqslant \overline{X} \leqslant b)$, where

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

is the sample mean. These translate to probabilities of the type

$$P(na \leqslant X_1 + \cdots + X_n \leqslant nb).$$

The CLT allows to approximate such probabilities.

One would expect the sample mean $\overline{X}$ to be close to the true mean $\mu = E[X_1]$. We will prove this fact, called the law of large numbers (LLN), at the end of this lecture. But for now, let us take it for granted and build some intuition for the CLT. (Note by the way that $E[X_1] = E[X_2] = E[X_3] = \cdots$ because we are assuming all these random variables have the same distribution, so they will have the same mean.)

So the law of large numbers says that $X_1 + \cdots + X_n$ is close to $n\mu$. But how close? To get an idea of the size of the difference between the two quantities let us compute its variance. Recall

[0] Last modified on November 04, 2020 at 11:31:11 -07'00'

that additive constants do not affect the variance and that the variance of a sum of independent random variables is the sum of their variances. So

$$\mathrm{Var}(X_1 + \cdots + X_n - n\mu) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2,$$

where $\sigma^2 = \mathrm{Var}(X_1)$ (which by the way equals $\mathrm{Var}(X_2)$ and $\mathrm{Var}(X_3)$ and so on, because the X_i's have the same distribution).

The above computation tells us that the size of $X_1 + \cdots + X_n - n\mu$ is roughly like $\sqrt{n\sigma^2}$ (remember that the variance is like the square of a distance). The central limit theorem tells us what the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}$$

converges to as $n \to \infty$. Here is the precise statement.

Theorem 21.1 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n, \ldots$ be independent identically distributed random variables. Assume $\sigma^2 = Var(X_1)$ is finite. Let $\mu = \mathrm{E}[X_1]$. Then*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}$$

converges in distribution, as $n \to \infty$, to a standard normal random variable. More precisely,

$$\lim_{n\to\infty} \mathrm{P}\Big\{ \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \leqslant b \Big\} = \Phi(b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-z^2/2}\, dz.$$

As we have seen before, similar limits hold for the probabilities of $\big\{\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} < b\big\}$, $\big\{\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} \geqslant a\big\}$, $\big\{\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} > a\big\}$, and also the probabilities of $\big\{a \leqslant \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} \leqslant b\big\}$, $\big\{a < \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} \leqslant b\big\}$, $\big\{a \leqslant \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} < b\big\}$, and $\big\{a < \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} < b\big\}$.

In words, the CLT says that the sum $X_1+\cdots+X_n$ becomes statistically very close to a Normal with the same mean and variance, i.e. to a Normal$(n\mu, n\sigma^2)$, as n grows large.

Note that the ratio in question can be rewritten in terms of the sample mean as

$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

The mean of $\overline{X}$ equals

$$\mathrm{E}[\overline{X}] = \mathrm{E}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{1}{n}\mathrm{E}[X_1 + \cdots + X_n] = \frac{n\mu}{n} = \mu$$

and its variance equals

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

So when reformulated in terms of the sample mean, the CLT states that the sample mean $\overline{X}$ becomes statistically very close to a Normal$(\mu, \sigma^2/n)$ as the sample size n grows. Note how the variance σ^2/n decays to 0. This means that $\overline{X}$ becomes closer and closer to its mean μ, which is consistent with what the law of large numbers says. This is in fact how we will prove the law of large numbers at the end of this lecture.

Instead of presenting the proof of the CLT we will see how it works in a few examples. Let us start with i.i.d. random variables that are normally distributed.

Example 21.2. We have seen already that the sum of independent normal random variables is itself a normal random variable. So if $X_1, \dots, X_n$ are independent random variables that are all $N(\mu, \sigma^2)$, then $X_1 + \cdots + X_n$ is Normal$(n\mu, n\sigma^2)$. This is much stronger than what the CLT claims. If the random samples already have a normal distribution, then there is no need to take n to be large to have the sum $X_1 + \cdots + X_n$ to be approximately normal. The sum $X_1 + \cdots + X_n$ is normal, for any sample size n. (And the sample mean $\overline{X}$ is Normal$(\mu, \sigma/\sqrt{n})$, for any n.)

The next case we will look at is the one where the X_i's are the most basic random variables, i.e. Bernoulli.

Example 21.3. We have seen that the sum of independent Binomials with the same success probability is a Binomial. In particular, if $X_1, \dots, X_n$ are i.i.d. Bernoulli(p) random variables, then $X_1 + \cdots + X_n$ is a Binomial(n, p). Recall that the mean of X_1 is $\mu = p$ and its variance is $\sigma^2 = p(1-p)$. So the CLT says that $(X_1 + \cdots + X_n - np)/\sqrt{np(1-p)}$ converges in distribution to a standard normal. This is exactly what the De Moivre-Laplace CLT says. So that CLT is a special case of the CLT we saw in this lecture.

In fact, you have already worked out the case for Poisson and Gamma random variables in Exercises 20.2 and 20.3! Let us see what would happen if the random variables are uniformly distributed.

Example 21.4. Suppose $X_1, \dots, X_n$ are i.i.d. Uniform(a, b) random variables. We have already seen that the mean of X_1 is $\mu = (a+b)/2$ and its variance is $\sigma^2 = (b-a)^2/12$. Let us compute the mgf of

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}.$$

We have

$$\begin{aligned}
M_{Z_n}(t) &= E\big[e^{\frac{t}{\sigma\sqrt{n}}X_1} \cdots e^{\frac{t}{\sigma\sqrt{n}}X_n} \cdot e^{-t\mu\sqrt{n}/\sigma}\big] \\
&= e^{-t\mu\sqrt{n}/\sigma} E[e^{\frac{t}{\sigma\sqrt{n}}X_1}]^n = e^{-t\mu\sqrt{n}/\sigma}\big[M_{X_1}\big(\frac{t}{\sigma\sqrt{n}}\big)\big]^n \\
&= e^{-t\mu\sqrt{n}/\sigma}\Big[\frac{e^{\frac{bt}{\sigma\sqrt{n}}} - e^{\frac{at}{\sigma\sqrt{n}}}}{(b-a)t/(\sigma\sqrt{n})}\Big]^n = \Big[e^{-\frac{t\mu}{\sigma\sqrt{n}}} \cdot \frac{e^{\frac{bt}{\sigma\sqrt{n}}} - e^{\frac{at}{\sigma\sqrt{n}}}}{(b-a)t/(\sigma\sqrt{n})}\Big]^n \\
&= \Big[\frac{e^{\frac{(b-\mu)t}{\sigma\sqrt{n}}} - e^{\frac{(a-\mu)t}{\sigma\sqrt{n}}}}{(b-a)t/(\sigma\sqrt{n})}\Big]^n.
\end{aligned}$$

Note that $b - \mu = b - (a+b)/2 = \frac{b-a}{2}$ and $a - \mu = -\frac{b-a}{2}$. So

$$M_{Z_n}(t) = \Big[\frac{e^{\frac{(b-a)t}{2\sigma\sqrt{n}}} - e^{\frac{-(b-a)t}{2\sigma\sqrt{n}}}}{(b-a)t/(\sigma\sqrt{n})}\Big]^n = \exp\Big\{n \log\Big[\frac{e^{\frac{(b-a)t}{2\sigma\sqrt{n}}} - e^{\frac{-(b-a)t}{2\sigma\sqrt{n}}}}{(b-a)t/(\sigma\sqrt{n})}\Big]\Big\}.$$

We want to take $n \to \infty$. Let $h = \frac{b-a}{\sigma\sqrt{n}}$. Then $n = \frac{(b-a)^2}{\sigma^2 h^2} = \frac{12}{h^2}$ and

$$M_{Z_n}(t) = \exp\Big\{\frac{12 \log\Big[\frac{e^{th/2} - e^{-th/2}}{th}\Big]}{h^2}\Big\}.$$

Taking $n \to \infty$ is the same as taking $h \to 0$. But we cannot just plug in $h = 0$ as we get $0/0$ expressions. So we use de l'Hôpital's rule to get

$$\begin{aligned}
\lim_{n\to\infty} M_{Z_n}(t) &= \exp\Big\{ \lim_{h\to 0} \frac{12 \log\Big[\frac{e^{th/2}-e^{-th/2}}{th}\Big]}{h^2} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{\Big(12 \log\Big[\frac{e^{th/2}-e^{-th/2}}{th}\Big]\Big)'}{2h} \Big\} \\
&= \exp\left\{ \lim_{h\to 0} \frac{6\Big(\frac{e^{th/2}-e^{-th/2}}{th}\Big)'}{h \cdot \frac{e^{th/2}-e^{-th/2}}{th}} \right\} \\
&= \exp\left\{ \lim_{h\to 0} \frac{6\Big(\frac{(t^2h/2)e^{th/2}+(t^2h/2)e^{-th/2}-te^{th/2}+te^{-th/2}}{t^2h^2}\Big)}{\frac{e^{th/2}-e^{-th/2}}{t}} \right\}.
\end{aligned}$$

Multiplying the top and bottom by th^2 then doing some cancelations and applying de l'Hôpital rule multiple times gives

$$\begin{aligned}
\lim_{n\to\infty} M_{Z_n}(t) &= \exp\Big\{ \lim_{h\to 0} \frac{3\big((th-2)e^{th/2}+(th+2)e^{-th/2}\big)}{h^2(e^{th/2}-e^{-th/2})} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big((th-2)e^{th}+(th+2)\big)}{h^2(e^{th}-1)} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big((th-2)e^{th}+(th+2)\big)'}{\big(h^2(e^{th}-1)\big)'} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big((th-2)te^{th}+te^{th}+t\big)}{h^2te^{th}+2h(e^{th}-1)} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big((t^2h-t)e^{th}+t\big)}{(h^2t+2h)e^{th}-2h} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big((t^2h-t)e^{th}+t\big)'}{\big((h^2t+2h)e^{th}-2h\big)'} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3\big(t^2e^{th}+(t^2h-t)te^{th}\big)}{(h^2t+2h)te^{th}+(2ht+2)e^{th}-2} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3t^3\big(he^{th}\big)}{(h^2t^2+4ht+2)e^{th}-2} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3t^3\big(he^{th}\big)'}{\big((h^2t^2+4ht+2)e^{th}-2\big)'} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3t^3\big(e^{th}+hte^{th}\big)}{(2ht^2+4t)e^{th}+(h^2t^2+4ht+2)te^{th}} \Big\} \\
&= \exp\Big\{ \lim_{h\to 0} \frac{3t^3\big(1+ht\big)}{6ht^2+h^2t^3+6t} \Big\} = e^{t^2/2}.
\end{aligned}$$

Since the mgf of Z_n converges to that of a standard normal, we deduce that Z_n converges in distribution to a standard normal, exactly like the CLT states.

The proof of the general CLT will be omitted, but it is very similar to how these special cases worked out, although it requires some technical workarounds since, unlike these examples, we do not have an explicit formula for the moment generating function to work with.

2. The law of large numbers

We can use the central limit theorem to prove the following result.

Theorem 21.5. *Suppose that $X_1, \cdots, X_n$ is a sequence of independent identically distributed random variables with mean μ and finite variance σ^2. Then for any $\varepsilon > 0$ we have*

$$P\{|\overline{X} - \mu| \leqslant \varepsilon\} = P\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \leqslant \varepsilon\Big\} \underset{n\to\infty}{\longrightarrow} 1.$$

In words, the theorem says that as the sample size n is increased, the sample mean becomes very likely to be close to the true mean.

Proof of the law of large numbers. Start by writing

$$\begin{aligned} P\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \leqslant \varepsilon\Big\} &= P\Big\{\Big|\frac{X_1 + \cdots + X_n - n\mu}{n}\Big| \leqslant \varepsilon\Big\} \\ &= P\Big\{\Big|\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}\Big| \leqslant \frac{\varepsilon\sqrt{n}}{\sigma}\Big\}. \end{aligned}$$

Fix any $\delta > 0$. We want to show that the above can be made larger than $1 - \delta$ by taking the n large enough. (This would say that you can make the above as close to 1 as you want by taking n appropriately large. In other words, that the above converges to 1 as $n \to \infty$, as claimed by the theorem.)

Fix an arbitrary positive number a. Then if $n > a^2\sigma^2/\varepsilon^2$ we will have $\varepsilon\sqrt{n}/\sigma > a$. This then implies that the interval $[-a, a]$ is inside the interval $[-\varepsilon\sqrt{n}/\sigma, \varepsilon\sqrt{n}/\sigma]$. So if

$$\Big|\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{\sigma^2 n}}\Big| \leqslant a$$

happens to be true, then

$$\Big|\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}\Big| \leqslant \frac{\varepsilon\sqrt{n}}{\sigma}$$

is also true. So the probability of the latter happening is larger than the probability of the former happening. Therefore,

$$P\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \leqslant \varepsilon\Big\} = P\Big\{\Big|\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}\Big| \leqslant \frac{\varepsilon\sqrt{n}}{\sigma}\Big\} \tag{21.1}$$

$$\geqslant P\Big\{-a \leqslant \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{\sigma^2 n}} \leqslant a\Big\}. \tag{21.2}$$

The CLT says that the last probability converges to

$$\Phi(a) - \Phi(-a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1$$

as $n \to \infty$. Since a is arbitrary we can take it as large as we want. But as a grows,

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

gets closer and closer to

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = 1.$$

Consequently, $2\Phi(a) - 1$ gets closer and closer to $2 \times 1 - 1 = 1$. So we can take a large enough to make $2\Phi(a) - 1 > 1 - \delta$. Then if we take n large enough we get the probability in (21.2) close enough to $2\Phi(a) - 1$ (now a number $> 1 - \delta$) to make it also $> 1 - \delta$. But then this makes the left-hand side of (21.1) larger than $1 - \delta$, which is what we wanted. □

Just like the De Moivre-Laplace CLT was used to get confidence intervals and test hypotheses about proportions, the general CLT can be used to build confidence intervals and to test hypotheses about the sample mean.

Example 21.6. 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 $\mu = 0.5$ minutes 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 waiting time until they all clear is on average 50 minutes. This is only an estimate, however. Suppose, for example, we want to estimate of the probabilities they take between 45 minutes and an hour to clear. If X_i is the waiting time of car number i, then we want to compute $P\{45 \leqslant X_1 + \cdots + X_{100} \leqslant 60\}$. For this particular problem, we can use the fact that the sum of independent Gamma distributed random variables with the same λ is another Gamma and that an Exponential(λ) random variable is a Gamma$(1, \lambda)$. Therefore, $X_1 + \cdots + X_{100}$ is a Gamma$(100, 2)$ random variable and we can use its pdf to compute the probability it is between 45 and 60. But we can also use the central limit theorem for this. Using the CLT is more general, since it does not require knowing the distribution of X_i. Only the mean μ and the variance σ^2.

In our example, the average total waiting time for the 100 cars is 50 minutes and the variance of the total waiting time is $100\sigma^2 = 25$. The theorem tells us that the distribution of

$$Z = \frac{X_1 + \cdots + X_{100} - 50}{\sqrt{25}}$$

is approximately standard normal. Thus,

$$\begin{aligned} P\{45 \leqslant X_1 + \cdots + X_{100} \leqslant 60\} &= P\{-5/5 \leqslant Z \leqslant 10/5\} \\ &\approx \Phi(2) - \Phi(-1) \\ &= \Phi(2) - (1 - \Phi(1)) = \Phi(2) + \Phi(1) - 1 \\ &\approx 0.9772 + 0.8413 - 1 = 0.8185, \end{aligned}$$

i.e. about 82%.

Suppose now we do not know the average waiting time per car at the toll station, but someone claims it is at least half a minute. Suppose we observed that 100 cars cleared in 55 minutes. Does this data back up the claim or does it contradict it?

To answer this question let us first go back to the theory. Suppose again that X_i is the waiting time of car number i. Then the total waiting time is $X_1 + \cdots + X_{100}$, which we observed to equal 55 minutes. The law of large numbers tells us then that $55/100 = 0.55$ minutes is an estimate. This seems to back up the claim, but could it be that this was just due to randomness? That is, if the claim were wrong and the average waiting time per car were less than half a minute, then what are the odds that 100 cars will wait for as long as 55 minutes? Clearly, the odds of this get smaller if the average waiting time per car is smaller. So to consider the worst case scenario, we can assume the average waiting time to be exactly half a minute. This means $\lambda = 1/0.5 = 2$ in our model. So now the question is to compute the probability

$$P\{X_1 + \cdots + X_{100} \geqslant 55\}$$

when $X_1, \dots, X_{100}$ are i.i.d. Exponential(2) random variables. This is the same as

$$P\{X_1 + \cdots + X_{100} \geqslant 55\} = P\Big\{ \frac{X_1 + \cdots + X_{100} - 50}{5} \geqslant \frac{55 - 50}{5} \Big\}$$

($50 = 100 \times 0.5 = n\mu$ and $5 = \sqrt{25} = \sqrt{\sigma^2 n}$ where $\sigma^2 = 1/\lambda^2 = 1/4$.) By the CLT this can be approximated by $1 - \Phi(1) \approx 1 - 0.8413 = 0.1587$, i.e. about 16%. This is quite high. So it could be that the average waiting time per car is half a minute (or even a little less) and yet, because of randomness, a sample of 100 cars takes as long as 55 minutes to clear.

Read sections 9.2 and 9.3 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 21.1. A carton contains 144 baseballs, each of which has a mean weight of 5 ounces and a standard deviation of 2/5 ounces. Find an approximate value for the probability that the total weight of the baseballs in the carton is no more than 725 ounces.

Exercise 21.2. 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) $P\Big\{ \sum_{i=1}^{20} X_i \leqslant 12 \Big\}$.

(b) The 90-th percentile of $\sum_{i=1}^{20} X_i$; i.e. the number a for which

$$P\Big\{ \sum_{i=1}^{20} X_i \leqslant a \Big\} = 0.9.$$

Exercise 21.3. Let X_i be the weight of the i-th passenger's luggage. Assume that the weights are independent from passenger to passenger, each with pdf

$$f(x) = 3x^2/80^3, \text{ for } 0 < x < 80,$$

and 0 otherwise. Approximate $P\Big\{ \sum_{i=1}^{100} X_i > 6025 \Big\}$.

Exercise 21.4. 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.

Exercises 9.6 and 9.7 on page 323 in the textbook by Anderson, Sepäläinen, and Valkó. (For 9.7 just use the CLT.)

Lecture 22

1. Cumulative distribution functions

We have seen several times by now that a random variable can be statistically described by explaining how to compute for it the probabilities $P\{X \leqslant b\}$, $P\{X < b\}$, $P\{X \geqslant a\}$, $P\{X > a\}$, $P\{a \leqslant X \leqslant b\}$, $P\{a < X \leqslant b\}$, $P\{a \leqslant X < b\}$, and $P\{a < X < b\}$ for all numbers $a < b$. It turns out that in fact one can compute all these probabilities knowing only how to calculate $P\{X \leqslant b\}$ for all b. This is called the cumulative distribution function (CDF) of X and it is usually denoted by a capital F:

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

(As usual, we write F_X if we want to emphasize that this is the CDF of X.)

For example, $P\{X > a\} = 1 - P\{X \leqslant a\} = 1 - F(a)$ and $P\{a < X \leqslant b\} = P\{X \leqslant b\} - P\{X \leqslant a\} = F(b) - F(a)$. We will see in a moment how all the other probabilities can be computed from the CDF. But first, let us mention some basic properties of a CDF.

Theorem 22.1. *Suppose F is a CDF of some random variable (discrete, continuous, or other). Then the following are all true.*

(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.

In fact, these properties identify a CDF. So when asked whether a function is a CDF or not, what is meant is to check that properties (a)-(c) are all satisfied. Note that (c) only asks for continuity from the right. The CDF may not be continuous from the left. A point where the CDF is not continuous from the left is called a jump point. Figure 22.1 depicts a possible CDF. Note how it goes to 0 as we go far to the left and to 1 as we go far to the right. It is non-decreasing and continuous at most points. It has a jump at the point x_0 but it is still continuous from the right at that point, meaning that if we look at where the graph of the function goes as we get closer and closer to x_0 but from the right side, then we see that the function approaches its own value at x_0, which in this example equals $1/2$. This does not happen if we approach x_0 from the left: the limiting value is $1/3$. The size of the jump is thus, $1/2 - 1/3 = 1/6$. Note also how the CDF

[0] Last modified on November 16, 2020 at 12:40:40 -07'00'

may not be always strictly increasing. In the example in the figure the CDF has a flat segment between x_1 and x_2.

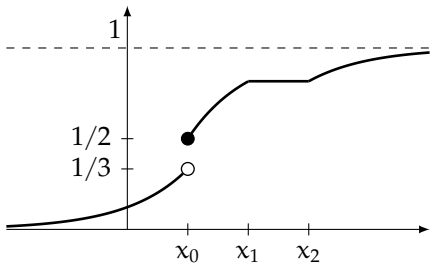


Figure 22.1.

Proof of Theorem 22.1. Property (a) just follows from the fact that if $x \leqslant y$ then $(-\infty, x] \subset (-\infty, y]$ and so the probability that X belongs to the former interval is smaller than the probability that X belongs to the latter interval.

The other two properties follow from a slightly deeper (but quite intuitive) fact about probability models, namely:

Lemma 22.2. *Let X be any random variable (discrete, continuous, or other). Let A_n be a sequence of increasing sets of real numbers: $A_1 \subset A_2 \subset A_3 \subset \cdots$. Then,*

$$\lim_{n\to\infty} P(X \in A_n) = P\Big(X \in \bigcup_{n\geqslant 1} A_n\Big).$$

(In words: the probability X is in the set A_n converges, as n grows, to the probability X is in one of the sets.)

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

$$\lim_{n\to\infty} P(X \in A_n) = P\Big(X \in \bigcap_{n\geqslant 1} A_n\Big).$$

(In words: the probability X is in the set A_n converges, as n grows, to the probability X is in all of the sets.)

The way we are about to apply this lemma should make it more clear what the lemma is saying and that it is intuitively a true statement. So we will omit the (not too difficult) proof.

Applying this lemma with the increasing sets $A_n = (-\infty, n]$ (whose union is the whole real line $\mathbb{R}$) gives that $F(n) = P(X \leqslant n) \to P(X \in \mathbb{R}) = 1$ as $n \to \infty$. Since F is non-decreasing this shows that in fact $F(x) \to 1$ as $x \to \infty$. (The difference between the two conclusions is that the first one only gave the limit along integers, which in general would not guarantee the limit along a different sequence would be the same. This is though the case here because F is non-decreasing.)

Applying the lemma with the decreasing sets $A_n = (-\infty, -n)$ (whose intersection is $\varnothing$) gives that $F(-n) = P(X \leqslant -n) \to P(X \in \varnothing) = 0$ as $n \to \infty$. And again, since F is non-decreasing this implies that $F(x) \to 0$ as $x \to -\infty$. We have thus shown that property (b) is true.

For property (c) fix any real number x and apply the lemma with the decreasing sets $A_n = (-\infty, x + 1/n]$ (whose intersection is $(-\infty, x]$) to get that $F(x + 1/n) = P(X \leqslant x + 1/n) \to P(X \leqslant x) = F(x)$. One more time, since F is non-decreasing this implies that $F(y) \to F(x)$ as y goes down to x. So property (c) is true and the theorem is proved. □

Now let us address the question of computing the different types of probabilities, mentioned at the very beginning of the lecture, in terms of the CDF. For this, let us use the notation $F(x-)$ to denote the limit of the function as we approach x from the left:

$$F(x-) = \lim_{y \nearrow x} F(y).$$

If F is continuous at x then $F(x) = F(x-)$. But if there is a jump at x, then $F(x) - F(x-)$ is exactly the size of the jump. The next theorem says that the size of the jump at x is exactly the probability the random variable equals x.

Theorem 22.3. *Suppose F is a CDF of some random variable (discrete, continuous, or other). Then for any real number x*

$$P\{X = x\} = F(x) - F(x-). \tag{22.1}$$

Looking back at the example in Figure 22.1, this theorem says that if X has that CDF then the probability $P(X = x_0)$ equals the size of the jump at x_0, which is $1/2 - 1/3 = 1/6$. Since that CDF is continuous at any point x other than x_0 we deduce that $P(X = x) = 0$ (because the size of the jump at such an x would be 0).

Proof of Theorem 22.3. First, note that if $y < x$ then

$$F(x) - F(y) = P(X \leqslant x) - P(X \leqslant y) = P(y < X \leqslant x).$$

Now apply Lemma 22.2, this time with the decreasing sets $A_n = (x - 1/n, x]$, the intersection of which is the set $\{x\}$ (x is the only point that belongs to all these intervals). So then $F(x) - F(x - 1/n) = P(x - 1/n < X \leqslant x) \to P(X \in \{x\}) = P(X = x)$ as $n \to \infty$. Again, since F is non-decreasing we can deduce that not only $F(x) - F(x - 1/n)$ converges to $P(X = x)$ as $n \to \infty$, but in fact more generally $F(x) - F(y)$ converges to $P(X = x)$ as y gets closer and closer to x from the left (not necessarily only along the sequence $x - 1/n$). This is what the theorem claims and so we are done with the proof. □

Now we see that

(a) $P(X \leqslant b) = F(b)$ (the value at b).

(b) $P(X < b) = P(X \leqslant b) - P(X = b) = F(b) - (F(b) - F(b-)) = F(b-)$ (the left limit at b).

The first is the definition of the CDF and the second is a consequence of the above theorem.

From that we can compute the rest of the probabilities of interest:

(c) $P(X > a) = 1 - P(X \leqslant a) = 1 - F(a)$.

(d) $P(X \geqslant a) = 1 - P(X < a) = 1 - F(a-)$.

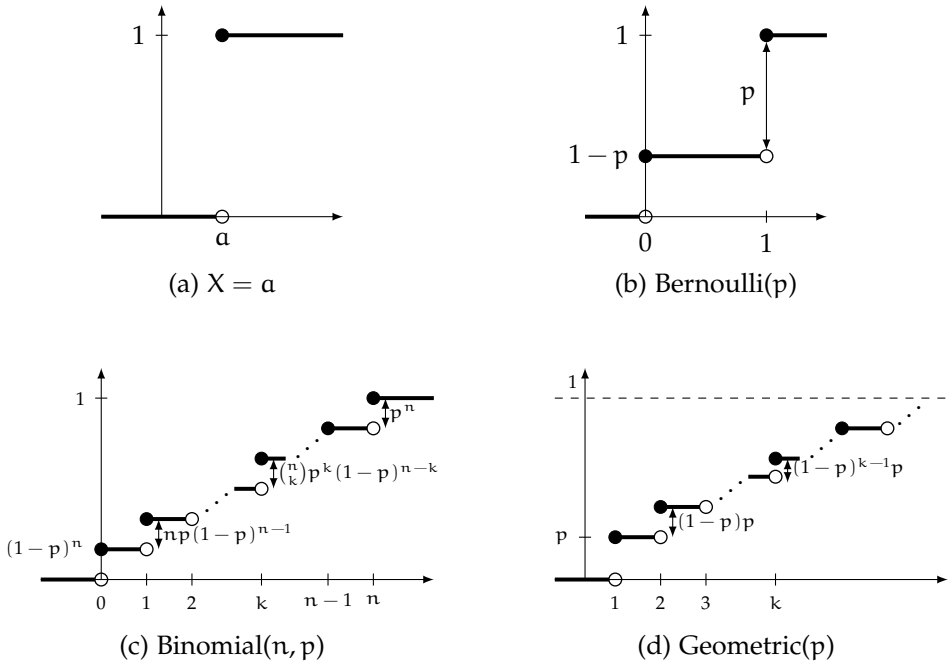
(e) $P(a < X \leqslant b) = P(X \leqslant b) - P(X \leqslant a) = F(b) - F(a)$ (the value at b minus the value at a).

(f) $P(a \leqslant X < b) = P(X < b) - P(X < a) = F(b-) - F(a-)$ (the left limit at b minus the left limit at a).

(g) $P(a \leqslant X \leqslant b) = P(X \leqslant b) - P(X < a) = F(b) - F(a-)$ (the value at b minus the left limit at a).

(g) $P(a < X < b) = P(X < b) - P(X \leqslant a) = F(b-) - F(a)$ (the left limit at b minus the value at a).

In particular, if the CDF F is constant over some interval $[x_1, x_2]$, as in the example in Figure 22.1, then $P(x_1 < X \leqslant x_2) = F(x_2) - F(x_1) = 0$. This says that there is zero probability of the random variable taking a value between x_1 and x_2.



(a) $X = a$ (b) Bernoulli(p)

(c) Binomial(n, p) (d) Geometric(p)

Figure 22.2. CDFs for some discrete distributions

Let us now work out the CDFs of some of the random variables we have learned.

Example 22.4. Let X be nonrandom. That is, $P\{X = a\} = 1$ for some number a. Such a random variable is called "deterministic." Then

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

See Figure 22.2(a). Note how the CDF is flat before a and flat after it. This means it has zero probability to take values $< a$ and zero probability to take values $> a$. The size of the jump at a is 1 and so the probability the random variable takes the value a is 1 (100%).

Example 22.5. Let X be Bernoulli with parameter $p \in [0, 1]$. Then

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

See Figure 22.2(b). Note how the CDF is flat before 0, flat between 0 and 1, and flat after 1. This means there is zero probability to take values < 0, strictly between 0 and 1, or > 1. In other words, the only possible values are 0 and 1. The size of the jump at 0 is $1 - p$, which is the probability the random variable takes the value 0. The size of the jump at 1 is p, which is the probability the random variable takes the value 1.

Example 22.6. Let X be binomial with parameters n and p. Then

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

See Figure 22.2(c). In words: the CDF starts off being 0 until we reach 0. This is consistent with the fact that the random variable does not take any negative values. Then the CDF makes a jump of size $(1-p)^n$ at 0 (and takes the value $(1-p)^n$ at 0). Then it remains constant (equal to $(1-p)^n$) until we reach 1. Again, this is consistent with the fact that the random variable does not take any values that are strictly between 0 and 1. At 1 it makes a jump of size $\binom{n}{1}p^1(1-p)^{n-1} = np(1-p)^{n-1}$, bringing the function to the value $(1-p)^n + np(1-p)^{n-1}$ at 1. Then again the CDF remains constant (equal to $(1-p)^n + np(1-p)^{n-1}$) until we reach 2 at which point the CDF makes a jump of size $\binom{n}{2}p^2(1-p)^{n-2}$, bringing the value of the function at 2 to $(1-p)^n + np(1-p)^{n-1} + \binom{n}{2}p^2(1-p)^{n-2}$. This continues until we reach n at which point the last jump the function makes is of size p^n, making the value at n equal to

$$\sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} = 1.$$

From n and on the CDF remains equal to 1, which is consistent with the fact that the random variable does not take values $> n$.

Example 22.7. Let X be geometric with parameter p. Then

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

See Figure 22.2(d). Here, we used the fact that

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

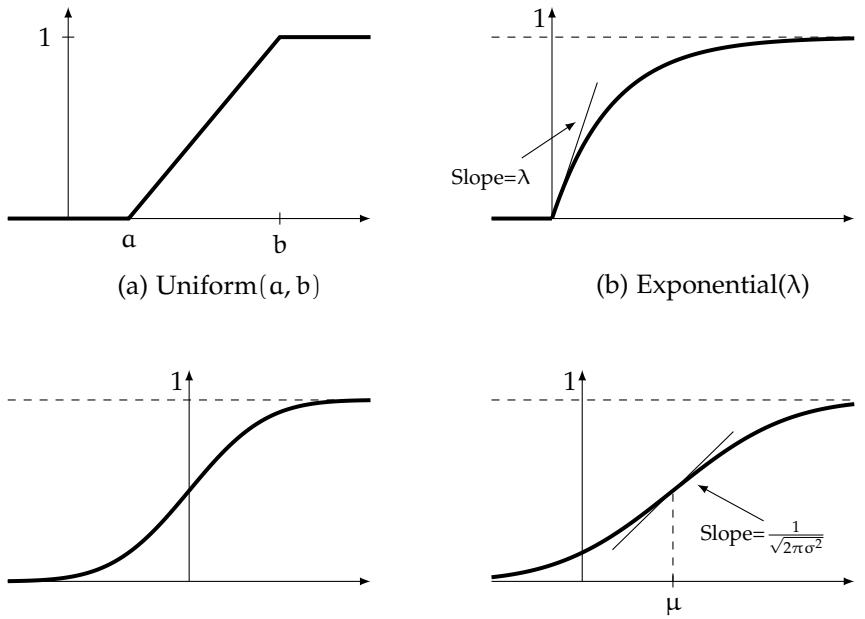
The CDF here behaves similarly to the previous ones, except that it never really reaches 1. The CDF starts off being 0 until we reach 1, at which point the CDF jumps by the amount p (the probability the geometric random variable equals 1). Then it remains constant until we reach 2, at which point it jumps by $(1-p)p$ (the probability the random variable equals 2). And so on: the CDF remains constant until the next integer is reached, at which point it makes a jump of size equal to the probability the random variable takes a value equal to that integer. Since the set of values the geometric random variable can take is infinite (all of the positive integers), this process continues for ever and even though the jumps get smaller and smaller, the CDF never reaches 1. But it does approach 1 as we keep going to the right.

Note how in all of the above cases, the CDF is flat (i.e. constant) except for when it makes a jump. Such a function is called a piece-wise constant function. This means that there is zero probability of the random variable taking values other than where the CDF jumps. And if x is a point where the CDF does jump (i.e. a possible value of the random variable), then the size of the jump is exactly $P(X = x)$, i.e. the value of the mass function. Piece-wise CDFs are exactly what the CDF of a discrete random variable looks like. And as we just explained, the jump locations are the possible values of the random variable and the jump sizes are the corresponding values of the mass function.

Now for a continuous random variable with pdf f we have

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

This says that just like we could retrieve the mass function from the CDF of a discrete random variable by looking at the locations and the sizes of the jumps, we can retrieve the pdf of a



(a) Uniform(a, b) (b) Exponential(λ)

(c) Standard Normal (d) Normal(μ, σ^2) with $\sigma^2 > 1$

Figure 22.3. CDFs for some continuous distributions

continuous random variable from its CDF by differentiating:

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

In particular, this implies that the CDF of a continuous random variable must be a continuous function, i.e. has no jumps. This is why a random variable with a pdf is called a continuous random variable! This also fits with our understanding: the size of the jump at x is the probability the random variable takes the value x. So:

$$\text{the size of the jump at } x = P(X = x) = \int_x^x f(z)\,dz = 0.$$

Since there are no jumps, we see that for any $a < b$

$$P(X \leqslant b) = P(X < b) = \int_{-\infty}^{b} f(z)\,dz, \quad P(X \geqslant a) = P(X > a) = \int_a^{\infty} f(z)\,dz,$$

and also

$$P(a < X \leqslant b) = P(a \leqslant X < b) = P(a < X < b) = P(a \leqslant X \leqslant b) = \int_a^b f(z)\,dz.$$

Let us now compute the CDFs of a few of the continuous random variables we know.

Example 22.8. Let X be uniform on the interval $[a, b]$. Then

$$F(x) = \begin{cases} 0 & \text{if } x \leqslant a, \\ \frac{1}{b-a}\int_a^x dz = \frac{x-a}{b-a} & \text{if } a \leqslant x \leqslant b, \\ 1 & \text{if } x \geqslant b. \end{cases}$$

See Figure 22.3(a). Note how the CDF is flat before a and after b. This corresponds to the fact that the random variable does not take values in these regions. The CDF is continuous, because the random variable is a continuous random variable (it has a pdf).

Example 22.9. Let X be exponential with parameter λ. Then

$$F(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ \int_0^x \lambda e^{-\lambda z}\,dz = 1 - e^{-\lambda x} & \text{if } x \geqslant 0. \end{cases}$$

See Figure 22.3(b). Note how the CDF is flat before 0 but never flat after 0. This corresponds to the fact that the random variable does not take negative values but can take any positive value. The CDF is again continuous.

Example 22.10. Let X a standard normal. Then its CDF is none other than the error function

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

(Since this CDF is often used, it was given the special name Φ instead of the generic F.) See Figure 22.3(c). Note how the CDF is never flat. This corresponds to the fact that the random variable can take any value.

If X is instead a normal random variable with mean μ and variance σ^2, then we have also seen that a change of variables brings things back to the standard case and that way we get that the CDF can be written in terms of Φ as

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(y-\mu)^2/(2\sigma^2)}\,dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-z^2/2}\,dz = \Phi\Big(\frac{x-\mu}{\sigma}\Big).$$

See Figure 22.3(d).

Now suppose we are given the formula for the pdf of a random variable and are asked to calculate the CDF. Since the pdf is the derivative of the CDF all we have to do is take the antiderivative of the CDF. However, antiderivatives are always defined up to constants and that is where the subtlety comes it: we need to keep in mind that the CDF calculates the cumulative area under the pdf curve. Let us see how this is done on an example.

Example 22.11. Suppose X has the pdf

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leqslant x < 1 \\ 1/2 & \text{if } 1 < x < 2 \\ 0 & \text{if } x \geqslant 2. \end{cases}$$

Before we calculate the CDF of X let us make a few observations. First, f is indeed a pdf because it is always nonnegative and it integrates to 1. To verify that it integrates to 1 we could either simply compute

$$\int_0^1 x\,dx + \int_1^2 \frac{1}{2}\,dx = 1$$

or observe that the graph of the pdf consists of a triangle with area $1 \times 1/2 = 1/2$ and a rectangle with area $1/2 \times 1 = 1/2$ and hence the total area indeed 1. Second, note how the domains for the various formulas are not necessarily of the form $a \leqslant x < b$, like it would be for a CDF. This is because a pdf does not have to be continuous nor left- nor right-continuous. In fact, the value of the pdf at finitely many points is immaterial because we always integrate it to compute probabilities and the integral over a point is just 0. Let us now calculate the CDF.

When $x < 0$, the pdf is identically 0 and then so is the CDF. When $0 \leqslant x < 1$, the CDF is given by

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{0}^{x} y\,dy = \frac{y^2}{2}.$$

In particular, make note that $P(X \leqslant 1) = F(1) = 1/2$. This is the area under the pdf to the left of 1. Now, when $1 \leqslant x < 2$, the CDF is given by

$$F(x) = P(X \leqslant x) = P(X \leqslant 1) + P(1 < X \leqslant x) = \frac{1}{2} + \int_{1}^{x} \frac{1}{2}\,dy = \frac{1}{2} + \frac{(x-1)}{2} = \frac{x}{2}.$$

Note how we integrated from 1 to x, but also added the area to the left of 1. Again, we make note that $P(X \leqslant 2) = F(2) = 1$. Since the CDF reached the value of 1, it will remain at 1 from there and on. Indeed, the pdf for $x \geqslant 2$ is 0 and so no new area will accumulate. In other words, for $x \geqslant 2$ we have

$$P(X \leqslant x) = P(X \leqslant 2) + \int_{2}^{x} 0\,dx = P(X \leqslant 2) = 1.$$

To summarize:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2/2 & \text{if } 0 \leqslant x < 1 \\ x/2 & \text{if } 1 \leqslant x < 2 \\ 1 & \text{if } x \geqslant 2. \end{cases}$$

We end this lecture with an interesting observation. Namely, there are random variables that are not discrete nor continuous. Namely, if the CDF is not piece-wise constant then the random variable is not discrete and if the CDF has jumps then the random variable is not continuous. So for example the random variable that corresponds to the CDF in Figure 22.1 is not discrete (since the CDF is not piece-wise constant) and is not continuous (since the CDF has a jump). In fact, there is a way to work with the CDF (which all random variables have) instead of worrying about whether or not the random variable has a mass function or a pdf. This (more elegant) way allows us to have one unifying treatment of all random variables, without the need to distinguish discrete random variables and continuous random variables as we have been doing this whole time. However, developing intuition for this approach takes time and a certain level of skill with calculus (and mathematical analysis, if one wants to be fully rigorous) and so we did not (and will not) go that route in this course.

Read section 3.2 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 22.1. Let F be the function defined by:

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

Let X be a random variable which corresponds to F.

(a) Verify that F is a cumulative distribution function.
(b) Compute $P\{X = 2\}$, $P\{X < 2\}$, and $P\{X \leqslant 2\}$.
(c) Compute $P\{X > 2\}$ and $P\{X \geqslant 2\}$.
(d) Compute $P\{-1 \leqslant X \leqslant 1/2\}$, $P\{1/3 \leqslant X < 1/2\}$, and $P\{X \in (1/3, 3/2]\}$.
(e) Compute $P\{4/3 \leqslant X \leqslant 5/3\}$, $P\{3/2 \leqslant X \leqslant 2\}$, and $P\{3/2 < X < 2\}$.
(f) Compute $P\{2 < X < 3\}$, $P\{2 \leqslant X < 3\}$, and $P\{3 \leqslant X < 5\}$.
(g) Compute $P\{3/2 \leqslant X < 3\}$, $P\{3/2 < X \leqslant 3\}$, and $P\{1/2 < X \leqslant 3\}$.
(h) Compute $P\{X = 2 \text{ or } 1/2 \leqslant X < 3/2\}$.
(i) Compute $P\{X = 2 \text{ or } 1/2 \leqslant X \leqslant 3\}$.

Exercises 3.5, 3.6, and 3.7 on page 127 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 23

In this lecture we will learn how to find the distribution of a function of a random variable. The main example is the following: If Z is a standard normal random variable, then what is the pdf of $X = Z^2$? This will lead us to a random variable that is often used in statistical applications, namely the Chi square (χ^2) random variable. We start with the easier case of a discrete random variable and try to understand the issue on two examples.

Example 23.1. Suppose X has the mass function

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = -2, \\ \frac{1}{3} & \text{if } x = 0, \\ \frac{1}{4} & \text{if } x = 1, \\ \frac{1}{4} & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

Define the new random variable $Y = X^3 + 1$. Then Y takes the values -7, 1, 2, and 9. Its mass function is

$$\begin{aligned} f_Y(-7) &= P\{X = -2\} = f_X(-2) = \frac{1}{6}, \\ f_Y(1) &= P\{X = 0\} = f_X(0) = \frac{1}{3}, \\ f_Z(2) &= P\{X = 1\} = f_X(1) = \frac{1}{4}, \quad \text{and} \\ f_Z(9) &= P\{X = 2\} = f_X(2) = \frac{1}{4}. \end{aligned}$$

Example 23.2. Now, with the same X as above let $Z = X^2 + 1$. The main difference between Y from the previous example and Z from this example is that $x^3 + 1$ is a one-to-one function (i.e. solving $y = x^3 + 1$ for x gives the unique solution $x = (y-1)^{1/3}$) while $x^2 + 1$ is not one-to-one (if $z \geqslant 1$ then solving $z = x^2 + 1$ for x gives the two solutions $x = \pm\sqrt{x-1}$).

[0] Last modified on November 15, 2022 at 13:40:57 -07'00'

Here, Z takes the values 1, 2, and 5. Its mass function is

$$f_Z(1) = P\{X = 0\} = f_X(0) = \frac{1}{3},$$
$$f_Z(2) = P\{X = 1\} = f_X(1) = \frac{1}{4}, \quad \text{and}$$
$$f_Z(3) = P\{X = -2 \text{ or } X = 2\} = f_X(-2) + f_X(2) = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$

Note how the case where the function is not one-to-one needed some extra care. We needed to pay attention to the fact that some values that Z takes arise from multiple values of X and hence we need to add a few values of the mass function of X to get the mass function of Z.

Now, let us work out the continuous case. The basic problem is this: If $Y = g(X)$, then how can we compute f_Y in terms of f_X? One way is to first compute the CDF F_Y from F_X and then take its derivative. Let us work things out on a few examples.

Example 23.3. Suppose X is uniform on $(0,1)$ and $Y = -\log X$. Then, we compute f_Y by first computing F_Y, and then using $f_Y = F_Y'$. Here are the details:

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

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

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

Since X is always between 0 and 1, $Y = -\log X$ is always positive. So $F_Y(y) = P(Y \leqslant y) = 0$ for $y \leqslant 0$. Consequently, $f_Y(y) = 0$ for $y < 0$ and for $y > 0$ we have

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

In other words, Y is an Exponential(1) random variable.

Remark 23.4. The above result is neat: if we can generate a Uniform$(0,1)$ random variable (e.g. by hitting the `RAND` [or maybe it is called `RND`] button on the calculator) then $-\log$ (natural logarithm) of that will give us samples that are exponentially distributed, with parameter $\lambda = 1$. In fact, exponential random variables are not special. One can generate any random variable using a Uniform$(0,1)$. For example, it is quite intuitive to generate a Bernoulli(p) random variable (or a p-coin) using a Uniform$(0,1)$ random variable (say U) by saying that we get a 1 (or heads) if $U \leqslant p$ and a 0 otherwise. The curious student can see how this idea works at the end of the lecture, but this will not be covered in class.

Example 23.5. Suppose $\mu \in \mathbb{R}$ and $\sigma > 0$ are fixed constants, and define $Y = \mu + \sigma X$. To find the density of Y in terms of that of X we compute

$$F_Y(y) = P(Y \leqslant y) = P(\mu + \sigma X \leqslant y) = P\Big(X \leqslant \frac{y-\mu}{\sigma}\Big) = F_X\Big(\frac{y-\mu}{\sigma}\Big).$$

Taking derivatives and using the chain rule we get

$$f_Y(y) = \frac{1}{\sigma} f_X\Big(\frac{y-\mu}{\sigma}\Big).$$

Now if X is a standard Normal random variable, then we get

$$f_Y(y) = \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{[(y-\mu)/\sigma]^2}{2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

That is, $Y \sim N(\mu, \sigma^2)$.

Example 23.6. Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$. Then, $y = e^x > 0$, $x = \log y$, and

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

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.

In all of the above examples, the function g was one-to-one. Let us see how things work when this is not the case.

Example 23.7. 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 (and don't forget to use the chain rule) 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 $y < 0$ (because Y is always positive, since it is the square of something). Therefore, also $f_Y(y) = 0$ when $y < 0$.

Example 23.8 (The Chi square distribution)**.** If Z is a standard normal random variable, then $X = Z^2$ is said to have a Chi square distribution with one degree of freedom. This is denoted by $X \sim \chi^2(1)$. Applying the last example (where the X there is our Z here and our X here is the Y there) we see that the pdf of X is given by

$$f_X(x) = \begin{cases} \dfrac{e^{-x/2}}{\sqrt{2\pi x}} = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} & \text{if } x > 0, \\ 0 & \text{if } x \leqslant 0. \end{cases}$$

In other words, $X \sim \text{Gamma}(1/2, 1/2)$!

If $Z_1, \ldots, Z_n$ are i.i.d. standard normal random variables, then $X = Z_1^2 + \cdots + Z_n^2$ is said to have the Chi square distribution with n degrees of freedom. This is denoted by $X \sim \chi^2(n)$. In words, a Chi square random variable with n degrees of freedom is the sum of the squares of n independent standard normal random variables. Since we just saw that the square of a standard normal is in fact a Gamma with parameters $1/2$ and 1 and since the sum of n independent Gamma$(1/2, 1/2)$ random variables is a Gamma$(n/2, 1/2)$ random variable, we see that in fact the Chi square distribution with n degrees of freedom is none other than a Gamma$(n/2, 1/2)$.

Example 23.9. If X is Uniform$(0, 1)$ and $Y = X^2$, then $X^2 = Y$ has one solution: $X = \sqrt{Y}$. This is because we know X is positive, and so $X = -\sqrt{Y}$ is not acceptable. Thus, even though $y = x^2$ seems at first like it puts us in the "not one-to-one" situation, we are in fact in the one-to-one situation and

$$F_Y(y) = P\{Y \leqslant y\} = P\{X^2 \leqslant y\} = P\{0 \leqslant X \leqslant \sqrt{y}\} = P\{X \leqslant \sqrt{y}\} = F_X(\sqrt{y}).$$

Differentiating (and using the chain rule) we get

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot f_X(\sqrt{y}) = \begin{cases} \dfrac{1}{2\sqrt{y}} & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

But what happens if X is Uniform$(-1,2)$ and $Y = X^2$? Well, then $X^2 = Y$ has two solutions when $0 < Y < 1$ and only one solution when $1 < Y < 4$. (Draw $y = x^2$ on the interval $(-1,2)$.) Then if $0 < y < 1$ we have

$$F_Y(y) = P\{Y \leqslant y\} = P\{X^2 \leqslant y\} = P\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\} = \frac{2\sqrt{y}}{3},$$

because $[-\sqrt{y}, \sqrt{y}] \subset (-1,2)$. On the other hand, if $1 < y < 4$ then

$$F_Y(y) = P\{Y \leqslant y\} = P\{X^2 \leqslant y\} = P\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\} = P\{-1 \leqslant X \leqslant \sqrt{y}\} = \frac{\sqrt{y}+1}{3},$$

because $[-\sqrt{y}, \sqrt{y}] \cap (-1,2) = (-1, \sqrt{y}]$. When $y < 0$ we clearly have $F_Y(y) = 0$ (becauwe Y does not take negative values) and similarly, when $y > 4$, $F_Y(y) = 1$ because Y is always smaller than 4.

Differentiating the above formula we get

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

Example 23.10. 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

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

There is no need to compute the integral, if we are after f_Y. Indeed, differentiating (and using the chain rule) we get

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

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 is supposed to integrate to 1.

In fact, if $y > 1$, then $1 - \sqrt{y} < 0$ and so the pdf is 0, not $3e^{-3x}$, while x is between $1 - \sqrt{y}$ and 0. Therefore,

$$F_Y(y) = \int_{0}^{1+\sqrt{y}} 3e^{-3x}\, dx$$

and differentiating (and using the chain rule) we get

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

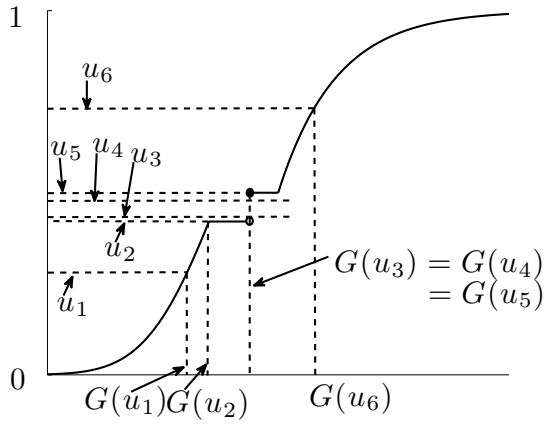


Figure 23.1. The function $G(u)$.

Example 23.11. 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 for any real number y

$$\begin{aligned} F_Y(y) &= P\{Y \leqslant y\} = P\{\tan\Theta \leqslant y\} = P\{\Theta \leqslant \arctan(y)\} \\ &= F_\Theta\big(\arctan(y)\big) = \frac{1}{\pi}\big(\arctan(y) + \pi/2\big). \end{aligned}$$

(Recall that Θ is uniform on $(-\pi/2, \pi/2)$.) Differentiating in the above display we get that the pdf of Y is

$$f_Y(y) = \frac{1}{\pi(1+y^2)}.$$

That is, Y is Cauchy distributed!

The rest of this lecture explains how one can generate a random variable with a known CDF using a Uniform$(0,1)$ random variable. This material is for the interested student only. It will not be covered in class.

Define

$$G(u) = \inf\{x : u \leqslant F(x)\} = \min\{x : u \leqslant F(x)\};$$

see Figure 23.1. Note that if $F^{-1}(u)$ exists, i.e. if we can find a unique x such that $F(x) = u$, then $G(u) = F^{-1}(u) = x$. The points u_1 and u_6 in the figure is an example of such a situation. However, G is still well defined at points u where we can either not find an x such that $F(x) = u$ (like for example the points u_3 and u_4) or we can find multiple x such that $F(x) = u$ (like the points u_2 and u_5).

Theorem 23.12. *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 CDF. 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 23.13. 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.

Example 23.14. To generate an exponential random variable with parameter λ we note that the CDF is $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ and 0 otherwise. In this case, for any $u \in (0,1)$ we can find a unique x such that $F(x) = u$. Namely, solving $u = 1 - e^{-\lambda x}$ we get $x = -\frac{\log(1-u)}{\lambda}$. Thus, according to the above theorem, $-\lambda^{-1}\log(1-U)$ has an exponential distribution with parameter λ, where $U \sim \text{Uniform}(0,1)$. So we have a way to generate an exponential random variable with a given parameter out of a Uniform(0,1) random variable. [Note that in this special case, $1-U$ is also Uniform$(0,1)$, and so we can simplify things a bit: $-\lambda^{-1}\log U$ also generates an Exponential(λ) random variable.]

Here is an exercise that the interested student can work on to make sure they understood the above.

Exercise. Let F be the function defined by:

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

Let U be a Uniform(0,1) random variable. Give a transformation G that would make $X = G(U)$ a random variable with CDF F.

Read section 5.2 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 23.1. Let X be a uniform random variable on $[-1, 1]$. Let $Y = e^{-X}$. What is the probability density function of Y?

Exercise 23.2. Let X be an exponential random variable with parameter $\lambda > 0$. What is the probability density function of $Y = X^2$?

Exercise 23.3. First, show that for any real number x the equation $z^3 + z + 1 = x$ has a unique solution in z. Now, let X be a standard normal random variable and let Z be the random variable that is the unique solution of $Z^3 + Z + 1 = X$. Find the probability density function of Z.

Exercise 23.4. Let X be an exponential random variable with parameter $\lambda > 0$. Find the probability density function of $Y = \log(X)$.

Exercise 23.5. Let X be an Uniform$(0, 1)$ random variable and let λ be a given positive number. Find the probability density function of $Y = -\lambda^{-1} \log(X)$.

Exercise 23.6. Let X be a Normal(μ, σ^2) random variable. Let a and b be two real numbers. Find the probability density function of $Y = aX + b$.

Exercise 23.7. 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 23.8. We throw a ball from the origin with velocity v_0 at 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:* recall that the laws of mechanics say that the distance is given by $R = \frac{v_0^2 \sin(2\Theta)}{g}$, where g is the gravity constant.

Exercise 5.6 on page 198 in the textbook by Anderson, Sepäläinen, and Valkó.

Lecture 24

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 it must be the case that knowing $Y = 2$ may 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 because of the symmetry of the standard normal). So knowing $Y = 2$ makes X a discrete random variable with mass function $f(2) = f(-2) = 1/2$. The purpose of the rest of this (last) lecture is to expand on this topic.

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 – otherwise, it does not make sense to condition on the fact that we observed that $Y = y$!).

By the law of total probability, adding $f_{X,Y}(x, y)$ over all values of x gives the marginal $f_Y(y)$. Therefore, we see that when viewed as a function in x (with y being fixed), $f_{X|Y}(x\,|\,y)$ 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 24.1 (Example 15.1, continued)**.** In this example, the joint mass function of (X, Y), and the resulting marginal mass functions, were given by the following table:

[0] Last modified on May 12, 2020 at 13:15:34 -06'00'

$x \setminus y$	**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 Y, given that $X = 0$:

$$f_{Y|X}(0\,|\,0) = \frac{f_{X,Y}(0,0)}{f_X(0)} = \frac{16}{25},\ f_{Y|X}(1\,|\,0) = \frac{f_{X,Y}(0,1)}{f_X(0)} = \frac{8}{25},$$
$$f_{Y|X}(2\,|\,0) = \frac{f_{X,Y}(0,2)}{f_X(0)} = \frac{1}{25},\ f_{Y|X}(y\,|\,0) = 0 \text{ for other values of } y.$$

Organized in a table, this gives

y	**0**	**1**	**2**
$f_{Y\|X}(y\,\|\,0)$	16/25	8/25	1/25

Note how these entries add up to on, like a pdf should. Also, note how the probabilities in this table are nothing but the relative frequencies in the first row of the previous table (the one for the joint mass function).

Similarly,

y	**0**	**1**	**2**
$f_{Y\|X}(y\,\|\,1)$	8/10	2/10	0
$f_{Y\|X}(y\,\|\,2)$	1	0	0

One can also compute $f_{X|Y}$ by considering the relative frequencies in each column of the joint mass function table. Though in this particular example, due to the symmetry in the joint mass function, it turns out that $f_{X|Y}(x\,|\,y) = f_{Y|X}(y\,|\,x)$. This is special to the current example and is not true in general.

Observe that if we know $f_{X|Y}$ and f_Y, then $f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)f_Y(y)$. Similarly, if we know $f_{Y|X}$ and f_X then $f_{X,Y}(x,y) = f_{Y|X}(y\,|\,x)f_X(x)$. This is really just 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). Here is an example.

Example 24.2. Let $X \sim \text{Poisson}(\lambda)$ and if for some integer value x we know $X = x$ then let $Y \sim \text{Binomial}(x,p)$. By the above observation, the joint mass function of X and Y is

$$f_{X,Y}(x,y) = f_{Y|X}(y\,|\,x)f_X(x) = \binom{x}{y}p^y(1-p)^{x-y}\cdot e^{-\lambda}\frac{\lambda^x}{x!},\quad 0 \leqslant y \leqslant x.$$

Here is an explanation of the binomial term in the above formula: if we know $X = x$ then Y is a Binomial(x,p). So the number of trials is x and the probability of success is p. The probability of $Y = y$ then equals $\binom{x}{y}p^y(1-p)^{x-y}$.

The mass function of Y is then the marginal

$$f_Y(y) = \sum_x f_{X,Y}(x,y) = \sum_{x=y}^{\infty} \binom{x}{y} p^y (1-p)^{x-y} \cdot e^{-\lambda} \frac{\lambda^x}{x!}$$
$$= \frac{p^y \lambda^y}{y!} e^{-\lambda} \sum_{x=y}^{\infty} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} = e^{-\lambda p} \frac{(\lambda p)^y}{y!}.$$

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 lands 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. It is thus reasonable to expect that this is again a Poisson random variable. Its average though is λp (since you had on average λ people originally and then only a fraction p of them stayed).

2. Conditional expectations

Define conditional expectations, as we did ordinary expectations. But use conditional probabilities in place of ordinary probabilities. So for a function h we have

$$E[h(X) \,|\, Y = y] = \sum_x h(x)\, f_{X|Y}(x \,|\, y).$$

Example 24.3 (Example 24.1, continued)**.** Here,

$$E[X \,|\, Y = 1] = 0 \times \frac{8}{10} + 1 \times \frac{2}{10} = \frac{2}{10} = \frac{1}{5}.$$

Similarly,

$$E[X \,|\, Y = 0] = 0 \times \frac{16}{25} + 1 \times \frac{8}{25} + 2 \times \frac{1}{25} = \frac{10}{25} = \frac{2}{5},$$

and

$$E[X \,|\, Y = 2] = 0.$$

Note that $E[X] = 0 \times 25/36 + 1 \times 10/36 + 2 \times 1/36 = 12/36 = 1/3$, which is none of the conditional expectations we just computed. If you know, for example, that $Y = 0$, then your best guess of X is $E[X \,|\, Y = 0] = 2/5$. But if you have no knowledge of the value of Y, then your best guest for X is $E[X] = 1/3$.

Conditional expectations ($E[X \,|\, Y = y]$) and the unconditional one ($E[X]$) are in fact related by Bayes' formula. Namely,

$$E[X] = \sum_x x\, P\{X = x\} = \sum_x x \sum_y P\{X = x, Y = y\}$$
$$= \sum_{x,y} x\, P\{X = x \,|\, Y = y\} P\{Y = y\} = \sum_y \Big(\sum_x x\, f_{X|Y}(x \,|\, y) \Big) P\{Y = y\}$$
$$= \sum_y E[X \,|\, Y = y] P\{Y = y\}$$

Applied to the above example this says

$$\begin{aligned} E[X] &= E[X \mid Y=0]P\{Y=0\} + E[X \mid Y=1]P\{Y=1\} + E[X \mid Y=2]P\{Y=2\} \\ &= \frac{2}{5} \times \frac{25}{36} + \frac{1}{5} \times \frac{10}{36} + 0 \times \frac{1}{36} = \frac{1}{3}. \end{aligned}$$

Example 24.4. 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 $f_{X|Y}(x \mid y)$. (That is, given that we know how many 6's we got, we want to compute the probabilities for the number of 3's.) The possible values for Y are the integers from 0 to n. If we know $Y = y$, for $y = 0, \ldots, n$, then the number of 3's cannot exceed $n - y$ and so 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 (in the remaining $n - y$ rolls, none of which is a 6) is

$$f_{X|Y}(x \mid y) = \binom{n-y}{x} \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{n-y-x},$$

for $x = 0, \ldots, n - y$ (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, if you wish, also compute $f_{X|Y}(x \mid y)$ using its definition, i.e. $f_{X,Y}(x, y)/f_Y(y)$.) Now, the expected value of X, given $Y = y$, is clear: $E[X \mid Y = y] = (n - y)/5$, for $y = 0, \ldots, n$.

Example 24.5 (Example 12.3, 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

$$E[W] = \sum_{x=1}^{6} E[W \mid X = x]P\{X = x\} = \sum_{x=1}^{6} \Big(-\frac{x}{2}\Big)\Big(\frac{1}{6}\Big) = -\frac{7}{4},$$

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.

3. Conditional probability density functions

We are now given two continuous random variables X and Y with probability density functions f_X and f_Y, respectively. For all y, define the conditional pdf of X given that $Y = y$ by

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \tag{24.1}$$

provided that $f_Y(y) > 0$.

As a function in x, $f_{X|Y}(x \mid y)$ is a probability density function. That is:

(1) $f_{X|Y}(x \mid y) \geqslant 0$;

(2) $\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\, dx = 1$.

Example 24.6. 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 $X \sim \text{Exponential}(1)$. 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 24.7. In the previous example the random variables were described starting from their joint pdf. Now, let us tell the story of the previous example in reverse and describe the pair of random variables by first describing Y and then the conditional of X given we know Y. So Y is a random variable with pdf $f_Y(y) = ye^{-y}$, $y > 0$. And then given that $Y = y > 0$, X is a uniform random variable on $(0, y)$. This means that X has the conditional pdf $f_{X|Y}(x\,|\,y) = \frac{1}{y}$, $0 < x < y$. Then, we can deduce that the joint pfd is $f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)f_Y(y) = e^{-y}$, for $0 < x < y$. From that we can deduce the pdf of X:

$$f_X(x) = \int_x^\infty e^{-y}\,dy = e^{-x},\ x > 0.$$

So $X \sim \text{Exponential}(1)$.

4. Conditional expectations in the continuous case

Just like in the discrete case, we define conditional expectations as we did ordinary expectations. But we use conditional probabilities in place of ordinary probabilities: for a function h

$$E[h(X)\,|\,Y = y] = \int_{-\infty}^{\infty} h(x)\,f_{X|Y}(x\,|\,y)\,dx.$$

Example 24.8 (Example 24.7, continued)**.** If we are given that $Y = y > 0$, then $X \sim \text{Uniform}(0, y)$. This implies that $E[X\,|\,Y = y] = y/2$. Now, say we are given that $X = x > 0$. Then, to compute $E[Y\,|\,X = x]$ we need to find

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)},\ 0 \leqslant x < y.$$

As a consequence, given $X = x > 0$,

$$E[Y\,|\,X = x] = \int_x^\infty ye^{-(y-x)}\,dy = \int_0^\infty (z + x)e^{-z}\,dz = 1 + x.$$

We can also compute, for $x > 0$, things like:

$$E[e^{Y/2}\,|\,X = x] = \int_x^\infty e^{y/2}e^{-(y-x)}\,dy = e^x\int_x^\infty e^{-y/2}\,dy = 2e^xe^{-x/2} = 2e^{x/2}.$$

In the continuous case Bayes' formula becomes:

$$E[Y] = \int_{-\infty}^{\infty} E[Y\,|\,X = x]\,f_X(x)\,dx.$$

The proof of this formula is similar to the proof in the discrete case. (Do it!)

In the above example this says that

$$E[Y] = \int_0^\infty (1 + x)e^{-x}\,dx = 2$$

and a direct computation of the expected value indeed confirms this result:

$$E[Y] = \int_{-\infty}^{\infty} y\, f(y)\, dy = \int_0^{\infty} y^2 e^{-y}\, dy = 2.$$

(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: $1 + 1^2 = 2$. You can also use the Γ function for this!)

5. Independence

We have seen that X and Y are independent (continuous or discrete) random variables if and only if the joint mass function or pdf $f_{X,Y}(x, y)$ factors into a function of x times a function of y and in that case we have $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. This is equivalent to having

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x)$$

and also to having

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

This is reasonable: if information about Y does not change the statistics of X (which means that $f_{X|Y}(x\,|\,y)$ does not really depend on y) then X and Y must be independent and vice versa.

Example 24.9. Let (X, Y) have density $f(x, y) = e^{-y}$ when $0 \leqslant x \leqslant y$ and 0 otherwise. We want to find the pdf of $Z = Y - X$ given $X = x$. For this let us first compute the marginal

$$f_X(x) = \int_x^{\infty} e^{-y}\, dy = e^{-x}, \quad \text{for } x > 0.$$

Next, we compute the conditional pdf of Y given $X = x$ for $x > 0$:

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, \quad \text{for } y > x.$$

Now, we observe that if we are given that $X = x > 0$, then $Z = Y - x$ and the pdf of Z (given $X = x$) can be computed for example using the CDF method by writing:

$$P(Z \leqslant z\,|\,X = x) = P(Y - x \leqslant z\,|\,X = x) = P(Y \leqslant x + z\,|\,X = x) = \int_{-\infty}^{x+z} f_{Y|X}(y\,|\,x)\, dy.$$

Differentiating in z we get that for $x > 0$

$$f_{Z|X}(z\,|\,x) = f_{Y|X}(x + z\,|\,x) = e^{-(x+z-x)} = e^{-z} \quad \text{for } z > 0.$$

Now we see that the pdf of Z given $X = x$ does not depend on x. Therefore, we conclude that $Z = Y - X$ and X are independent, a fact that is not at all clear at the outset.

6. 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:

$$f_{X|Y\in B}(x) = P\{X = x\,|\,Y \in B\} = \frac{P\{X = x, Y \in B\}}{P\{Y \in B\}} = \frac{\sum_{y\in B} f_{X,Y}(x, y)}{\sum_{y\in B} f_Y(y)}\,.$$

The analogous formula in the continuous case is for the pdf of X, given $Y \in B$:

$$f_{X|Y\in B}(x) = \frac{\int_B f_{X,Y}(x,y)\,dy}{\int_B f_Y(y)\,dy}. \tag{24.2}$$

Once we know the pdf (or mass function), formulas for expected values become clear:

$$E[X\,|\,Y \in B] = \sum_x x\, f_{X|Y\in B}(x) = \frac{\sum_x \sum_{y\in B} x\, f_{X,Y}(x,y)}{\sum_{y\in B} f_Y(y)},$$

in the discrete case, and

$$E[X\,|\,Y \in B] = \int_{-\infty}^{\infty} x\, f_{X|Y\in B}(x)\,dx = \frac{\int_{-\infty}^{\infty} x\left(\int_B f_{X,Y}(x,y)\,dy\right)dx}{\int_B f_Y(y)\,dy}, \tag{24.3}$$

in the continuous case. Observe that this can also be written as:

$$\begin{aligned} E[X\,|\,Y \in B] &= \frac{\int_B \left(\int_{-\infty}^{\infty} x\, f_{X|Y}(x\,|\,y)\,dx\right) f_Y(y)\,dy}{P\{Y\in B\}} \\ &= \frac{\int_B E[X\,|\,Y=y]\, f_Y(y)\,dy}{P\{Y\in B\}}. \end{aligned} \tag{24.4}$$

Example 24.10. Let (X,Y) have joint density function $f_{X,Y}(x,y) = e^{-y}$, $0 < x < y$. We want to find the expected value of X, conditioned on $Y \leqslant 5$. First, we find the conditional pdf. One part we need to compute is $P\{Y \leqslant 5\}$. The pdf of Y is

$$f_Y(y) = \int_0^y e^{-y}\,dx = ye^{-y},\ y > 0,$$

and, using integration by parts, we have

$$P\{Y \leqslant 5\} = \int_0^5 ye^{-y}\,dy = 1 - 6e^{-5}.$$

Now, we can go ahead with computing the conditional pdf using (24.3). If $Y \leqslant 5$, then also $X < 5$ (since $X < Y$) and

$$f_{X|Y\leqslant 5}(x) = \frac{\int_0^5 f_{X,Y}(x,y)\,dy}{1-6e^{-5}} = \frac{\int_x^5 e^{-y}\,dy}{1-6e^{-5}} = \frac{e^{-x}-e^{-5}}{1-6e^{-5}},\ 0 < x < 5.$$

(Check that this pdf integrates to 1!) Finally, using integration by parts, we can compute:

$$E[X\,|\,Y \leqslant 5] = \int_{-\infty}^{\infty} x\, f_{X|Y\leqslant 5}(x)\,dx = \frac{\int_0^5 x(e^{-x}-e^{-5})\,dx}{1-6e^{-5}} = \frac{1-18.5e^{-5}}{1-6e^{-5}} \approx 0.912.$$

Remark 24.11. We have $f_X(x) = \int_x^\infty e^{-y}dy = e^{-x}$, $x > 0$. Thus, $X \sim$ Exponential(1) and $E[X] = 1$. Note next that the probability that $Y \leqslant 5$ is $1 - 6e^{-5} \approx 0.96$, which is very close to 1. So knowing that $Y \leqslant 5$ gives very little information. This explains why $E[X\,|\,Y \leqslant 5]$ is very close to $E[X]$. Try to compute $E[X\,|\,Y \leqslant 1]$ and see how it is not that close to $E[X]$ anymore. Try also to compute $E[X\,|\,Y \leqslant 10]$ and see how it is even closer to $E[X]$ than $E[X\,|\,Y \leqslant 5]$.

We could have done things in a different order to find $E[X\,|\,Y \leqslant 5]$. First, we find the conditional expectation $E[X\,|\,Y = y]$. To do so, we need to find $f_{X|Y}$ and thus to find first $f_Y(y) = \int_0^y e^{-y}\,dy = ye^{-y}$, $y > 0$. Hence, $f_{X|Y}(x\,|\,y) = 1/y$, for $0 < x < y$. Now, we see

that $\mathrm{E}[X \mid Y = y] = \int_0^y x \frac{1}{y}\,dx = \frac{y}{2}$. (This, of course, is not surprising since given $Y = y$ we found that $X \sim \mathrm{Uniform}(0, y)$.) Finally, we can apply (24.4) and use integration by parts to compute:

$$\mathrm{E}[X \mid Y \leqslant 5] = \frac{\int_0^5 \frac{y}{2} y e^{-y}\,dy}{\mathrm{P}\{Y \leqslant 5\}} = \frac{\frac{1}{2}\int_0^5 y^2 e^{-y}\,dy}{\int_0^5 y e^{-y}\,dy} = \frac{1 - 18.5e^{-5}}{1 - 6e^{-5}}\,.$$

Read sections 10.1, 10.2, and 10.3 in the textbook by Anderson, Sepäläinen, and Valkó.

Homework Problems

Exercise 24.1. 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 24.2. Suppose X has pdf $f_X(x) = 1/x^2$ when $x \geqslant 1$ and 0 otherwise. If $X = x$, let Y be uniformly distributed between 0 and x. Find the pdf of Y.

Exercise 24.3. Let (X, Y) have joint pdf $f(x, y) = k|x|$ if $-1 \leqslant y \leqslant x \leqslant 1$ and 0 elsewhere. Find k to make this a legitimate pdf. Then find the marginal pdfs of X and Y, the conditional pdf of Y given $X = x$, and the conditional pdf of X given $Y = y$.

Exercise 24.4. 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 $X \leqslant 1/2$.

Exercise 24.5. 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 24.6. The density for the time T required for the failure of a light bulb is $f(t) = \lambda e^{-\lambda t}$, $t \geqslant 0$. Fix $s > 0$. Find the conditional density function for $T - s$, given that $T > s$, and interpret the result intuitively.

Exercise 24.7. Let X and Y be independent random variables, each with pdf $f(x) = (1/2)e^{-x}$ if $x \geqslant 0$, $f(x) = 1/2$ if $-1 \leqslant x \leqslant 0$, and $f(x) = 0$ if $x < -1$. Let $Z = X^2 + Y^2$. Find $E[Z \,|\, X = x]$.

Exercises 10.1, 10.4, 10.5, 10.6, 10.7, and 10.10 on pages 366, 367, and 368 in the textbook by Anderson, Sepäläinen, and Valkó.

Solutions

Exercise 1.1 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.

Exercise 1.2 We can choose $\Omega = \{\mathtt{B},\mathtt{G},\mathtt{R}\}$ where $\mathtt{B}$ denotes the black chip, $\mathtt{G}$ the green one and $\mathtt{R}$ the red one.

Exercise 1.3 (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 1.4 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 1.5

(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 with the alphabetical part starting with an A.

Exercise 2.1

(a) $\{4\}$

(b) $\{0, 1, 2, 3, 4, 5, 7\}$

(c) $\{0, 1, 3, 5, 7\}$

(d) $\varnothing$

Exercise 2.2

(a) Let $x \in (A \cup B) \cup C$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cup B) \cup C &\Leftrightarrow x \in A \cup B \text{ or } x \in C \\ &\Leftrightarrow x \in A \text{ or } x \in B \text{ or } x \in C \\ &\Leftrightarrow x \in A \text{ or } x \in (B \cup C) \\ &\Leftrightarrow x \in A \cup (B \cup C) \end{aligned}$$

This proves the assertion.

(b) Let $x \in A \cap (B \cup C)$. Then we have the following equivalences:

$$\begin{aligned} x \in A \cap (B \cup C) &\Leftrightarrow x \in A \text{ and } x \in B \cup C \\ &\Leftrightarrow (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C) \\ &\Leftrightarrow (x \in A \cap B) \text{ or } (x \in A \cap C) \\ &\Leftrightarrow x \in (A \cap B) \cup (A \cap C) \end{aligned}$$

This proves the assertion.

(c) Let $x \in (A \cup B)^c$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cup B)^c &\Leftrightarrow x \notin A \cup B \\ &\Leftrightarrow (x \notin A \text{ and } x \notin B) \\ &\Leftrightarrow x \in A^c \text{ and } x \in B^c \\ &\Leftrightarrow x \in A^c \cap B^c \end{aligned}$$

This proves the assertion.

(d) Let $x \in (A \cap B)^c$. Then we have the following equivalences:

$$\begin{aligned} x \in (A \cap B)^c &\Leftrightarrow x \notin A \cap B \\ &\Leftrightarrow (x \notin A \text{ or } x \notin B) \\ &\Leftrightarrow x \in A^c \text{ or } x \in B^c \\ &\Leftrightarrow x \in A^c \cup B^c \end{aligned}$$

This proves the assertion.

Exercise 2.3

(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) $A \cup B \cup C$

(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 2.4 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 $(A \cap C) \cap (B \cap C) = A \cap B \cap C = \varnothing \cap C = \varnothing$. 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 2.5 It is enough to prove the exercise for the case of two sets, i.e. $n = 2$. Once that is done one can use induction to prove the same result for any integer $n \geqslant 2$. When we have two sets $A_1 \subset A_2$, the elements of A_1 are also elements of A_2 and hence they are in both sets and are thus in the intersection $A_1 \cap A_2$. Conversely, the elements in $A_1 \cap A_2$ must be in A_1. Therefore, we have shown that $A_1 \subset A_1 \cap A_2 \subset A_1$ and hence $A_1 \cap A_2 = A_1$. A similar reasoning works to prove that $A_1 \cup A_2 = A_2$.

Exercise 2.6 There are 150 people who favor the Health Care Bill, do not approve of the President's performance and are not registered Democrats.

Exercise 2.7

(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 2.8 Write

$$A_1 \cap (A_2 \cup \cdots \cup A_n) = (A_1 \cap A_2) \cup \cdots (A_1 \cap A_n) = \varnothing \cup \cdots \varnothing = \varnothing.$$

Exercise 3.1 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,

$$P(A) = P(R \cup Y \cup G) = P(R) + P(Y) + P(G) = \frac{3 \times 5}{24 \times 18} + \frac{8 \times 7}{24 \times 18} + \frac{13 \times 6}{24 \times 18} = \frac{149}{432} = 0.345.$$

Exercise 3.2 Let's consider the case where we toss a coin twice and let $A = \{$we get H on the first toss$\}$ and $B = \{$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 3.3 See Example 1.25 in the textbook by Anderson, Seppäläinen, and Valkó.

Exercise 4.1 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 4.2 Suppose the rooks are numbered 1 through 8. Then there are $64 \cdot 63 \cdots 57$ ways to place them on the chessboard. If we want the rooks not to check each other they have to all go on different columns and on different rows. So first, choose a column for each rook. There are 8! ways to do that (first rook can go any where from A to H, second rook has now 7 options, etc). Then we place the rook that is on column A on one of the rows (8 ways). Next, we place the rook that is supposed to go on column B on one of the remaining rows (7 ways), and so on. In total we see there are $8! \times 8!$ ways to do this. So the probability the 8 rooks are not checking each other is $8! \times 8!/(64 \cdots 57)$ as claimed.

Exercise 4.3 Let us think of the group of women as one entity. Hence, this problem boils down to arranging $m + 1$ objects in a row (m men and one group of women). We have $(m + 1)!$ ways to do it. Then, within the group of women, we have $w!$ ways to seat them. As we have $(m + w)!$ ways to seat the $m + w$ people, the probability is

$$P(\text{women together}) = \frac{(m + 1)!w!}{(m + w)!}.$$

Exercise 4.4 It is easier to compute the probability all 15 are good and then subtract that from 1. So the answer is

$$1 - \frac{75 \cdot 74 \cdot \cdots \cdot 61}{100 \cdot 99 \cdots 87 \cdot 86}.$$

Exercise 4.5 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 4.6

(a) For the draw with replacement, there are 52^{10} possible hands. If we want no two cards to have the same face value, 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, to draw exactly 9 cards of the same suite we have $4 \cdot 13 \cdot 12 \cdots 5$ ways (4 to pick the suite and then 13 for the the face value of the first card, 12 for the second, etc). Then for the last card we have $3 \cdot 13$ ways (3 to pick the suite and 13 face values). But then we have to decide where this remaining card goes (first, second, third, etc?). We have 10 options for that. (We do not need to do this for the 9 cards because the way we counted them took their order into account.) So there are $4 \cdot 13 \cdot 12 \cdots 5 \cdot 3 \cdot 13 \cdot 10$ ways to get 10 cards with exactly 9 that are of the same suite. Similarly, there are $4 \cdot 13 \cdot 12 \cdots 4$ ways to get 10 cards of the same suite. So the probability in question is

$$\frac{4 \cdot 13 \cdot 12 \cdots 5 \cdot 3 \cdot 13 \cdot 10 + 4 \cdot 13 \cdot 12 \cdots 4}{52 \cdot 51 \cdots 43} = 0.00000712.$$

We will see in the next lecture how to count things when order does not matter. Then we could have solved part (b) like this: We have $\binom{52}{10}$ possible hands. The number of hands that

have exactly 9 cards of the same suit is $4 \cdot \binom{13}{9} \cdot 39$ (4 possible suits, 9 cards out of this suit and 1 additional card from the 39 remaining) and the number of hands that have 10 cards of the same suite is $4 \cdot \binom{13}{10}$. Hence, the probability is

$$\frac{4 \cdot \binom{13}{9} \cdot 39 + 4 \cdot \binom{13}{10}}{\binom{52}{10}} = 0.00000712.$$

Exercise 4.7 First we will pick the three balls that have numbers smaller than 8. There are $7 \cdot 6 \cdot 5$ ways to do that. Notice that this way of counting takes the order of the balls into account. Next, we pick the ball that has a number larger than 8. There are two ways to do that. But then we have to decide on the order of the ball relative to the three we already picked (recall, those were already ordered). So there are 4 possible spots for this ball. Lastly we have the ball numbered 8. For this one we have to decide on which ball it is, relative to the four we already picked. There are 5 options for this. So in total, there are $7 \cdot 6 \cdot 5 \cdot 2 \cdot 4 \cdot 5$ ways to pick the balls (in order) so that the second largest number is an 8. The probability is then

$$\frac{7 \cdot 6 \cdot 5 \cdot 2 \cdot 4 \cdot 5}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} = 0.2778.$$

Again, we can do this without regard to order. Then we have $\binom{10}{5}$ possible ways to draw the balls. And 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 is

$$\frac{\binom{2}{1} \cdot \binom{7}{3}}{\binom{10}{5}} = 0.2778.$$

Exercise 4.8 Let A be the event we get exactly three aces and K the event we get exactly three kings. We are after $P(A \cup K)$. For this we need to compute $P(A)$, $P(K)$, and $P(A \cap K)$.

To compute the probability of A we need to count the number of ways we can get exactly three aces. First, we decide on which of the four aces are going to be among the eight cards. There are four ways to do this. Indeed, it is enough to tell which ace will not be chosen. Then we need to decide on the placement of these three aces among the eight cards. So we need to choose three distinct spots (for the three distinct aces) among the eight possible spots. There are $8 \cdot 7 \cdot 6$ ways to do that. Once that is done, we need to pick the remaining five cards (whose spots are now known, since we have allocated which of the eight cards will be an ace). There are 48 options for the first of the remaining spots, 47 for the second, etc. So in total, there are $4 \cdot 8 \cdot 7 \cdot 6 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44$ ways to get exactly three aces. Therefore,

$$P(A) = \frac{4 \cdot 8 \cdot 7 \cdot 6 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45}.$$

Since aces and kings are just names, $P(K)$ is equal to $P(A)$.

To compute the probability of $A \cap K$ we need to count the number of ways we can get exactly three aces and three kings. The computation is similar to the above and we get

$$P(A \cap K) = \frac{4 \cdot 8 \cdot 7 \cdot 6 \cdot 4 \cdot \cdot 5 \cdot 4 \cdot 3 \cdot 44 \cdot 43}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45}.$$

The final answer is then

$$P(A \cup K) = \frac{4 \cdot 8 \cdot 7 \cdot 6 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 2 - 4 \cdot 8 \cdot 7 \cdot 6 \cdot 4 \cdot 5 \cdot 4 \cdot 3 \cdot 44 \cdot 43}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45}.$$

Again, once we learn in the next lecture how to count without regard to order, we can compute the above in a different way. Namely, for $\mathsf{P}(\mathsf{A})$ we say that there are $\binom{4}{3}$ ways to pick the three aces and $\binom{48}{5}$ to pick the remaining five cards and there are $\binom{52}{8}$ ways to pick eight cards out of 52. So

$$\mathsf{P}(\mathsf{A}) = \frac{\binom{4}{3} \cdot \binom{48}{5}}{\binom{52}{8}}.$$

As above, $\mathsf{P}(\mathsf{K}) = \mathsf{P}(\mathsf{A})$ and then for $\mathsf{P}(\mathsf{A} \cap \mathsf{K})$ we say that there are $\binom{4}{3} \cdot \binom{4}{3}$ ways to pick the three aces and the three kings and then $\binom{44}{2}$ ways to pick the remaining two cards. Hence,

$$\mathsf{P}(\mathsf{A} \cap \mathsf{K}) = \frac{\binom{4}{3} \cdot \binom{4}{3} \cdot \binom{44}{2}}{\binom{52}{8}}.$$

And the final answer is

$$\mathsf{P}(\mathsf{A} \cup \mathsf{K}) = \frac{\binom{4}{3} \cdot \binom{48}{5} - \binom{4}{3} \cdot \binom{4}{3} \cdot \binom{44}{2}}{\binom{52}{8}}.$$

It is left to the student to compute all of the above numbers and check that the final answer is the same whether we compute things with order or without order.

Exercise 5.1

(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 5.2

(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 prize, you would definitely choose the first one which has a larger probability of winning.

Exercise 5.3 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

$$P(\text{four of a kind}) = \frac{13 \cdot 48}{\binom{52}{5}} = 0.00024.$$

(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

$$P(\text{three of a kind}) = \frac{13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4^2}{\binom{52}{5}} = 0.0211.$$

(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

$$P(\text{straight flush}) = \frac{10 \cdot 4}{\binom{52}{5}} = 0.0000154.$$

(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

$$P(\text{flush}) = \frac{4 \cdot \binom{13}{5} - 40}{\binom{52}{5}} = 0.00197.$$

(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

$$\mathrm{P}(\text{straight}) = \frac{10 \cdot 4^5 - 40}{\binom{52}{5}} = 0.00392.$$

Exercise 5.4

(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

$$w\mathfrak{m}w\mathfrak{m}w\mathfrak{m}w\mathfrak{m} \text{ or } \mathfrak{m}ww\mathfrak{m}w\mathfrak{m}w\mathfrak{m}w,$$

where w stands for a woman and $\mathfrak{m}$ 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 5.5

(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 from 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 5.6

(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 5.7

(a) For each digit, except the zero, we can build a 4-digit number. Hence, there are 9 possible numbers with identical digits.

(b) Two cases can occur. First, the number of ways to build a 4-digit number that has two pairs of 2 identical digits, different from 0, is $\binom{9}{2}\binom{4}{2}$. Indeed, we choose two digits among 9 and two places among four to place say the smaller pair. 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 $9 \cdot 8 \cdot 7 \cdot 6$ numbers with 4 different digits without 0. There are $9 \cdot 8 \cdot 7 \cdot 3$ 4-digit numbers with a 0. Indeed, there are 3 positions for the 0 and then there are $9 \cdot 8 \cdot 7$ three digit numbers with non-0 digits. Hence, there are $9 \cdot 8 \cdot 7 \cdot 6 + 9 \cdot 8 \cdot 7 \cdot 3 = 4536$ 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. Alternatively, we can say that there are $\binom{10}{4}$ ways to choose 4 different digits (which we can then put in order from left to right) and $\binom{9}{3}$ ways to choose 0 and 3 different non-0 digits. And then the number we are after is $\binom{10}{4} - \binom{9}{3}$ which happens to equal $\binom{9}{4}$.

(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 no n-digit numbers with strictly increasing digits when $n \geqslant 10$ (since some digits must repeat in this case). 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 5.8

(a) The equality is immediate from the definition of $\binom{n}{k}$. But here is another way to see it. 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*.

(b) One can prove this using algebra. But here is a 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.

Exercise 5.9

(a) The hint says it all!

(b) There are $\binom{n}{k}$ subsets of size k. This and the hint answer the question.

(c) There are $\binom{n}{j}$ ways to pick j of the n possible $(x+y)$'s. This is hence the coefficient in front of $x^j y^{n-j}$.

(d) It is $\binom{7}{3}2^3(-4)^4 = 71680$.

Exercise 6.1 To show that $\mathbb{Z}_+$ (the set of nonnegative integers) is countable we need to give a bijection between it and the natural numbers $\mathbb{N}$ (i.e. positive integers). One simple bijection is this: $g : \mathbb{N} \to \mathbb{Z}_+ : k \mapsto k - 1$. (Check that this is a bijection between $\mathbb{N}$ and $\mathbb{Z}_+$.) In words, this means we "count" $\mathbb{Z}_+$ by saying that 0 is the first number, 1 is the second, and so on.

Similarly, a bijection from $\mathbb{N}$ to $\mathbb{Z}$ (the set of integers) is this: map k to $k/2$ if k is an even positive integer and to $-(k-1)/2$ if it is an odd positive integer. (Again, check that this is a bijection between $\mathbb{N}$ and $\mathbb{Z}$.) In words, this means we say 0 is the first number ($k = 1$ is mapped to $-(1-1)/2 = 0$), 1 is the second number ($k = 2$ is mapped to $2/2 = 1$), -1 is the third number, (3 is mapped to $-(3-1)/2 = -1$), 2 is the fourth number (4 is mapped to $4/2 = 2$), and so on, alternating between negative and positive integers as we go.

Exercise 6.2 A state space an be taken to be $\Omega = \mathbb{N} \cup \{\infty\}$, the positive integers and "infinity". Infinity is added to denote the case when a 4 never appears.

When asking about the probability a 4 appears for the first time on the n-th roll, it is enough to consider that we rolled the die n times. There are 6^n possible outcomes in this case, all of which are equally likely. As to the desired outcomes they are the outcomes in which 4 did not come up in the first $n-1$ rolls and then came up in the n-th roll. So there are $5^{n-1} \times 1$ desired outcomes. The probability of a 4 appearing on the n-th roll for the first time is thus $\frac{5^{n-1}}{6^n} = \frac{1}{6} \cdot (\frac{5}{6})^{n-1}$.

Applying (6.3) with $m = 0$ and $a = 5/6$ we get that

$$\sum_{n=1}^{\infty} \frac{1}{6} \cdot (\frac{5}{6})^{n-1} = \frac{1}{6} \sum_{n=1}^{\infty} (\frac{5}{6})^{n-1} = \frac{1}{6} \sum_{k=0}^{\infty} (\frac{5}{6})^{k} = \frac{1}{6} \cdot \frac{1}{1-5/6} = 1.$$

So adding the probabilities over all positive integers gives one and leaves the probability of the remaining outcome "4 never appears" to equal 0. In particular, this shows that a 4 will appear at some point, with probability 100%.

Exercise 7.1 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 7.2 Let W_2 be the event that the first two balls are white and let W_6 be the event that the sample contains exactly six white balls. We want to find $P(W_2 \,|\, W_6) = P(W_2 \cap W_6)/P(W_6)$. So we need to compute the two probabilities appearing in the ratio.

First, we work out the case of drawing without replacement. The probability of the first six balls are white and the remaining four are black is $\frac{30\times29\times28\times27\times26\times25\times15\times14\times13\times12}{45\times\cdots\times36}$. This is also the probability the last six balls are white and the remaining ones are black, or any other combination in which there are six white balls and four black ones. So the probability the sample has six white and four black balls in some order is the ratio we just computed, multiplied by the number of ways we could arrange 6 white balls and 4 black ones. This number of arrangements is the same as the number of ways in which we can pick 6 places (for the white balls) out of 10 available slots. So it is $\binom{10}{6}$. Thus,

$$P(W_6) = \frac{30 \times 29 \times 28 \times 27 \times 26 \times 25 \times 15 \times 14 \times 13 \times 12}{45 \times \cdots \times 36} \cdot \binom{10}{6}.$$

To compute $P(W_2 \cap W_6)$ we see that we can repeat the above logic, but now we only multiply by the number of ways we can arrange 4 white balls and 4 black ones, because the first two of the ten balls are now known to be white. So

$$P(W_6) = \frac{30 \times 29 \times 28 \times 27 \times 26 \times 25 \times 15 \times 14 \times 13 \times 12}{45 \times \cdots \times 36} \cdot \binom{8}{4}.$$

The answer, in the case of drawing with replacement, is therefore

$$P(W_2 \,|\, W_6) = \frac{\binom{8}{4}}{\binom{10}{6}}.$$

For the case of drawing with replacement the probability the computation goes similarly. When we know the places of the six white balls and the four black ones, the probability of drawing such a combination is $\frac{30^6\times15^4}{45^{10}}$. So

$$P(W_6) = \frac{30^6 \times 15^4}{45^{10}} \cdot \binom{10}{6}.$$

Similarly,

$$P(W_2 \cap W_6) = \frac{30^6 \times 15^4}{45^{10}} \cdot \binom{8}{4}.$$

Consequently,

$$P(W_2 \,|\, W_6) = \frac{\binom{8}{4}}{\binom{10}{6}},$$

which is the same probability as the one for the case of draws without replacement.

Exercise 7.3 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 7.4 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 7.5 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 7.6 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 7.7 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

$$\begin{aligned} P(A_n) &= P(A_n|A_{n-1})P(A_{n-1}) + P(A_n|A_{n-1}^c)P(A_{n-1}^c) \\ &= \frac{1}{2}P(A_{n-1}) + \frac{1}{4}(1 - P(A_{n-1})) \\ &= \frac{1}{4}P(A_{n-1}) + \frac{1}{4}, \end{aligned}$$

where the event A_n^c 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

$$\begin{aligned} P(R_n) &= P(R_n|A_n)P(A_n) + P(R_n|A_n^c)P(A_n^c) \\ &= \frac{1}{2}P(A_n) + \frac{1}{4}(1 - P(A_n)) \\ &= \frac{1}{4}P(A_n) + \frac{1}{4} = P(A_{n+1}). \end{aligned}$$

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 8.1

(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 8.2 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 8.3

(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 8.4 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 8.5 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 8.6 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 8.7 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 9.1 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

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = HHH, \\ 1 & \text{if } \omega = HHT, HTH, THH, \\ 2 & \text{if } \omega = HTT, THT, TTH, \\ 3 & \text{if } \omega = TTT. \end{cases}$$

Exercise 9.2 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 9.3 The possible values of X are 0, 1, and 2. We cannot have three or more heads each followed by a tail if we are tossing the coin five times. Now, $X = 0$ means no heads were followed by a tail. So as soon as an heads appears, the remaining tosses must be heads. There are six possible such outcomes and $2^5 = 32$ overall possible outcomes. So

$$\begin{aligned} P(X=0) &= P(\{HHHHH, THHHH, TTHHH, TTTHH, TTTTH, TTTTT\}) \\ &= \frac{6}{32} = \frac{3}{16}. \end{aligned}$$

Similarly, $X = 1$ means we had one or more heads then one tails and then possibly several heads. The probability of each or these possibilities is different and there are a lot of these possibilities. Since the only remaining outcome is $X = 2$ we try to analyze its probability in the hope that it is easier to compute and then we can use that to compute $P(X = 1) = 1 - P(X = 0) - P(X = 2)$.

Indeed, $X = 2$ means we had HT twice and the number of possibilities for that is a lot smaller than for the case $X = 1$. Indeed, we have

$$\begin{aligned} P(X=2) &= P(\{HTHTT, HTHTH, HTHHT, HTTHT, HHTHT, THTHT\}) \\ &= \frac{6}{32} = \frac{3}{16}. \end{aligned}$$

And this leaves us with

$$P(X=1) = 1 - \frac{3}{16} - \frac{3}{16} = \frac{5}{8}.$$

Exercise 9.4 $P(X=k) = \binom{3}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{3-k}$, 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 9.5

(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 9.6 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 9.7

(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 10.1

(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 $p = 0.002$. So we could in fact compute $P\{X = 15\}$ using the Binomial mass function:

$$P\{X = 15\} = \binom{10000}{15}(0.002)^{15}(0.998)^{9985} \simeq 0.0516 = 5.16\%.$$

However, since $p = 0.002$ is small, $n = 10,000$ is large and $np = 20$ is not too large, nor too small, we can approximate X by a Poisson random variable with parameter $\lambda = np = 20$. Then, we have

$$P\{X = 15\} \simeq e^{-\lambda}\frac{\lambda^{15}}{15!} = e^{-20}\frac{20^{15}}{15!} \simeq 0.0516 = 5.16\%.$$

(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 $p = 0.002$. So we can again use the Binomial distribution and get

$$P(Y = 3) = \binom{2000}{3}(0.002)^{3}(0.998)^{1997} \simeq 0.1955 = 19.55\%.$$

But we can also approximate by a Poisson distribution of parameter $\lambda = np = 4$. Then, we have

$$P\{Y = 3\} \simeq e^{-\lambda}\frac{\lambda^{3}}{3!} = e^{-4}\frac{4^{3}}{3!} \simeq 0.1954 = 19.54\%.$$

Exercise 2.20 in the textbook

(a) $(5/6)^5 + 5 \cdot (1/6) \cdot (5/6)^4 + 10 \cdot (1/6)^2 \cdot (5/6)^3$.

(b) $(5/6)^4$.

(c) $(5/6)^4 - (5/6)^{19}$.

Exercise 11.1 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, $c = \frac{3}{32}$.

Exercise 11.2 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, $c = \frac{4}{\pi}$.

Exercise 11.3 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 11.4 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.

Exercise 11.5 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.

Exercise 12.1 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 12.2 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

(Remember that the player gets the \$1 back if he wins.)

The probability mass function of H is

x	-5	-4	-1	4	14	34
$f_H(x) = P\{H = x\}$	$\frac{22}{52}$	$\frac{15}{52}$	$\frac{7}{52}$	$\frac{4}{52}$	$\frac{2}{52}$	$\frac{2}{52}$

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 a 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 12.3 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

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1}.$$

To compute this we repeat the calculation done in Exercise 12.5 but with the specific value $r = 1$. Namely, we write

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}\sum_{k=1}^{\infty} kp^2(1-p)^{k-1} = \frac{1}{p}\sum_{j=2}^{\infty}(j-1)p^2(1-p)^{j-2}.$$

For the last equality we just replaced k with $j-1$ and so while k goes from 1 and up, j goes from 2 and up.

Now observe that $f(j) = (j-1)p^2(1-p)^{j-2}$, $j = 2, 3, \ldots$, is the mass function of a negative binomial with parameters $r = 2$ and p. Therefore, the sum equals to 1 (since a mass function must add up to one) and so we are left with $E[X] = 1/p$.

In the above, we used a fact that was not proved when the negative binomial was introduced. Namely, that the mass function of a negative binomial adds up to one. Let us prove it in the case $r = 2$ (which is what we needed in this exercise). Namely, we want to show that

$$\sum_{k=1}^{\infty} kp^2(1-p)^{k-1} = 1.$$

Or, equivalently, that

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{1}{p^2}.$$

For this observe that $k(1-p)^{k-1}$ is the derivative with respect to p of $-(1-p)^k$. Hence, we can write

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = -\sum_{k=1}^{\infty}\Big((1-p)^k\Big)' = -\Big(\sum_{k=1}^{\infty}(1-p)^k\Big)' = -\Big(\frac{1}{p}\Big)' = \frac{1}{p^2}\,.$$

Note however that for the second equality we said that the derivative of an infinite sum of functions is the sum of the derivatives of the functions. This is a highly non-trivial fact that we are sweeping under the rug, because the tools required to prove it are beyond the scope of this class.

Exercise 13.1 We compute

$$\mathrm{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Also

$$\mathrm{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\cdot\frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

Exercise 13.2 First of all, if n is odd, we have

$$\mathrm{E}[X^n] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}}\,dx = 0,$$

by the symmetry of the function $x \mapsto x^n e^{-\frac{x^2}{2}}$ (the function is odd).

When n is even we proceed by induction. When $n = 0$ we have $\mathrm{E}[1] = 1$ and the formula is true. So now assume the result is true for all even numbers up to $n-2$ and let us compute $\mathrm{E}[X^n]$. Using an integration by parts, we have

$$\begin{aligned}
\mathrm{E}[X^n] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{n-1} x e^{-\frac{x^2}{2}}\,dx \\
&= \frac{1}{\sqrt{2\pi}}\left(-x^{n-1}e^{-\frac{x^2}{2}}\right)\Big|_{-\infty}^{\infty} + \frac{(n-1)}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{n-2}e^{-\frac{x^2}{2}}\,dx \\
&= (n-1)\mathrm{E}[X^{n-2}] = (n-1)\cdot(n-3)(n-5)\cdots 1.
\end{aligned}$$

Hence, by induction the result is true for every even integer n. In particular, notice that setting $n = 2$ gives that $\mathrm{E}[X^2] = 1$ for a standard normal random variable.

Exercise 13.3 By the definition of expectation and using integration by parts, we have

$$\begin{aligned}
\mathrm{E}[c(X)] &= \int_{-\infty}^{\infty} c(x)f(x)\,dx = 2\int_0^3 x e^{-x}\,dx + \int_3^{\infty}(2+6(x-3))xe^{-x}\,dx \\
&= 2\left(-xe^{-x}\right)\big|_0^3 + 2\int_0^3 e^{-x}\,dx + \left(-(2+6(x-3))xe^{-x}\right)\big|_3^{\infty} + \int_3^{\infty}(12x-16)e^{-x}\,dx \\
&= -6e^{-3} + 2\left(-e^{-x}\right)\big|_0^3 + 6e^{-3} + \left(-(12x-16)e^{-x}\right)\big|_3^{\infty} + \int_3^{\infty} 12e^{-x}\,dx \\
&= 2 - 2e^{-3} + 20e^{-3} + \left(-12e^{-x}\right)\big|_3^{\infty} \\
&= 2 + 18e^{-3} + 12e^{-3} = 2 + 30e^{-3}.
\end{aligned}$$

Exercise 14.1 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 15.1

(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 15.2

(a) The set of possible values for X_1 and X_2 is $\{1, \dots, 6\}$. By definition, we always have $X_1 \leqslant X_2$. We have to compute $f(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\}$. 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 $f(x_1, x_1) = \frac{1}{36}$ in this case. If $x_1 < x_2$, one die 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 $f(x_1, x_2) = \frac{1}{18}$. To summarize: 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).

$x_1 \backslash x_2$	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 15.3

(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).

$x_1 \backslash x_2$	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 16.1 We can see that the distribution of (X, Y) is uniform on the square $[-1, 1]^2$. Hence, we can use a ratio of areas 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 area of the domain corresponding to $XY > \frac{1}{4}$ as

$$2\int_{\frac{1}{4}}^{1} dx \int_{1/4x}^{1} dy = 2\int_{\frac{1}{4}}^{1} dx \left(1 - \frac{1}{4x}\right) = 2\left(x - \frac{\ln(x)}{4}\right)\Bigg|_{\frac{1}{4}}^{1} = \frac{3 - \ln(4)}{2}.$$

To compute the probability we divide by the area of the square, i.e. by 4, to get

$$P\{XY \leqslant \frac{1}{4}\} = 1 - P\{XY > \frac{1}{4}\} = 1 - \frac{(3 - \ln(4))/2}{4} = \frac{5 + \ln(4)}{8}.$$

(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 area $\frac{5}{4}$ 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 easily identify the surface as the union of two triangles of area $\frac{1}{2}$ 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 16.2 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 16.12 on page 106 in the Lecture Notes.

Exercise 16.3 The random variables X and Y have joint density function $f(x,y) = \frac{1}{4}$ if $-1 \leqslant x \leqslant 1$ and $-1 \leqslant y \leqslant 1$, $f(x,y) = 0$ otherwise. Hence,

$$\begin{aligned} E[\max(X,Y)] &= \frac{1}{4}\int_{-1}^{1} dx \int_{-1}^{1} dy\ \max(x,y) = \frac{1}{4}\int_{-1}^{1} dx \left(\int_{-1}^{x} dy\, x + \int_{x}^{1} dy\, y \right) \\ &= \frac{1}{4}\int_{-1}^{1} dx \left(x(x+1) + \frac{1-x^2}{2} \right) = \frac{1}{8}\int_{-1}^{1} dx\,(x+1)^2 = \frac{1}{8}\frac{(x+1)^3}{3}\Bigg|_{-1}^{1} = \frac{1}{3} \end{aligned}$$

Exercise 16.4

(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 16.5

(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 16.6 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 16.7 (a) To compute the probability we need to calculate the integral of the joint pdf over the region $\{(x,y) : x^2 + y \leqslant 1\}$. We also have to make sure x and y stay between 0 and 2 because the pdf is zero otherwise. Note that the region of integration says that for a given x, y goes from 0 to $1 - x^2$. This guarantees that y is less than 2, but if x happens to be in $(1,2)$, then $1 - x^2$ is negative and there is no room for y. (Recall, it has to be between 0 and 2.) So x has to vary from 0 to 1 and then y varies from 0 to $1 - x^2$. The upshot is that the probability we are after equals

$$\begin{aligned}
\frac{1}{8}\int_0^1 \left(\int_0^{1-x^2} (x+y)\,dy \right) dx &= \frac{1}{8}\int_0^1 \left(x(1-x^2) + \frac{1}{2}(1-x^2)^2 \right) dx \\
&= \frac{1}{16}\int_0^1 (2x - 2x^3 + 1 - 2x^2 + x^4)\,dx \\
&= \frac{1}{16}(1 - 1/2 + 1 - 2/3 + 1/5) = \frac{31}{480}.
\end{aligned}$$

For part (b) we need to use the definition of conditional probability. Say A is the event that exactly one of X and Y is $\leqslant 1$ and let B be the event that at least one of the two is $\leqslant 1$. Then we are after $P(A \mid B) = P(A \cap B)/P(B)$. Note that $A \subset B$ and so $P(A \cap B) = P(A)$. Also, to compute the probability of B it is easier to observe that the reverse of B is the event where both X and Y are > 1. So

$$\begin{aligned}
P(B) = 1 - P(B^c) &= 1 - \frac{1}{8}\int_1^2 \left(\int_1^2 (x+y)\,dy \right) dx \\
&= 1 - \frac{1}{8}\int_1^2 (x + 3/2)\,dx = 1 - \frac{1}{8}(3/2 + 3/2) \\
&= \frac{5}{8}.
\end{aligned}$$

The probability that exactly one of X and Y is $\leqslant 1$ is twice the probability that $X \leqslant 1$ and $Y > 1$. That is,

$$\begin{aligned} P(A) &= 2 \cdot \frac{1}{8} \int_0^1 \left(\int_1^2 (x+y)\, dy \right) dx \\ &= \frac{1}{4} \int_0^1 (x + 3/2)\, dx = \frac{1}{4}(1/2 + 3/2) = \frac{1}{2}. \end{aligned}$$

Consequently, $P(A \mid B) = \frac{1/2}{5/8} = 4/5$.

Finally, to answer part (c) observe that we just showed that the probability of $X \leqslant 1$ and $Y > 1$ is $P(A)/2 = 1/4$ and the probability of $X > 1$ and $Y > 1$ is $1 - 5/8 = 3/8$. Adding the two and using the law of total probability we find that $P(Y > 1) = 1/4 + 3/8 = 5/8$. So $P(Y \leqslant 1) = 3/8$. By symmetry $P(X \leqslant 1) = 3/8$. But now

$$P(X \leqslant 1)P(Y > 1) = \frac{3}{8} \cdot \frac{5}{8} \neq \frac{1}{4} = P(X \leqslant 1, Y > 1).$$

So the two variables are dependent. (We could also compute the pdfs of X and Y out of the joint pdf and check that the joint pdf is not a product of the two marginals.)

Exercise 17.1 First of all, 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 $E[Y] = 0$,

$$\mathrm{Var}(Y) = E[Y^2] = E[\cos^2(X)] = \frac{1}{2\pi}\int_0^{2\pi} \cos^2(x)\,dx = \frac{1}{4\pi}(x - \frac{\sin(2x)}{2})\Big|_0^{2\pi} = \frac{1}{2}.$$

Simiarly, 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,

$$P(Y > 1/2) = P(\cos(X) > 1/2) = P(0 < X < \pi/3) + P(5\pi/3 < X < 2\pi) = \frac{1}{3},$$

$$P(Z > 1/2) = P(\sin(X) > 1/2) = P(\pi/6 < X < 5\pi/6) = \frac{1}{3},$$

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 17.2 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 18.1

(a) It is not an mgf since it can take negative values and mgfs are supposed to be positive.

(b) It is not an mgf since $M(0) \neq 1$ and an mgf is supposed to be 1 at $t = 0$.

(c) It is the mgf of an exponential random variable with parameter $\lambda = 1$. (See Example 19.1)

(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 18.2 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 18.3

$$M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = E[e^{bt}e^{taX}] = e^{bt}E[e^{taX}] = e^{bt}M_X(at).$$

Exercise 18.4 We have

$$M_X(t) = f_X(0) + f_X(1)e^t + f_X(2)e^{2t} \quad \text{and} \quad M_Y(t) = f_Y(0) + f_Y(1)e^t + f_Y(2)e^{2t}.$$

If $M_X(t) = M_Y(t)$ for all t then

$$f_X(0) + f_X(1)e^t + f_X(2)e^{2t} = f_Y(0) + f_Y(1)e^t + f_Y(2)e^{2t}$$

for all t. Take $t \to -\infty$ to get that $f_X(0) = f_Y(0)$. Now that we know this the above equation becomes

$$f_X(1)e^t + f_X(2)e^{2t} = f_Y(1)e^t + f_Y(2)e^{2t}$$

for all t. Divide through by e^t then take again $t \to -\infty$ to get $f_X(1) = f_Y(1)$. Since X and Y take values 0, 1, and 2 only and we have shown that their mass functions match at 0 and at 1 and since the mass functions both have to add up to 1 this forces $f_X(2) = f_Y(2)$. So the two random variables have the same mass function.

This is a demonstration of Theorem 18.11 in a simple case.

Exercise 19.1

(a) The pdf is $\lambda e^{-\lambda x}$, for $x > 0$, and 0 otherwise. The mean is $1/\lambda$ and we are given that it equals 2.5. So $\lambda = 2/5 = 0.4$. The probability the random variable exceeds 3 is

$$\int_3^\infty \frac{2}{5} e^{-2x/5}\, dx = e^{-6/5} = e^{-1.2} = 0.3012.$$

(b) The probability none of the five earthquakes exceeds size 6, i.e. that all of them have size less than 6, is (by independence)

$$\Big(\int_0^6 \frac{2}{5} e^{-2x/5}\, dx\Big)^5 = (1 - e^{-2.4})^5 = 0.6215.$$

Then the probability that at least one of the five earthquakes exceeds size 6 is $1 - 0.6215 = 0.3784$.

(Note by the way that this problem did not consider modeling the time between earthquake strikes, which can be done using an exponential random variable, but then one needs to use data/mathematical models to estimate its parameter.)

Exercise 19.2 The mean is $\alpha/\lambda = 3.2$ and the variance is $\alpha/\lambda^2 = 6.4$.

Exercise 19.3 The pdf is given by $\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$ and 0 for $x \leqslant 0$. We need to figure out the parameters α and λ to give the precise formula of the pdf. We know that the mean is $\alpha/\lambda = 4$ and the variance is $\alpha/\lambda^2 = 8$. Taking a ratio we get that $\lambda = 1/2$. Then this gives $\alpha = 2$. Since $\Gamma(2) = 1! = 1$ we now get that the pdf is given by $\frac{1}{4} x e^{-x/2}$ for $x > 0$ and 0 for $x \leqslant 0$.

Exercise 19.4 The mean is $\alpha/\lambda = 20,000$ and the variance is $\alpha/\lambda^2 = 400,000$. (So the standard deviation from the mean is $\sqrt{400,000} = 632.46$, which is quite a small variation, compared to the size of the mean.)

Exercise 19.5 The mean is

$$E[L] = 30E[X] + 2E[X^2]$$

To compute the variance we will need to compte

$$E[L^2] = E[(30X + 2X^2)^2] = 900E[X^2] + 120E[X^3] + 4E[X^4].$$

So we need to know $E[X]$, $E[X^2]$, $E[X^3]$, and $E[X^4]$. For this, we will use the fact that we know the mgf is given by

$$M(t) = \Big(\frac{\lambda}{\lambda - t}\Big)^\alpha = (1 - 2t)^{-3}$$

and therefore

$$M'(t) = 6(1-2t)^{-4}, \quad M''(t) = 48(1-2t)^{-5},$$
$$M'''(t) = 480(1-2t)^{-6}, \quad \text{and} \quad M''''(t) = 5760(1-2t)^{-7}.$$

From this we get

$$E[X] = M'(0) = 6, \quad E[X^2] = M''(0) = 48,$$
$$E[X^3] = M'''(0) = 480, \quad \text{and} \quad E[X^4] = M''''(0) = 5760.$$

Now we can compute

$$E[L] = 30E[X] + 2E[X^2] = 30 \cdot 6 + 2 \cdot 48 = 276,$$
$$E[L^2] = 900E[X^2] + 120E[X^3] + 4E[X^4] = 900 \cdot 48 + 120 \cdot 480 + 4 \cdot 5760 = 123840, \quad \text{and}$$
$$\mathrm{Var}(L) = E[L^2] - E[L]^2 = 123840 - 276^2 = 47664.$$

(The standard deviation of L is thus $\sqrt{47664} = 218.32$, which is rather large compared to the value of the mean $E[L] = 276$.)

Exercise 20.1 First, note that

$$M_{X_n/n}(t) = E[e^{tX_n/n}] = E[e^{(t/n)X_n}] = M_{X_n}(t/n).$$

Since X_n is a Negative Binomial$(r, \lambda/n)$ we have $M_{X_n}(t) = \left(\frac{\frac{\lambda}{n}e^t}{1-(1-\frac{\lambda}{n})e^t}\right)^r$ when $t < -\log(1-\lambda/n)$ and $M_{X_n}(t) = \infty$ when $t \geqslant -\log(1-\lambda/n)$. Therefore,

$$M_{X_n/n}(t) = M_{X_n}(\frac{t}{n}) = \left(\frac{\frac{\lambda}{n}e^{\frac{t}{n}}}{1-(1-\frac{\lambda}{n})e^{\frac{t}{n}}}\right)^r,$$

when $t/n < -\log(1-\lambda/n)$, i.e. $t < -n\log(1-\lambda/n)$, and it equals ∞ when $t \geqslant -n\log(1-\lambda/n)$.

Letting $h = 1/n$ then using de l'Hôpital's rule we see that

$$\lim_{n\to\infty} n\log(1-\lambda/n) = \lim_{h\to 0}\frac{\log(1-\lambda h)}{h} = -\lambda.$$

So if $t < \lambda$ and n is large then $t < -n\log(1-\lambda/n)$ and $M_{X_n/n}(t)$ has the above formula. Therefore, when $t < \lambda$, letting again $h = 1/n$ and using de l'Hôpital's rule we get

$$\lim_{n\to\infty} M_{X_n/n}(t) = \left(\lim_{h\to 0}\frac{\lambda h e^{ht}}{1-(1-\lambda h)e^{ht}}\right)^r = \left(\frac{\lambda}{\lambda - t}\right)^r.$$

This is the mgf of a Gamma(r, λ).

We have thus shown that the mgf of X_n/n, at any $t < \lambda$, converges that of a Gamma(r, λ). Since the interval $(-\infty, \lambda)$ has 0 strictly inside it, we can apply Theorem 20.3 (e.g. take $\delta = \lambda$) to conclude that X_n/n converges in distribution to a Gamma(r, λ).

Exercise 20.2 For $i = 1, \ldots, n$, we have $M_{X_i}(t) = \exp(\lambda(e^t - 1))$. Hence,

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\cdot\cdots\cdot M_{X_n}(t) = \exp(n\lambda(e^t - 1)).$$

This is the mgf of a Poisson$(n\lambda)$. So that is what $X_1 + \ldots + X_n$ is.

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 the usual trick of setting $h = 1/n$ and applying de l'Hôspital's rule, we get that as $n \to \infty$ this function converges to $e^{t^2/2}$, which is the mgf of a standard normal. The conclusion is therefore that $(X_1 + \cdots + X_n - n\lambda)/\sqrt{n\lambda}$ converges in distribution to a standard normal.

Note how the λ in the numerator is the mean of the X_i's and the λ in the denominator is the variance of the X_i's. Therefore, what we showed in this exercise is the central limit theorem for Poisson random variables! (See the next lecture.)

Exercise 20.3 The solution is similar to the previous exercise. Here, $M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$ when $t < \lambda$. Hence,

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\cdot\cdots\cdot M_{X_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{rn}$$

. This is the moment generating function of a Gamma(nr, λ) and so this is what $X_1 + \cdots + X_n$ is.

and setting $Z_n = \frac{X_1+\cdots+X_n-nr/\lambda}{\sqrt{nr/\lambda^2}}$, we have

$$\begin{aligned} M_{Z_n}(t) &= E[e^{tZ_n}] = e^{-t\sqrt{nr}}E[e^{\frac{\lambda t}{\sqrt{nr}}(X_1+\cdots+X_n)}] \\ &= e^{-t\sqrt{nr}}M_{X_1+\cdots+X_n}(\frac{\lambda t}{\sqrt{nr}}) = e^{-t\sqrt{nr}}\Big(\frac{\lambda}{\lambda - \lambda t/\sqrt{nr}}\Big)^{rn} \\ &= e^{-t\sqrt{nr}}\Big(\frac{1}{1-t/\sqrt{nr}}\Big)^{rn}. \end{aligned}$$

Let $h = 1/\sqrt{n}$, rewrite the above as

$$M_{Z_n}(t) = \exp\Big\{-\frac{t\sqrt{r}}{h} - \frac{r\log(1-ht/\sqrt{r})}{h^2}\Big\} = \exp\Big\{-\frac{ht\sqrt{r}+r\log(1-ht/\sqrt{r})}{h^2}\Big\},$$

take $n \to \infty$ (which means $h \to 0$) and apply de l'Hôspital's rule to get

$$\begin{aligned} \lim_{n\to\infty} M_{Z_n}(t) &= \exp\Big\{-\lim_{h\to 0}\frac{\big[ht\sqrt{r}+r\log(1-ht/\sqrt{r})\big]'}{2h}\Big\} \\ &= \exp\Big\{-\lim_{h\to 0}\frac{t\sqrt{r}-r\frac{t/\sqrt{r}}{1-ht/\sqrt{r}}}{2h}\Big\} \\ &= \exp\Big\{-t\sqrt{r}\lim_{h\to 0}\frac{-ht/\sqrt{r}}{2h(1-ht/\sqrt{r})}\Big\} \\ &= \exp\Big\{t^2\lim_{h\to 0}\frac{1}{2(1-ht/\sqrt{r})}\Big\} = e^{t^2/2}. \end{aligned}$$

The limit is the mgf of a standard normal and so the conclusion is that $(X_1 + \cdots + X_n - nr/\lambda)/\sqrt{nr/\lambda^2}$ converges in distribution to a standard normal.

Note how the r/λ in the numerator is the mean of the X_i's and the r/λ^2 in the denominator is the variance of the X_i's. Therefore, what we showed in this exercise is the central limit theorem for Gamma random variables (which includes as a special case Exponential random variables)! (See the next lecture.)

Exercise 21.1 The probability is approximately $\Phi\big((725 - 144 \times 5)/(0.4\sqrt{144})\big) \approx 0.8512$.

Exercise 21.2 The mean of a Unifrom$(0, 1)$ is $\mu = 1/2$ and the variance is $\sigma^2 = 1/12$.

(a) The probability is approximately $\Phi\big((12 - 20 \times 0.5)/\sqrt{20/12}\big) \approx 0.9393$.

(b) a is such that $\Phi\big((a - 20 \times 0.5)/\sqrt{20/12}\big) = 0.9$ which gives $a = 11.65$.

Exercise 21.3 The mean is

$$\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx = \frac{3}{80^3} \int_0^{80} x^3\, dx = \frac{3 \times 80^4}{4 \times 80^3} = 60.$$

Also

$$E[X_1^2] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx = \frac{3}{80^3} \int_0^{80} x^4\, dx = \frac{3 \times 80^5}{5 \times 80^3} = 3840.$$

So the variance is

$$\sigma^2 = 3840 - 60^2 = 240.$$

Thus, the probability is approximately $1 - \Phi\big((6025 - 100 \times 60)/\sqrt{100 \times 240}\big) \approx 0.4359$.

Exercise 21.4 The probability is approximately

$$\Phi\Big(\frac{8080 - 10,000 \times 0.8}{\sqrt{10,000 \times 0.8 \times 0.2}}\Big) - \Phi\Big(\frac{7940 - 10,000 \times 0.8}{\sqrt{10,000 \times 0.8 \times 0.2}}\Big) \approx 0.910.$$

Exercise 22.1

(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 CDF.

(b) $P(X = 2)$ is the size of the jump at 2, i.e. equals

$$F(2) - F(2-) = (\frac{1}{6} \cdot 2 + \frac{1}{3}) - \frac{1}{3} = \frac{1}{3},$$

$P(X \leqslant 2) = F(2) = 2/3$, and $P(X < 2) = F(2-) = 1/3$.

(c) $P(X > 2) = 1 - P(X \leqslant 2) = 1 - 2/3 = 1/3$ and

$$P(X \geqslant 2) = 1 - P(X < 2) = 1 - 1/3 = 2/3.$$

(d) $P(-1 \leqslant X \leqslant 1/2) = F(1/2) - F((-1)-) = (1/2)^2/3 - 0 = 1/12$, $P(1/3 \leqslant X < 1/2) = F((1/2)-) - F((1/3)-) = 1/12 - 1/27 = 5/108$, and $P(X \in (1/3, 3/2]) = F(3/2) - F(1/3) = 1/3 - (1/3)^2/3 = 8/27$.

(e) $P(4/3 \leqslant X \leqslant 5/3) = F(5/3) - F((4/3) = 1/3 - 1/3 = 0$, then $P(3/2 \leqslant X \leqslant 2) = F(2) - F((3/2)-) = 2/3 - 1/3 = 1/3$, and $P(3/2 < X < 2) = F(2-) - F(3/2) = 1/3 - 1/3 = 0$.

(f) We have $P(2 < X < 3) = F(3-) - F(2) = 5/6 - 2/3 = 1/6$, then $P(2 \leqslant X < 3) = F(3-) - F(2-) = 5/6 - 1/3 = 1/2$, and then $P(3 \leqslant X < 5) = F(5-) - F(3-) = 1 - 5/6 = 1/6$.

(g) $P(3/2 \leqslant X < 3) = F(3-) - F((3/2)-) = 5/6 - 1/3 = 1/2$, then $P(3/2 < X \leqslant 3) = F(3) - F(3/2) = 5/6 - 1/3 = 1/2$, and lastly $P(1/2 < X \leqslant 3) = F(3) - F(1/2) = 5/6 - 1/12 = 3/4$.

(h) $\{X = 2\}$ and $\{1/2 \leqslant X < 3/2\}$ are disjoint events. So the probability of the union is the sum of probabilities and is thus equal to $1/3 + F((3/2)-) - F((1/2)-) = 1/3 + 1/3 - 1/12 = 7/12$.

(e) As 2 is included in $[1/2, 3]$, the probability in question is the same as $P(1/2 \leqslant X \leqslant 3) = F(3) - F((1/2)-) = 5/6 - 1/12 = 3/4$.

Exercise 23.1 Since $y = e^{-x}$ is one-to-one we have

$$F_Y(y) = P(Y \leqslant y) = P(e^{-X} \leqslant y) = P(X \geqslant -\log y) = 1 - F_X(-\log y).$$

Differentiating we get

$$f_Y(y) = f_X(-\log y) \cdot \frac{1}{y} = \begin{cases} \frac{1}{2y} & \text{if } y \in (\frac{1}{e}, e), \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 23.2 Since an exponential random variable is always positive and $y = x^2$ is one-to-one on the interval $(0, \infty)$, we have

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

Differentiating we get

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} = \begin{cases} \frac{\lambda e^{-\lambda\sqrt{y}}}{2\sqrt{y}} & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0. \end{cases}$$

Exercise 23.3 The first part is an exercise in calculus. Indeed, the derivative of the function $h(z) = z^3 + z + 1$ is $3z^2 + 1 > 0$. So the function is strictly increasing and hence its graph cannot cross any horizontal line more than once. Furthermore, $h(z) \to -\infty$ as $z \to -\infty$ and $h(z) \to \infty$ as $z \to \infty$. So the graph of h must in fact cross every horizontal line at least once. Putting the two arguments together we see that the function must cross every horizontal line exactly once. Hence, $z^3 + z + 1 = x$ has a unique solution. We will denote this unique solution by $z = h^{-1}(x)$.

Next, we write

$$F_Z(z) = P(Z \leqslant z) = P(h(Z) \leqslant h(z)) = P(X \leqslant h(z)) = F_X(h(z)) = \Phi(h(z)).$$

Taking derivatives and noticing that $\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ we get

$$f_Z(z) = \Phi'(h(z)) \cdot h'(z) = \frac{1}{\sqrt{2\pi}} (3z^2 + 1) e^{-\frac{(z^3+z+1)^2}{2}}.$$

Exercise 23.4 We have

$$F_Y(y) = P(Y \leqslant y) = P(\log X \leqslant y) = P(X \leqslant e^y) = F_X(e^y).$$

Differentiating we get

$$f_Y(y) = f_X(e^y) \cdot e^y = \lambda e^{y - \lambda e^y}.$$

Exercise 23.5 Since X is always positive we see that $Y = -\lambda^{-1} \log X$ has CDF

$$F_Y(y) = P(Y \leqslant y) = P(-\lambda^{-1} \log X \leqslant y) = P(X \geqslant e^{-\lambda y}) = 1 - F_X(e^{-\lambda y}).$$

Differentiating we get

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

Exercise 23.6 We have

$$F_Y(y) = P\{Y \leqslant y\} = P\{aX + b \leqslant y\} = P\{X \leqslant (y - b)/a\} = F_X\big((y - b)/a\big).$$

Differentiating we get

$$f_Y(y) = \frac{1}{a} \cdot f_X\big((y - b)/a\big) = \frac{1}{a} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[(y-b)/a - \mu]^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi(a\sigma)^2}} e^{-\frac{[y - (a\mu + b)]^2}{2(a\sigma)^2}}.$$

So $Y \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

The real information here is that a linear transformation of a normal is yet another normal. The fact that the mean of Y is $a\mu + b$ and the variance is $a^2\sigma^2$ should not come as a surprise. (Why not?)

Exercise 23.7 Define $g : [1,\infty) \to [1,\infty)$ by

$$g(x) = \begin{cases} 2x & \text{if } x \geqslant 2, \\ x^2 & \text{if } 1 \leqslant x < 2, \end{cases}$$

Since this is a strictly increasing function, it is one-to-one. Then

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

We know that Y is never less than 1 and so $F_Y(y) = 0$ for $y < 1$. Next, we have two cases. If $1 \leqslant y < 4$ then then $g(X) \leqslant y$ means $X^2 \leqslant y$ and so $-\sqrt{y} \leqslant X \leqslant \sqrt{y}$. Since X is never negative we see that this is equivalent to saying that simply $X \leqslant \sqrt{y}$. Therefore,

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

Differentiating, we get

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} = \frac{1}{y} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2y^{3/2}}.$$

On the other hand, if $y > 4$ then $g(X) \leqslant y$ means either $X < 2$ or $X \geqslant 2$ and $2X \leqslant y$. Therefore,

$$F_Y(y) = P\{X < 2\} + P\{2 \leqslant X \leqslant y/2\} = P\{X < 2\} + P\{X \leqslant y/2\} - P\{X < 2\} = P\{X \leqslant y/2\}.$$

Differentiating, we get

$$f_Y(y) = \frac{1}{2} \cdot f_X(y/2) = \frac{1}{2} \cdot \frac{4}{y^2} = \frac{2}{y^2}.$$

To summarize, we have

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

Exercise 23.8 Let $g : [0, \frac{\pi}{2}] \to [0, \frac{v_0^2}{g}]$ be defined by $g(\theta) = \frac{v_0^2}{g}\sin(2\theta)$. The function g is not one-to-one since the equation $g(\theta) = r$ has two solutions: $\theta = \frac{1}{2}\arcsin(\frac{gr}{v_0^2})$ and $\theta = \frac{\pi}{2} - \frac{1}{2}\arcsin(\frac{gr}{v_0^2})$. Hence, for $0 \leqslant r \leqslant \frac{v_0^2}{g}$,

$$\begin{aligned} F_R(r) &= P(R \leqslant r) = P\left(\left\{0 \leqslant \theta \leqslant \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right\} \text{ or } \left\{\frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right) \leqslant \theta \leqslant \frac{\pi}{2}\right\}\right) \\ &= \frac{2}{\pi}\left(\frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right) + \frac{2}{\pi}\left(\frac{\pi}{2} - \left(\frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{gr}{v_0^2}\right)\right)\right) \\ &= \frac{2}{\pi}\arcsin\left(\frac{gr}{v_0^2}\right). \end{aligned}$$

Differentiating, we get that for $0 \leqslant r \leqslant \frac{v_0^2}{g}$,

$$f_R(r) = \frac{2g}{\pi v_0^2} \cdot \frac{1}{\sqrt{1 - \frac{g^2r^2}{v_0^4}}} = \frac{2g}{\pi}\frac{1}{\sqrt{v_0^4 - g^2r^2}}.$$

Also, since the sine of an angle between 0 nad $\pi/2$ is always between 0 and 1 we see that R has to be between 0 and $\frac{v_0^2}{g}$ and so $f_R(r)$ is 0 outside this interval.

Exercise 24.1 First we need the conditional mass function of X given $X \geqslant 2$. This is given by

$$f_{X\,|\,X\geqslant 2}(k) = \frac{P(X=k, X\geqslant 2)}{P(X\geqslant 2)} = \frac{P(X=k)}{1-P(X=0)-P(X=1)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}}{1-(1-p)^n-np(1-p)^{n-1}}$$

for k between 2 and n, and 0 otherwise. Then the expected value we are after equals

$$E[X\,|\,X\geqslant 2] = \sum_{k=2}^{n} k f_{X\,|\,X\geqslant 2}(k) = \frac{1}{1-(1-p)^n-np(1-p)^{n-1}} \sum_{k=2}^{n} k\cdot\binom{n}{k}p^k(1-p)^{n-k}.$$

To compute the sum we recall that the expected value of a binomial equals np. So if the sum were from 0 to n then we would have had np for the sum. The $k=0$ term is 0 anyway and so the sum is simply equal to np minus the $k=1$ term, i.e. we get

$$E[X\,|\,X\geqslant 2] = \sum_{k=2}^{n} k f_{X\,|\,X\geqslant 2}(k) = \frac{np-np(1-p)^{n-1}}{1-(1-p)^n-np(1-p)^{n-1}}.$$

Exercise 24.2 First, let us write down the joint pdf of X and Y:

$$f_{X,Y}(x,y) = f_X(x)\cdot f_{Y|X}(y\,|\,x) = \frac{1}{x^2}\cdot\frac{1}{x} = \frac{1}{x^3}, \quad \text{when } x\geqslant 1 \text{ and } 0\leqslant y\leqslant x,$$

and 0 otherwise. Now we can integrate the x away to get the pdf of Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{\substack{x\geqslant 1\\ x\geqslant y}} \frac{1}{x^3}\,dx.$$

Saying that $x\geqslant 1$ and $x\geqslant y$ is the same as saying that x is larger than the maximum of y and 1. So

$$f_Y(y) = \int_{\max(1,y)}^{\infty} \frac{1}{x^3}\,dx = \frac{1}{2(\max(1,y))^2}.$$

(If you are not comfortable with using the notation $\max(1,y)$, then simply do the integral in both cases: $0\leqslant y\leqslant 1$, where the integral is from 1 to infinity, and $y\geqslant 1$, where the integral is from y to infinity. And then you get that $f_Y(y) = 1/2$ if $0\leqslant y\leqslant 1$ and equals $1/(2y^2)$ if $y\geqslant 1$.)

Exercise 24.3 To find k we write

$$1 = \iint f(x,y)\,dx\,dy = k\int_{-1}^{1}\left(\int_{y}^{1}|x|\,dx\right)dy.$$

Instead of using calculus to compute the inside integral we will use geometry. The integral is the area under the graph of $|x|$ between y and 1. If $-1<y<0$ this is the sum of the area of two 45-90-45 triangles. One with base (and height) $-y$ and the other with base and hight 1. (Draw the picture.) Therefore, the inside integral equals $\frac{1}{2}(1+y^2)$. On the other hand, when $0<y<1$ the area becomes the difference of the areas of the two triangles. So the integral equals $\frac{1}{2}(1-y^2)$. (Do the integral with calculus and double check these answers.) Now we have

$$1 = \frac{k}{2}\int_{-1}^{0}(1+y^2)\,dy + \frac{k}{2}\int_{0}^{1}(1-y^2)\,dy = \frac{k}{2}\int_{0}^{1}(1+y^2)\,dy + \frac{k}{2}\int_{0}^{1}(1-y^2)\,dy = k\int_{0}^{1}dy = k.$$

(In the second equality we used symmetry: change variables $z=-y$ and then rename the new variable z back to y, since it is just a dummy variable. In the third equality we just combined the two integrals bu saying that the sum of the integrals is the integral of the sum. This way we ended up not even needing to integrate y^2.)

The marginals are given by

$$f_X(x) = \int_{-1}^{x} |x| \, dy = |x| \, (x+1), \quad \text{for } -1 < x < 1$$

and 0 otherwise and

$$f_Y(y) = \int_{y}^{1} |x| \, dx = \begin{cases} \frac{1}{2}(1+y^2) & \text{if } -1 < y < 0, \\ \\ \frac{1}{2}(1-y^2) & \text{if } 0 \leqslant y < 1. \end{cases}$$

The conditional pdfs are given by

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{2|x|}{1+y^2} & \text{if } -1 < y < 0, \, y \leqslant x \leqslant 1 \\ \\ \frac{2|x|}{1-y^2} & \text{if } 0 \leqslant y \leqslant x < 1, \end{cases}$$

and 0 otherwise and

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{x+1} \quad \text{if } -1 < y \leqslant x < 1,$$

and 0 otherwise.

Exercise 24.4 a) We have already computed the marginals for this pdf in Example 16.4. Applying the definitions we get

$$f_{X|Y}(x \,|\, y) = \frac{2x}{1-y^2} \quad \text{and} \quad f_{Y|X}(y \,|\, x) = \frac{2y}{x^2}$$

for $0 < y < x < 1$. Therefore,

$$E[X \,|\, Y = y] = \int_{y}^{1} x \cdot \frac{2x}{1-y^2} \, dx = \frac{2}{3} \cdot \frac{1-y^3}{1-y^2}$$

for $0 < y < 1$ and

$$E[Y \,|\, X = x] = \int_{0}^{x} y \cdot \frac{2y}{x^2} \, dy = \frac{2x}{3}$$

for $0 < x < 1$.

b) Since we know the conditional pdf of Y given $X = x$ we have

$$E[Y^2 \,|\, X = x] = \int_{0}^{x} y^2 \cdot \frac{2y}{x^2} \, dy = \frac{x^2}{2},$$

c) There are a couple of ways to compute this. For example, we start by computing the pdf of Y given $X \leqslant 1/2$:

$$f_{Y|X\leqslant 1/2}(y) = \frac{\int_{-\infty}^{1/2} f_{X,Y}(x,y) \, dx}{\int_{-\infty}^{1/2} f_X(x) \, dx} = \frac{\int_{y}^{1/2} 8xy \, dx}{\int_{0}^{1/2} 4x^3 \, dx} = 16y(1-4y^2)$$

for $0 < y < 1/2$, and 0 otherwise. (Can check at this point that this indeed integrates to 1.) Now that we have the conditional pdf we compute

$$E[Y \,|\, X \leqslant 1/2] = \int_{-\infty}^{\infty} y \, f_{Y|X\leqslant 1/2}(y) \, dy = \int_{0}^{1/2} y \cdot 16y(1-4y^2) \, dy = \frac{16}{3} \cdot (1/2)^3 - \frac{64}{5}(1/2)^5 = \frac{4}{15}.$$

Exercise 24.5 Given that $X = x$ the Y should be uniformly distributed over its possible range of values. But the range of values of the Y depend on what x is. Drawing a picture of the parallelogram we see the following:

If $0 \leqslant x \leqslant 1$ then the Y can range from 0 to x and since it is uniformly distributed on this interval its mean is $E[Y \mid X = x] = x/2$. If $1 \leqslant x \leqslant 2$ the Y ranges from 0 to 1 and so its mean is $E[Y \mid X = x] = 1/2$. And if $2 \leqslant x \leqslant 3$ the Y can range from $x - 2$ to 1 and so its mean is $E[Y \mid X = x] = (x - 1)/2$.

Exercise 24.6 Let us write $Y = T - s$. We will first compute the CDF of Y, given that $T > s$. For $t \geqslant 0$ this is given by

$$\begin{aligned} P(Y \leqslant t \mid T > s) &= P(T - s \leqslant t \mid T > s) = \frac{P(s < T \leqslant t + s)}{P(T > s)} \\ &= \frac{\int_s^{t+s} \lambda e^{-\lambda x}\, dx}{\int_s^{\infty} \lambda e^{-\lambda x}\, dx} \\ &= \frac{e^{-\lambda s} - e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= 1 - e^{-\lambda t}. \end{aligned}$$

If $t < 0$ then the above is 0 because given that $T > s$ we have that $Y = T - s$ is positive and the probability it is less than a negative number is 0.

Now, differentiating in t we get that

$$f_{Y \mid T > s}(t) = \lambda e^{-\lambda t}, \quad \text{for } t \geqslant 0.$$

So $Y = T - s$ is an Exponential(λ) random variable, just like T. This is in fact the memoryless property of the exponential distribution that we have seen way back.

Exercise 24.7 If we are given that $X = x$ then $Z = x^2 + Y^2$ and we just need to find

$$E[Z \mid X = x] = x^2 + E[Y^2 \mid X = x].$$

But since Y is independent of X the last expected value simply equals $E[Y^2]$ which we can compute as

$$E[Y^2] = \int_{-1}^{0} y^2 \cdot \frac{1}{2}\, dy + \int_0^{\infty} y^2 \cdot \frac{1}{2} e^{-y}\, dy = \frac{1}{6} + \frac{1}{2} \cdot 2 = \frac{7}{6}.$$

So we find that

$$E[Z \mid X = x] = x^2 + \frac{7}{6}.$$

Appendix B

Tables

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

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 B.1. ©1991 Introduction to Probability and Mathematical Statistics, 2nd Edition, by Bain & Engelhardt

Discrete random variables:

Name	pmf	mean	variance	mgf
Bin(n,p)	$\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\dots,n$	np	$np(1-p)$	$(1-p+pe^t)^n$
Neg Bin(r,p)	$\binom{k-1}{r-1}(1-p)^{k-r}p^r$ $k=r,r+1,\dots$	$\frac{r}{p}$	$\frac{(1-p)r}{p^2}$	$\begin{cases}\left(\frac{pe^t}{1-(1-p)e^t}\right)^r & \text{if } t<-\log(1-p),\\ \infty & \text{otherwise.}\end{cases}$
Poi(λ)	$e^{-\lambda}\frac{\lambda^k}{k!}$ $k=0,1,2,\dots$	λ	λ	$e^{\lambda(e^t-1)}$

Ber(p)=Bin$(1,p)$ and Geo(p)=Neg Bin$(1,p)$.

Continuous random variables:

Name	pdf	mean	variance	mgf
Uniform(a,b)	$\frac{1}{b-a}$ $a\leqslant x\leqslant b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Exponential(λ)	$\lambda e^{-\lambda x}$ $x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\begin{cases}\frac{\lambda}{\lambda-t} & \text{if } t<\lambda,\\ \infty & \text{otherwise.}\end{cases}$
Gamma(α,λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$ $x>0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\begin{cases}\left(\frac{\lambda}{\lambda-t}\right)^\alpha & \text{if } t<\lambda,\\ \infty & \text{otherwise.}\end{cases}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$e^{\mu t+\sigma^2t^2/2}$

$\chi^2(n)$=Gamma$(n/2,1/2)$.