MATH 302 "INTRODUCTION TO PROBABILITY" LECTURE NOTES

Preamble

These are typed-up lecture notes. They are supposed to accompany the lectures and the book *A first course in probability, 9th ed.* by Sheldon Ross, they are not supposed to supplant them! Sheldon Ross has had editors, teachers, and thousands of students look over his text for more than a decade, so it is pretty well polished. You, on the other hand, are one of the first persons to see these notes. It is entirely possible that these notes contain the occasional mistake. In general, if a calculation looks off to you, it is much more likely that I made a mistake than that logic has failed to do it's magic, so don't panic. Please contact me[1] (at the lectures or via email) if you do find a mistake, a fishy argument, or even a typo, so I can correct it! You help everybody if you do.

1. What is probability theory good for?

Probability theory is used in practically every branch of science, engineering, medicine, economics, finance, and the social sciences. Probability lies at the foundation of statistics. Probability theory is also really useful if you like to gamble.

In a nutshell, probability theory is *the science of incomplete knowledge.* Say you have a six-sided die. What is the outcome of the roll going to be? Well, if you know the initial position, the velocity at which it is thrown, the position of the table, the constants of friction of the table and the air, etc. you could use Newton's laws and a good computer to predict the throw *exactly.* But when we are playing a game of Monopoly, we are not going to have all this data, nor the computational power. Probability theory is aimed at making the best predictions given incomplete data or limited resources.

Example 1. Some real or conceptual experiments or observations where probability theory is useful for predictions or explanations:

(a) Rolling 2 dice.
(b) Dealing a poker hand.
(c) Turning a roulette wheel.
(d) The half-life of a radio-active isotope.
(e) The position of a molecule in a solution.
(f) The sex of an offspring.
(g) The number of mutations in a strand of DNA.
(h) The number of busy lines of a call-centre.

[1] You can find my contact information (and other course info) in the course outline on my website: `www.math.ubc.ca/~thulshof/teaching`.

(i) The number of cars sold in Vancouver in 2014.
(j) A person’s preference for Pepsi or Coke.

△

Example 2. Roll 2 fair dice. What is the probability that their total is 8? We can start by looking at all possible outcomes, and determine how frequent 8 is:

Die #1 ∖ Die #2	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

From the table we can see that there are $6 \times 6 = 36$ different outcomes for two rolled dice. Of those 36 outcomes, 5 result in a total of 8. With a fair dice we assume that all outcomes of a single roll are equally likely (i.e., all outcomes have probability $1/6$), so the probability that 2 dice rolls add up to 8 is $5/36$.[2]

To get acquainted with the notation of probability theory, let’s write that down as a formula:

$$\mathbb{P}(\text{ sum of two dice rolls} = 8) = \frac{5}{36}.$$

In general we will write $\mathbb{P}(\text{event})$ to denote the probability of that event. △

2. Counting

The example with the two dice really just boils down to counting how many outcomes of a given type there are. As we will see in the coming semester, a huge part of probability theory deals with counting. Sometimes counting is as simple as writing down a little table, but other times it can get pretty complex. There are counting problems that are so complex that a whole branch of mathematics has been devoted to them. The mathematical theory of counting is known as *combinatorics*.

Example 3. Some of the problems that are studied with combinatorics:

- The number of 5 note melodies you can play on a piano.
- The number of 10 digit phone numbers that contain the sequence “555”.
- The number of ways in which we can colour the countries on a map so that neighbouring countries have different colours, using four colours.

[2] In probability theory, we will most times express probabilities as numbers between 0 and 1, rather than as percentages. The reason for this is that it is more convenient. Doing calculations with percentages is cumbersome because we will sometimes need to keep track of many factors of 100. Anyways, if we calculate that the probability is x, where $x \in [0,1]$, then we can always express this in percentages by multiplying x with 100%. The probability of rolling a total of 8 with 2 dice is thus $5/36 \times 100\% = 13.888\ldots\%$.

- The number of possible outcomes of the Kentucky Derby.

△

2.1. **The multiplication rule.** Before we start studying how to calculate probabilities, we will first have to study some of the basics of combinatorics. In Example 2 we already saw one of the most important ideas of combinatorics in action:

Principle 1 [The multiplication rule]. *If we do two experiments and experiment #1 has n possible outcomes, while experiment #2 has m possible outcomes, then both experiments together have $n \times m$ possible outcomes.*

We can of course generalise this principle to more than two experiments.[3] In general, when we do r experiments and experiment $\#i$ has n_i possible outcomes, than the total number of outcomes of all r experiments equals $n_1 \times n_2 \times \cdots \times n_r$.

Question 1. *How many possible postal codes are there?*

Solution. Postal codes have the following form:

$$\text{letter number letter} \quad \text{number letter number}$$

so by the multiplication rule we get that the total number of possible postal codes is

$$26 \times 10 \times 26 \quad \times \quad 10 \times 26 \times 10 = 26^3 \times 10^3 = 17576000.$$

◊

Question 2. *How many 5 note melodies can you play on a piano? What if the first three notes have to be played using the black keys?*

Solution. A piano has 85 keys, so by the multiplication rule, we can play $85 \times 85 \times 85 \times 85 \times 85 = 85^5 = 4437053125$ different "melodies". If we restrict ourselves to the black keys, of which there are 36, for the first three notes, then we can play $36 \times 36 \times 36 \times 85 \times 85 = 36^3 \times 85^2 = 337089600$ different "melodies". ◊

2.2. **Permutations.** Another counting problem comes from the following question.

Question 3. *Suppose there are 5 horses entering in a race, call them Day Star, Apollo, Elwood, Sea Biscuit, and Baden-Baden. How many different orders are there for the horses to cross the finish line?*

Solution. We are going to consider each place of the finishing order in turn:

[3] We are going to use the word *experiment* often, and not necessarily in the sense that you may be familiar with. In probability theory, an experiment can be any action, event or procedure, past, present or future, that can yield or could have yielded more than one outcome. So observing the half-life of an atomic nucleus is an experiment, but so is rolling a die, or the number of deaths that resulted from shark attacks between 1879 – 1889. Even making a sandwich with bacon, lettuce, and tomatoes is an experiment if you want it to be…

- Which horse comes in 1st? There are 5 horses, so there are 5 possibilities. Suppose that Apollo finished 1st.
- Given that Apollo finished 1st, who finished 2nd? All horses except Apollo could finish 2nd, so there are 4 possibilities. Suppose Baden-Baden finished 2nd.
- Given that Apollo came in 1st and Baden-Baden 2nd, who finished 3rd? There are still 3 horses unaccounted for, so there are 3 possibilities. Suppose Sea Biscuit came in 3rd.
- Given that Apollo, Baden-Baden, and Sea Biscuit are 1st, 2nd, and 3rd, who finished 4th? There are 2 possibilities. Suppose Day Star came in 4th.
- Elwood is the only horse left, so it has to come in 5th: there is only 1 possibility.

As a result, we can apply the multiplication rule:

$$\#\text{ of different orders} = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

◊

The mathematical term for a reordering of a list is a *permutation*. In general we have that the number of permutations of a list of *distinguishable* elements can be counted according to the following principle:

PRINCIPLE 2 [Counting permutations]. *Given a list with n distinguishable element there are*

$$n! := n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

different permutations of that list.

We write $n!$ for the number of permutations, we say "*n factorial.*" It turns out that it is convenient to define[4] $0! := 1$.

Distinguishability is important when counting permutations: horse races wouldn't be too much fun if we couldn't tell the horses apart at the finish line. Distinguishability is not always a given though. In chemistry and particle physics, for instance, we may want to count the different configurations of electrons around a nucleus, but there is no way to tell these electrons apart. This will affect the number of different states, so we will need to take this into account, otherwise our calculations can be off by a huge margin.[5]

Clearly, if we have a list of n elements that are totally indistinguishable, then any permutation is the same, so there is only one permutation (e.g. the list $\{1,1,1,1,1,1,1,1\}$ has only one permutation). It can also happen that the elements fall into different categories, where the elements in a given category are indistinguishable from each other, but distinguishable from elements in other categories.

[4] Whenever we make a definition, we are going to write := to denote that the equality is true because we say so, whereas we will use = only when the equality holds because of logic.

[5] You may have noticed that the numbers we are dealing with when we do these combinatorial calculations have a tendency towards the gigantic. When doing calculations it is wise to keep on writing thinks like 63! or 31^{12} until the very end of the calculation to avoid copying errors. Sometimes numbers will be so big that your calculator can't handle them anymore, like 545!. Don't bother approximating the number if this is the case, just leave it as is in your final answer.

QUESTION 4 [The labelling trick]. *How many different "words" can we spell with the letters of MISSISSIPPI?.*

Solution. If all letters were distinguishable, i.e., if they were for instance labeled as follows:

$$M_1 I_1 S_1 S_2 I_2 S_3 S_4 I_3 P_1 P_2 I_4,$$

then there would be $11! = 39916800$ different orderings. Let's imagine we put all these labeled words in a big list.

If we look only at the ways the labeled S's may appear in a given word (ignoring the positions of the other letters for the moment) then it is clear that there are $4! = 24$ different orderings. So if, in the list of all labeled "words" we remove the labels on the S's, then we will see 24 identical copies of each word in the list. Similarly, there are $4! = 24$ ways in which the labeled I's may appear, so if we also remove the labels on the I's, we will now see $4! \times 4!$ identical copies of each word in the list. There are $2! = 2$ ways in which the labeled P's may appear, so removing the labels on the P's, we will see $4! \times 4! \times 2!$ identical copies of each word in the list. There is only one M, so removing its label has no effect on the multiplicity of words.

As a result, our list contains

$$\frac{\#\text{ labeled words}}{\#\text{ identical copies of each word}} = \frac{11!}{4! \times 4! \times 2!} = 34650$$

distinct words spelled with the letters MISSISSIPPI. ◊

We can generalise the above solution:

PRINCIPLE 3 [Counting permutations with categories]. *Given a list with n elements that fall into r categories of indistinguishable elements with sizes $n_1, n_2, \ldots, n_r$, there are*

$$\frac{n!}{n_1!\, n_2! \cdots n_r!}$$

different orderings for that list.

2.3. Combinations.

The third main counting problem that we will encounter in this course is that of counting combinations. Consider the following question:

QUESTION 5. *A vase contains 5 marbles: 1 red, 1 green, 1 blue, 1 yellow, and 1 pink. If we draw 3 marbles from the vase, how many different colour combinations can we get?*

Solution. Let's first keep track of the order in which we draw the marbles. For the first marble there are 5 possibilities, for the second there are 4 possibilities, and for the third there are 3 possibilities, so there are $5 \times 4 \times 3 = 60$ different ordered draws possible. But we are not interested in the order. As far as we are concerned, the draw "red – blue – pink" is equivalent to the draw "pink – red – blue", so we overcounted the number of combinations when we kept track of the order. How much have we overcounted? For each set of 3 marbles, there are $3! = 6$ possible orders, so we counted each combination 6 times when we kept track of the order. Therefore,

$$\begin{aligned}\#\text{ combinations of 3 out of 5 colours} &= \frac{\#\text{ ordered combinations of 3 out of 5 colours}}{\#\text{ orders of 3 colours}}\\ &= \frac{5\times4\times3}{3\times2\times1} = 10.\end{aligned}$$

◊

Again we can generalise the above example:

Principle 4 [Counting combinations]. *For $r \le n$, we can make a combination of r elements out of a list of n indistinguishable elements in*

$$\binom{n}{r} := \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!}$$

different ways.

We write $\binom{n}{r}$, we say "n choose r". For the second equality in the above equation we used that

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

Observe that $\binom{n}{r} = \binom{n}{n-r}$. Since we defined $0! := 1$ this implies that $\binom{n}{0} = \binom{n}{n} = 1$.

Question 6. *How many different poker hands are there?*

Solution. There are 52 cards in a deck, and a poker hand is 5 cards, so $n = 52$ and $r = 5$. As a result,

$$\#\text{ poker hands} = \binom{52}{5} = \frac{52\times51\times50\times49\times48}{5!} = 2598960.$$

◊

Question 7. *Why does four-of-a-kind beat a full house?*

Solution. Assume a fair deal, so each hand is equally likely.

We start by counting the number of combinations of five cards that are a four-of-a-kind. A deck of cards contains 13 quartets. We need to draw one of those, so there are $\binom{13}{1}$ combinations. Having drawn a quartet, there are 48 cards left in the deck. We need to draw one of those. There are $\binom{48}{1}$ ways of doing that. Therefore,

$$\#\text{ four-of-a-kind hands} = \binom{13}{1}\times\binom{48}{1} = 624.$$

Moreover, the probability of getting four-of-a-kind is

$$\mathbb{P}(\text{hand is four-of-a-kind}) = \frac{\#\text{ four-of-a-kind hands}}{\#\text{ poker hands}} = \frac{624}{2598960} = 0.00024.$$

Now we count the number of hands that are a full house. Again, a deck contains 13 quartets. From one of those quartets we want to draw three cards, from a second quartet we want to draw two cards. There are $\binom{13}{1}$ ways of choosing the quartet from which we draw three cards. Then, there are 12 quartets left, so there are $\binom{12}{1}$ ways of choosing the quartet from which to draw two cards. There are $\binom{4}{3}$ ways of drawing 3 cards from 4, and there are $\binom{4}{2}$ ways of drawing 2 cards from 4. Therefore,

$$\#\text{ full house hands} = \binom{13}{1} \times \binom{4}{3} \times \binom{12}{1} \times \binom{4}{2} = 3744.$$

So there are 6 times as many hands that are a full house than there are four-of-a-kind hands. Since all hands are equally likely, drawing a full house is 6 times more likely than drawing a four-of-a-kind. Indeed, the probability of getting a full house is

$$\mathbb{P}(\text{hand is full house}) = \frac{\#\text{ full house hands}}{\#\text{ poker hands}} = \frac{3744}{2598960} = 0.00114.$$

◊

2.3.1. *The binomial theorem.* The number of combinations $\binom{n}{r}$ are also known as the *binomial coefficients*. The binomial coefficients come about naturally when we expand equations of the form $(x+y)^n$.

Example 4. We expand $(x+y)^2$:

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

Well, that is not a very convincing example to show the relation between $(x+y)^n$ and $\binom{n}{r}$, so lets try a larger value of n. We are going to expand $(x+y)^4$:

$$\begin{aligned}(x+y)^4 &= (x+y)^2(x+y)^2\\ &= (x^2+2xy+y^2)(x^2+2xy+y^2)\\ &= x^2(x^2+2xy+y^2)+2xy(x^2+2xy+y^2)+y^2(x^2+2xy+y^2)\\ &= x^4+2x^3y+x^2y^2+2x^3y+4x^2y^2+2xy^3+x^2y^2+2xy^3+y^4\\ &= 1x^4+4x^3y+6x^2y^2+4xy^3+1y^4\\ &= \binom{4}{4}x^4y^{4-4}+\binom{4}{3}x^3y^{4-1}+\binom{4}{2}x^2y^{4-2}+\binom{4}{1}x^1y^{4-1}+\binom{4}{0}x^0y^{4-0}\\ &= \sum_{k=0}^{4}\binom{4}{k}x^ky^{4-k}.\end{aligned}$$

△

In the last step we introduced the capital-sigma notation to simplify the formula.[6]

It turns out that we can generalise the above example to hold for any integer power of $(x+y)$:

Theorem 1 [The binomial theorem].

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

Proof. We prove[7] this theorem using the ‘labelling trick’. We are going to label all the x’s and y’s:

$$(x_1+y_1)(x_2+y_2)\cdots(x_n+y_n). \tag{1}$$

We expand the above formula. By the multiplication rule we get 2^n distinguishable terms. Each of these terms will contain either x_i or y_i for any $i = 1,2,\ldots,n$, but never both. For instance, we get the term $y_1x_2y_3y_4y_5x_6y_7y_8\cdots x_{n-2}y_{n-1}y_n$. Since there are $\binom{n}{r}$ ways of choosing r labeled x’s from n labeled x’s, there will be $\binom{n}{r}$ terms with r x’s and $(n-r)$ y’s in the expansion of (1). So if we erase all the labels we get

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

as claimed. □

[6] Recall that the capital-sigma notation works like this: $\sum_{i=a}^{b} x_i = x_a + x_{a+1} + \cdots + x_{b-1} + x_b$. Basic example:

$$\sum_{s=1}^{4} s^2 = 1^2+2^2+3^2+4^2 = 1+4+9+15 = 29.$$

[7] Ross presents another proof that uses induction. You may find it helpful to study both proofs.

Question 8. *How many subsets can we make from a set of n elements?*

Solution. We will use the binomial theorem with $x = 1$ and $y = 1$. Since there are $\binom{n}{k}$ subsets of size k we can make a total of

$$\sum_{k=0}^{n}\binom{n}{k} = \sum_{k=0}^{n}\binom{n}{k}1^k 1^{n-k} = (1+1)^n = 2^n$$

subsets.[8] ◊

2.3.2. *The number of ways of dividing things.* The next combinatorial subject we will discuss is that of counting possible divisions.

Question 9 [The wall trick]. *Suppose we have 11 identical marbles, and we want to put them into 4 different boxes. How many ways are there of doing that?*

Solution. We are going to solve this using the so-called *wall trick*. Instead of putting the marbles in the boxes, we are going to put the boxes *around* the marbles. We start by putting all our marbles in a row:

❂ ❂ ❂ ❂ ❂ ❂ ❂ ❂ ❂ ❂ ❂

Now we put our four boxes around them:

|❂ ❂||❂ ❂ ❂||❂||❂ ❂ ❂ ❂ ❂|

We can encode this sequence by writing M (marble) for ❂, and W (wall) for $||$:

$$\{M, M, W, M, M, M, W, M, W, M, M, M, M, M\}.$$

We did not encode the walls at the beginning and end of the sequence, since they are always in the same place. We can shorten the encoding even more by only keeping track of the positions of the walls with respect to the set of walls and marbles, i.e., $\{3, 7, 9\} \subset \{1, 2, \ldots, 14\}$. A possible division then, is characterised by the positions of 3 walls among $11+4-1$ possible locations.

So it turns out that if we want to count the number of possible divisions of 11 marbles among 4 boxes, this is equivalent to the number of ways that we can choose a set of 3 among $11 + 4 - 1$ elements, i.e., there are $\binom{14}{3}$ possible divisions. ◊

Once again, we can generalise this principle:

Principle 5 [Counting possible divisions].

$$\#\ \textit{ways of dividing n indistinguishable elements into r subsets} = \binom{n+r-1}{r-1}$$

[8] Note that we also counted the subset of size 0, i.e., we counted the *empty set* $\varnothing$, corresponding to $k = 0$. Since $\binom{n}{0} = 1$, there are $2^n - 1$ nonempty subsets.

2.4. **Urn problems.**

[9] With our basic knowledge of combinatorics we can already do some useful calculations. The problems that we are going to discuss are generally known as *urn problems* because we are going to imagine our problem as a problem where we have to pick coloured marbles from an urn (vase).[10]

Let's start with an "actual" urn problem:

QUESTION 10. *An urn contains 15 marbles: 8 are red and 7 are green. We pick 5 marbles from the urn at random. What is the probability that we pick 3 red marbles and 2 green marbles?*

Solution. We start by observing

$$\mathbb{P}(\text{ drawing 3 red and 2 green}) = \frac{\#\text{ draws of 3 red and 2 green}}{\#\text{ draws of 5}} = \frac{A}{B}.$$

We apply the combination rule for the denominator:

$$B = \binom{15}{5} = 3003.$$

To count the numerator, we first apply the multiplication rule:

$$A = \#\text{ draws of 3 from 8 red} \times \#\text{ draws of 2 from 7 green} = \binom{8}{3}\binom{7}{2} = 1176.$$

Therefore,

$$\mathbb{P}(\text{ drawing 3 red and 2 green}) = \frac{1176}{3003} = 0.3916.$$

◊

Now on to a less obvious urn problem:

QUESTION 11. *An election in a small town is between two candidates with the following outcome:*

- *Candidate A gets 1422 votes.*
- *Candidate B gets 1405 votes.*

Candidate A is about to be declared the winner, but then it is discovered that 101 votes were accidentally counted twice. There is no reason to believe that the miscount was intentional, so each twice counted vote could have been for either candidate. What is the probability that after a recount it turns out that Candidate B is declared the winner?

[9] This section does not correspond to a section in Ross. This section is more about solving a certain type of problem, and not so much about theory.

[10] This kind of reductive thinking involving marbles and urns has a long history going back to Jacob Bernoulli, who wrote about it first in 1713. In his writing he used the latin word 'urna'. Urna means 'clay vessel', especially of the type used to collect votes or ballots. Nowadays the word has different meanings, but we are going to ignore this fact.

Solution. We suppose each of the $1422 + 1405 = 2827$ votes is equally likely to have been counted twice. We write

$$X = \#\text{ double votes for A.}$$

We want to calculate the probability of the following event:

$$\begin{aligned} E &= \text{event that B is the winner} \\ &= \{1422 - X < 1405 - (101 - X)\} \\ &= \{1422 - 1405 + 101 < 2X\} \\ &= \{118 < 2X\} \\ &= \{X > 59\}. \end{aligned}$$

That is, we want to calculate

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(X > 59) \\ &= \mathbb{P}(X = 60) + \mathbb{P}(k = 61) + \cdots + \mathbb{P}(X = 101) \\ &= \sum_{k=60}^{101} \mathbb{P}(X = k). \end{aligned}$$

So now we have to calculate the probability $\mathbb{P}(X = k)$. We can imagine the votes for A as being 1422 red marbles, and the votes for B as 1405 green marbles. Then we have the following equivalence:

$$\mathbb{P}(X = k) = \mathbb{P}(\text{ drawing } k \text{ red and } (101 - k) \text{ green}) = \frac{\binom{1422}{k}\binom{1405}{101-k}}{\binom{2827}{101}}.$$

Now we calculate (with a computer)

$$\mathbb{P}(E) = \sum_{k=60}^{101} \frac{\binom{1422}{k}\binom{1405}{101-k}}{\binom{2827}{101}} = 0.059,$$

so there is about 6% chance that candidate B wins the election. ◊

Question 12. *An ecologist captures 60 water beetles in a pond, marks each with a dot of paint and returns them to the pond. A day later she captures 50 beetles and finds that 12 are marked and 38 are unmarked.*

What is the best estimate for N, the number of beetles in the pond?

Solution. (A) Assume that

$$\text{ratio in 2nd sample} = \text{ratio in pond.}$$

Then

$$\frac{12}{50} = \frac{60}{N}, \qquad \Rightarrow \qquad N = \frac{60 \times 50}{12} = 250.$$

(B) The assumption made in (A) is not needed. We can calculate the probability p_N,

$$p_N = \mathbb{P}(\text{ catch 12 marked and 38 unmarked beetles from } N)$$

and then see which value of N gives the highest probability for that to happen. This is known as a *maximum likelihood estimate.*

We can calculate p_N using an urn model: consider an urn with 60 red marbles and $N-60$ green marbles, then

$$p_N = \mathbb{P}(\text{draw 12 red and 38 green from } N) = \frac{\binom{60}{12}\binom{N-60}{38}}{\binom{N}{50}}.$$

Now we can check with a computer that p_N is maximised when $N = 250$. ◊

Question 13 [The pigeonhole principle]. *Suppose n people attend a party. What is the probability that at least two people shake the exact same number of hands during that party? (Assume that no one shakes the same hand twice, and they don't shake their own hands.)*

Solution. This could be a very hard problem to solve if we want to enumerate the different ways in which hands are shook. Suppose that person #1 shakes the hands of persons #2 through #11 during the party. That implies that persons #2 through #11 also shook at least one hand. Now suppose person #2 shook 20 hands that night. We know one of them was person #1, but the other 19 handshakes could have gone to either persons that shook hands with person #1 (except for person #2 of course), or to persons that didn't shake hands with person #1. This is going to matter if we want to keep track. Things will get exponentially more complicated by the time we reach person $\#n$.

We could spend hours trying to figure it out for ourselves, or we could program a computer to go through the enumeration for us, but it turns out that both approaches are needlessly complicated.

We can use the *pigeonhole principle* instead. The pigeonhole principle states that if there are m pigeons in n pigeonholes and $m > n$, then there has to be at least one pigeonhole with more than one pigeon in it.[11]

In terms of our problem, we can look at the pigeonhole principle as follows. Imagine each person as a marble, so there are n marbles in total. Imagine now that we are going to put these marbles into urns as follows: marble $\#k$ goes into urn $\#i$ if person $\#k$ shook exactly i hands. A person can shake $0, 1, 2, \dots, n-1$ hands at the party, so there are n urns, i.e.,

$$\underset{\text{urn \#0}}{\sqcup} \qquad \underset{\text{urn \#1}}{\sqcup} \qquad \underset{\text{urn \#2}}{\sqcup} \qquad \cdots \qquad \underset{\text{urn \#(n-1)}}{\sqcup}\ .$$

[11] A simplified version of the pigeonhole principle: suppose your sock drawer only contains blue and brown socks. If you pick out your socks without looking, how many socks must you take to be sure that you get a pair? The answer is three socks: if you take two socks, then it could happen that you get a pair, but it could also happen that you take a blue one and a brown one. But now if you take a third sock, then you either have one blue sock and two brown socks, or two blue socks and one brown sock. Either way you get a pair.

But note that urn #0 and urn #(n-1) cannot both contain a marble, since if urn #(n-1) contains a marble, then there is one person at the party that shook hands with everyone, which would imply that urn #0 is empty. Conversely, if urn #0 contains a marble, then urn #(n-1) cannot contain a marble for the same reason. Therefore, there are effectively $(n-1)$ urns and n marbles. Now, by the pigeonhole principle, there must be an urn that contains more than one marble. Therefore, there must be at least two persons that shook the same number of hands that night. Therefore, the probability of this event is simply 1. That is to say, it *cannot happen* that everybody at a party shakes a different number of hands than everybody else. $\Diamond$

3. The axioms of probability theory

In the previous chapter we looked at combinatorics, and the types of probabilities that we can calculate using them. In particular, we noted that the combinatorial principles are useful for treating experiments where every single outcome has the *same* probability. But oftentimes it is either too cumbersome or simply impossible to express our experiment as one with equally probable outcomes. What we want is a probabilistic theory where we can decide for ourselves what the probability of a single outcome is going to be, so that we can match it to our data or our theory. To that end, we are going to start at the very foundation and work our way up towards a theory of probability that has a large amount of flexibility.

3.1. Sample spaces and events.

Our first task will be to give a general description for the events of which we want to calculate the probabilities. Let's start by defining what a sets, a subsets and events are:

Definition 1 [Sets and subsets].

(a) *A* set *is a well-defined collection of distinct objects (a.k.a. elements).*
(b) *If A and B are sets, then B is a* subset *of A if and only if all elements of B are also present in A. We write $B \subset A$ if B is a subset of A.*
(c) *We define the* empty set *as the set with no elements and write $\varnothing$.*
(d) *We define the* cardinality *of the set A as the number of elements in A and write $|A|$.*

Example 5. Some examples of sets:

(a) $A = \{1,2,3,4,5,6,7,8\}$ is a set. $B = \{2,4,5\}$ is a subset of A, i.e., $B \subset A$. The cardinality of A is 8, i.e., $|A| = 8$.
(b) $C = \{c,d,e,f,g\}$ is a set. $D = \{c\}$ is a subset of C, i.e., $D \subset C$. The empty set is also a subset of C (or of *any* set, for that matter), i.e., $\varnothing \subset C$. Finally, $|C| = 5$, $D = 1$, and $|\varnothing| = 0$.
(c) $E = [0,3]$ is the set of all real numbers between 0 and 3 (i.e., 2.33349373637..., with any number of decimal points). $F = [2,2.5]$ is a subset of E. So is $G = \{1,2\}$ (the

set consisting only of the numbers 1 and 2.). The cardinality of E and F is both infinity,[12] i.e., $|E| = |F| = \infty$. The cardinality of G is 2.

(d) $H = \{\text{shoes, toothbrush, comb, t-shirt, socks}\}$ is a set (e.g., the set of things in your suitcase). The cardinality of H is 5.

△

Definition 2 [Sample space and events]. *For an experiment we define*

(a) The sample space S *as the set of all possible outcomes of the experiment.*

(b) An event E *as any set that is a subset of* S*, i.e.,* E *is an event if and only if* $E \subset S$.

Let's investigate these definitions with a couple of examples:

Example 6. (a) Roll a die. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. A possible event could be $E = \{\text{roll is even}\} = \{2, 4, 6\}$. Note that $E \subset S$.

(b) Toss two coins. The sample space is $S = \{\text{HH, HT, TT, TH}\}$. A possible event could be $E = \{\text{second coin is tails}\} = \{\text{HT, TT}\}$.

(c) The outcome of a race between 5 horses. The sample space is $S = \{\text{all } 5! \text{ permutations of } (1, 2, 3, 4, 5)\}$. An event could be $E = \{\text{horse 2 finishes before horse 4}\}$.

(d) The amount of rainfall on a day in millimetres per square meter. The sample space is $S = \{t : 0 \leq t < \infty\} = [0, \infty)$. An event could be $E = \{\text{it doesn't rain}\} = \{0\}$ or $F = \{\text{between 2 and 4 millimetres of rain}\} = (2, 4)$.

△

The sample space and the events are all sets. Sets have their own rules of arithmetic, and since we will be working with sets from here on, it is necessary that we master the arithmetic of sets. Set arithmetic is not unlike number arithmetic. Number arithmetic has addition $(+)$ and multiplication $(\times)$ as its most basic operations. Set arithmetic has its own basic operations: *union* $(\cup)$ and *intersection* $(\cap)$.

Definition 3 [Set arithmetic]. *Given two sets, E and F we define*

(a) *The union* $E \cup F$ *as the set of all elements that is either in* E *or in* F *or in both.*

(b) *The intersection* $E \cap F$ *as the set of all elements that is both in* E *and in* F.

When reading set-arithmetic equations and formulas, it may help you to read $A \cup B$ as 'A or B happens' and of $A \cap B$ as 'A and B happen'.

Let's again look at some examples:

Example 7. (a) Let $E = \{1, 2, 4\}$ and $F = \{2, 3, 4, 6\}$. The union of E and F, is set their combined elements , i.e., $E \cup F = \{1, 2, 3, 4, 6\}$ (we remove double elements). The intersection of E and F is the set of elements that is in both, so $E \cap F = \{2, 4\}$.

(b) Let $E = \{\text{all odd rolls of a die}\} = \{1, 3, 5\}$ and let $F = \{\text{all rolls greater than 4}\} = \{5, 6\}$. The union $E \cup F = \{1, 3, 5, 6\}$ and the intersection $E \cap F = \{5\}$.

[12] The study of the cardinality of intervals of real lines is complicated and more sophisticated that it is represented here. If you're interested, look up 'cardinal numbers' on Wikipedia.

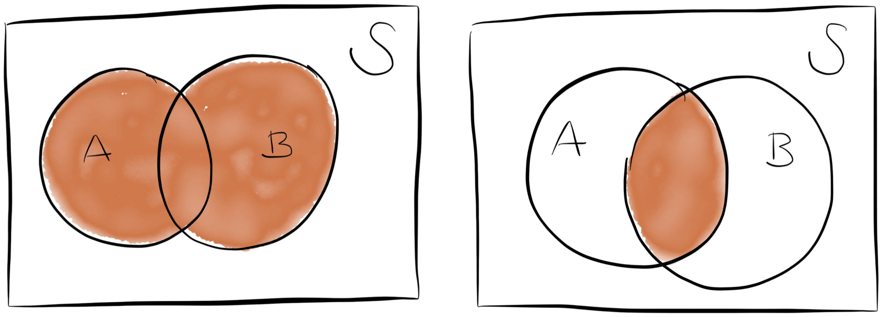

FIGURE 1. Two Venn diagrams. The rectangle represents the sample space S, the two circles represent the events A and B. On the left, the union $A \cup B$ has been shaded, on the right, the intersection $A \cap B$ is shaded.

(c) Let $E = \{\text{apple, lime, banana, orange, grape, lemon}\}$ and let $F = \{\text{all citrus fruits}\}$. The union $E \cup F = \{\text{apple, banana, grape, and all citrus fruits}\}$ and the intersection $E \cap F = \{\text{lime, orange, lemon}\}$.

(d) Let $E = [2,7]$ (all real numbers between 2 and 7), and let $F = [3,9]$. The union $E \cup F = [2,9]$ and the intersection $E \cap F = [3,7]$.

△

When doing math it is often a good idea to draw some schematic diagrams of the thing you're trying to study. We can draw such schematic diagrams for set arithmetic as well. One of the simplest ways of doing this is using so-called *Venn diagrams*. In Venn diagrams the sample space is represented as a square, and the events are represented as (possibly intersecting) circles. Figure 1 depicts a pair of shaded Venn diagrams.

Sometimes we want to exclude a set, or write it as the negation of an event. For those cases we have the following definitions:

DEFINITION 4 [Relative complement and complement]. *Given a sample space S and two sets, E and F we define*

(a) The relative complement *of F in E as $E \setminus F$, the event that contains all elements of E that are* not *in F.*

(b) The complement *of E as E^c, the event that contains all elements of S that are not in E, i.e., $E^c = S \setminus E$.*

See figure 2 for Venn diagrams of the above definition.

Finally, we have the following important definition:

DEFINITION 5 [Mutually exclusive sets]. *Two sets A and B are* mutually exclusive *if and only if $A \cap B = \varnothing$. If A and B are mutually exclusive, we will sometimes write this as $A \dot{\cup} B$.*

See figure 3 for a Venn diagram. Observe that by this definition, A and A^c are mutually exclusive for *any* set A.

Given the definitions given above, the following arithmetic rules for sets can be verified:

Principle 6 [Basic rules of set arithmetic]. *Given a sample space S and events E, F, and G,*

(a) $(E \cup F) \cup G = E \cup (F \cup G) = E \cup F \cup G$.
(b) $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$.
(c) $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$.
(d) $(E^c)^c = E$.
(e) $S^c = \varnothing$.

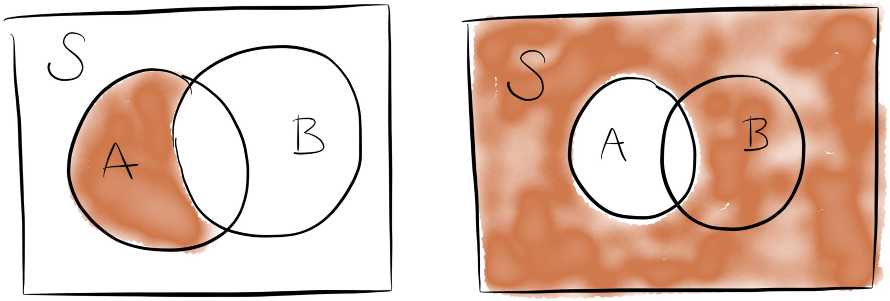

Figure 2. Two Venn diagrams. The rectangle represents the sample space S, the two circles represent events A and B. On the left $A \smallsetminus B$, the relative complement of the B in A has been shaded, on the right A^c, the complement the A is shaded.

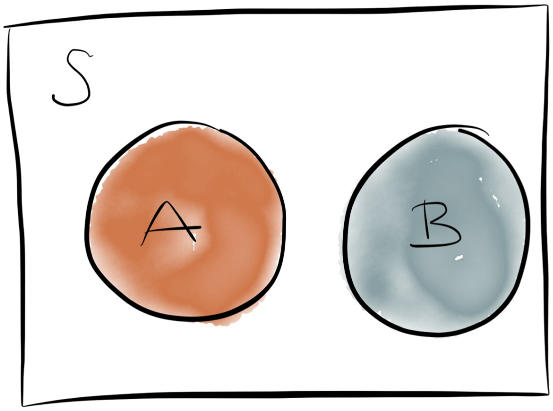

Figure 3. A Venn diagram for two mutually exclusive events.

There is another principle that is less obvious:

PRINCIPLE 7 [DeMorgan's laws]. *Given a sample space S and sets E_i, for $i = 1, 2, \dots, n$,*

$$\left(\bigcup_{i=1}^{n} E_i\right)^c = \bigcap_{i=1}^{n} E_i^c \qquad \text{and} \qquad \left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c.$$

To convince yourself of DeMorgan's laws, draw a Venn diagram for the case $n = 2$, i.e., illustrate that $(A \cup B)^c = A^c \cap B^c$ and that $(A \cap B)^c = A^c \cup B^c$. There is a proof of DeMorgan's laws at the end of Section 2.2 in Ross.

3.2. The axioms of probability.

What we want is to express the *probability* of an event E as $\mathbb{P}(E)$. In particular, we want a function $\mathbb{P}(\,\cdot\,)$ that takes events as its input and that gives a number between 0 and 1 as its output. Besides this, we want as much flexibility as possible, but we don't want to make any 'unnatural' assumptions[13] on the events or on the probability function $\mathbb{P}$. To this end, Andrey Kolmogorov came up with the following three basic assumptions on the probability function $\mathbb{P}$, also known as *axioms* on which we can base the entire study of probability theory:

THE AXIOMS OF PROBABILITY. *Given a sample space S and and any event $E \subset S$,*

- AXIOM 1:
$$0 \leq \mathbb{P}(E) \leq 1.$$
- AXIOM 2:
$$\mathbb{P}(S) = 1.$$
- AXIOM 3: *For any sequence of mutually exclusive events $E_1, E_2, \dots$ (i.e., $E_i \cap E_j = \varnothing$ when $i \neq j$),*
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

From these three axioms we can derive all sorts of general properties of probability functions. Then, if we have an experiment that we want to describe in terms of probabilities, we can take our data or theory about the experiment and derive from it a probability function $\mathbb{P}$. We will then check whether our choice of $\mathbb{P}$ satisfies the axioms. If it does, then it is a 'true' probability, and all these general properties will hold for our choice of $\mathbb{P}$. If it doesn't satisfy the axioms, then we can't be sure that those properties hold (most likely they won't). So it is very important that any probability we choose from now on satisfies the axioms.

We can roughly translate these three axioms into words as follows:

- "AXIOM 1:" Any outcome of an experiment will have a probability that is between 0 (if it never happens) and 1 (if it always happens).
- "AXIOM 2:" If we do an experiment, then 'something' has to happen.

[13] E.g., we want to exclude the possibility that a probability is negative, or that it is possible that a single experiment can have more than one outcome at the same time.

- "Axiom 3:" If two outcomes cannot happen at the same time (e.g. heads and tails) then their probabilities can be added.

The third translation is actually a fair deal weaker than axiom 3,[14] but it does serve to illustrate the 'naturalness' of the axiom.

Let's consider some examples of functions that satisfy the three axioms of probability:

Example 8. Suppose S is finite: $S = \{1, 2, \dots, m\}$.

(a) Given numbers p_j, $j = 1, \dots, m$ that satisfy

$$0 \le p_j \le 1 \qquad \text{and} \qquad p_1 + p_2 + \cdots + p_m = 1,$$

a probability is defined by setting

$$\mathbb{P}(\{j\}) = p_j \qquad \text{and} \qquad \mathbb{P}(E) = \sum_{j \in E} \mathbb{P}(\{j\})$$

for any event $E \subset S$.

(b) Equally likely outcomes: take $p_j = \frac{1}{m} = \frac{1}{|S|}$ for each j. Then a probability is defined by setting

$$\mathbb{P}(E) = \sum_{j \in E} \frac{1}{|S|} = \frac{|E|}{|S|} = \frac{\#\text{ of outcomes in } E}{\#\text{ total outcomes}}.$$

(Note that we have used this construction of a probability many times already in the first chapter. This verifies that what we did there was ok.)

(c) Pick an integer n and a number $p \in [0, 1]$. By the binomial theorem,

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

So if we set $S = \{0, 1, \dots, n\}$ and $p_j = \binom{n}{j} p^j (1 - p)^{n-j}$ we get a probability by setting

$$\mathbb{P}(E) = \sum_{j \in E} \binom{n}{j} p^j (1 - p)^{n-j}.$$

(This is a very important probability, also known as the *binomial distribution*. We will return to it many times.)

△

We can determine some of the immediate consequences of the axioms of probability:

Principle 8. *For any event E,*

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E) \qquad \textit{and} \qquad \mathbb{P}(\varnothing) = 0.$$

[14] In axiom 3 we use the union of an infinite number of events: this may seem 'unnatural', and some mathematicians in the past have objected to this axiom (just as some have objected to infinite sums, infinite integrals, etc.), but it will make our life a lot easier later on, so we will be pragmatic and use it nonetheless.

Proof: Using that $E \cup E^c = S$ and that $E \cap E^c = \varnothing$, we can apply AXIOMS 2 and 3 to determine

$$1 = \mathbb{P}(S) = \mathbb{P}(E \cup E^c) = \mathbb{P}(E) + \mathbb{P}(E^c),$$

so it follows that

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E),$$

and in particular,

$$\mathbb{P}(\varnothing) = \mathbb{P}(S^c) = 1 - \mathbb{P}(S) = 0.$$

□

PRINCIPLE 9. *If $E \subset F$, then*

$$\mathbb{P}(E) \leq \mathbb{P}(F).$$

Proof: We can write F as a union of two mutually exclusive events: $F = E \dot\cup (F \cap E^c)$ and use first AXIOM 3 and then AXIOM 1:

$$\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c) \geq \mathbb{P}(E).$$

□

PRINCIPLE 10. *Given two events E and F,*

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).$$

Proof: We can write $E \cup F$ as three mutually exclusive events:

$$E \cup F = (E \cap F^c) \dot\cup (E \cap F) \dot\cup (E^c \cap F) =: A \dot\cup B \dot\cup C.$$

By AXIOM 3,

$$\mathbb{P}(E \cup F) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C).$$

But also note that by the definitions of A, B and C and AXIOM 3,

$$\mathbb{P}(E) = \mathbb{P}(A) + \mathbb{P}(B) \qquad \text{and } \mathbb{P}(F) = \mathbb{P}(B) + \mathbb{P}(C).$$

Adding up these two equations gives

$$\begin{aligned} \mathbb{P}(E) + \mathbb{P}(F) &= \mathbb{P}(A) + 2\mathbb{P}(B) + \mathbb{P}(C) \\ &= \mathbb{P}(E \cup F) + \mathbb{P}(B) \\ &= \mathbb{P}(E \cup F) + \mathbb{P}(E \cap F). \end{aligned}$$

Subtracting $\mathbb{P}(E \cap F)$ on both sides completes the proof. □

QUESTION 14. *In Beverly Hills, 10% of the population is rich, 5% is famous, and 3% is rich and famous. Pick a person at random from the population of Beverly Hills. What is the probability that this person is either rich or famous?*

Solution. Define the events

$$R := \{\text{person is rich}\} \qquad \text{and} \qquad F := \{\text{person is famous}\}.$$

It is given that

$$\mathbb{P}(R) = 0.1, \qquad \mathbb{P}(F) = 0.05, \qquad \text{and} \qquad \mathbb{P}(R \cap F) = 0.03,$$

Applying Principle 10 we get that

$$\begin{aligned}\mathbb{P}(\{\text{person is rich or famous}\}) &= \mathbb{P}(R \cup F)\\ &= \mathbb{P}(R) + \mathbb{P}(F) - \mathbb{P}(R \cap F)\\ &= 0.1 + 0.05 - 0.03\\ &= 0.12,\end{aligned}$$

so the probability of picking someone either rich or famous is 12%. ◊

Principle 10 can be generalised to the following identity:

Principle 11 [The inclusion-exclusion identity].

$$\begin{aligned}\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} \mathbb{P}(E_i) - &\sum_{0 \le i_1 \le i_2 \le n} \mathbb{P}(E_{i_1} \cap E_{i_2}) + \cdots\\ &+ (-1)^{r+1} \sum_{0 \le i_1 \le i_2 \le \cdots \le i_r \le n} \mathbb{P}(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r})\\ &+ \cdots + (-1)^{n+1} \mathbb{P}(E_1 \cap \cdots \cap E_n).\end{aligned}$$

This is a pretty complicated identity, but for small n it is fairly simple, e.g., for $n = 3$,

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

See figure **??** for a sketch.

Question 15 [The matching problem]. *In preparation of a ‘Secret Santa’ party, n friends draw a slip of paper with a name on it from a hat with one slip for each of the friends. What is the probability that nobody draws the lot with their own name on it?*

Solution. Let E_i denote the event that friend i draws his or her own lot. The probability that no one person draws their own name can be written as

$$\mathbb{P}(A_n) := \mathbb{P}(\text{no one draws their own lot}) = 1 - \mathbb{P}\left(E_1 \cup E_2 \cup \cdots \cup E_n\right). \tag{2}$$

If $n = 2$ then by Principle 10,

$$\begin{aligned}\mathbb{P}(A_2) = 1 - \mathbb{P}(E_1 \cup E_2) &= 1 - \mathbb{P}(E_1) - \mathbb{P}(E_2) + \mathbb{P}(E_1 \cap E_2)\\ &= 1 - \frac{1}{2} - \frac{1}{2} + \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

(Here we used that $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) = \frac{1}{2}$ because if there are two lots and friend 1 draws his own name, then friend 2 has to draw her name as well.

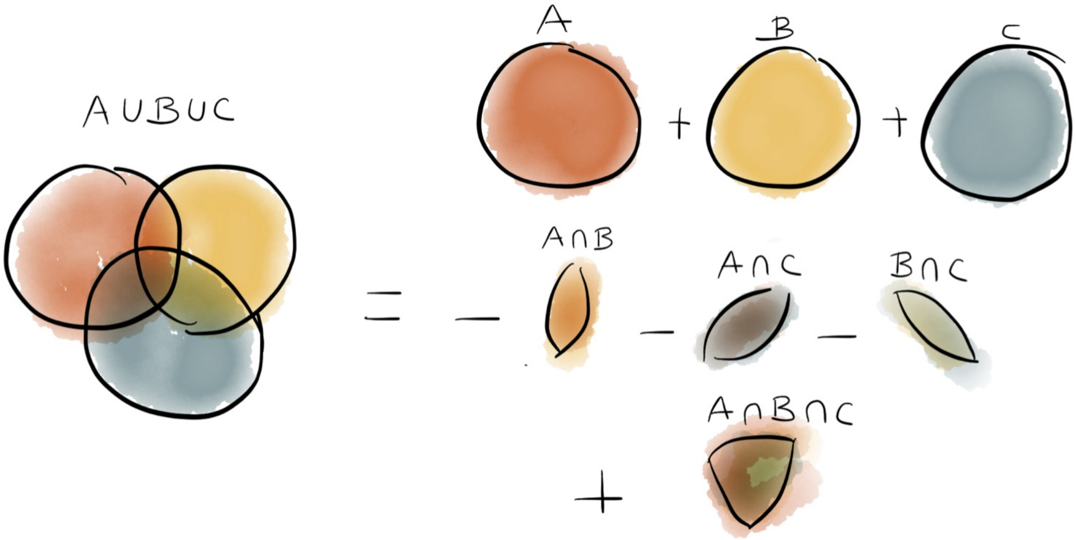

FIGURE 4. A schematic drawing of the inclusion-exclusion identity in terms of Venn diagrams.

For general n we can apply the inclusion exclusion identity (Principle 11) to the probability on the right-hand side of (2). We want to calculate for all r between 1 and n the sum

$$\sum_{1\le i_1<i_2<\cdots<i_r\le n} \mathbb{P}(E_{i_1}\cap E_{i_2}\cap\cdots\cap E_{i_r}).$$

We know from combinatorics that there are $n!$ ways of distributing n lots. If friends $i_1, i_2, \ldots, i_r$ draw their own lots, then there are $(n-r)!$ ways of distributing the remaining $(n-r)$ lots among the remaining $(n-r)$ friends, so the probability that friends $i_1, i_2, \ldots, i_r$ draw their own lots is equal to

$$\mathbb{P}(E_{i_1}\cap E_{i_2}\cap\cdots\cap E_{i_r}) = \frac{\text{\# ways to draw lots of } (n-r) \text{ friends}}{\text{\# ways to draw lots of } n \text{ friends}} = \frac{(n-r)!}{n!}.$$

Also, there are $\binom{n}{r}$ ways of choosing r friends from n, so

$$\sum_{1\le i_1<\cdots<i_r\le n} \mathbb{P}(E_{i_1}\cap\cdots\cap E_{i_r}) = \binom{n}{r}\frac{(n-r)!}{n!} = \frac{n!}{r!(n-r)!}\frac{(n-r)!}{n!} = \frac{1}{r!}.$$

Therefore we can apply the inclusion-exclusion identity to determine

$$\mathbb{P}\left(E_1\cup E_2\cup\cdots\cup E_n\right) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{(-1)^{n+1}}{n!}.$$

This implies that

$$\mathbb{P}(\text{no one draws their own lot}) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} = \sum_{r=0}^{n} \frac{(-1)^r}{r!}.$$

From calculus we know that the Taylor expansion of e^x around $x = 0$ is given by

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

If we choose $x = -1$ we get

$$\mathrm{e}^{-1} = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!}.$$

So when n is large, we have that

$$\mathbb{P}(\text{no one draws their own lot}) \approx \mathrm{e}^{-1} \approx 0.3679.$$

◊

QUESTION 16 [The birthday 'paradox']. *Suppose n persons are in a room. At what value of n is the probability that at least two persons have the same birthday equal to 1/2?*

Solution. So let's ignore February 29. We want to calculate $\mathbb{P}(E)$ where $E = \{$at least two persons have the same birthday$\}$. To determine the probability for smaller values of n, we first use Principle 8, i.e., $\mathbb{P}(E) = 1 - \mathbb{P}(E^c)$. Note that

$$E^c = \{\text{nobody has the same birthday as anybody else}\}.$$

Now it is easy to use combinatorics to determine the value of n such that $\mathbb{P}(E^c) = \frac{1}{2}$:

$$\begin{aligned}
\mathbb{P}(E^c) &= \frac{\#\text{ ways to assign } n \text{ different birthdays}}{\#\text{ ways to assign } n \text{ birthdays}} \\
&= \frac{365 \cdot 364 \cdots (365 - (n-1))}{(365)^n} \\
&= \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{365 - (n-1)}{365} \\
&= 1 \cdot \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) \\
&= \prod_{j=1}^{n-1} \left(1 - \frac{j}{365}\right) \\
&= \frac{1}{365^n} \frac{365!}{(365 - (n-1) - 1)!}.
\end{aligned}$$

Here in the second-to-last step we used the capital-pi notation for a product.[15] From here on it is not too hard to calculate the value of n with a computer. But let's continue by hand to try our hand at some approximation techniques.

Product formulas are typically hard to evaluate, so we are going to use a trick to turn the product into a sum. Recall the following identity:

$$\log(xy) = \log(x) + \log(y),$$

and more generally,

$$\log\left(\prod_{i=1}^{n} a_i\right) = \sum_{i=1}^{n} \log a_i.$$

That is, the logarithm turns products into sums.[16] Furthermore, $x = \mathrm{e}^{\log(x)}$, so we can write

$$\mathbb{P}(E^c) = \mathrm{e}^{\log \prod_{j=1}^{n-1}\left(1-\frac{j}{365}\right)} = \mathrm{e}^{\sum_{j=1}^{n-1}\log\left(1-\frac{j}{365}\right)}.$$

Now we use that $\log(1-x) \approx -x$ is a good approximation[17] when x is small, to get

$$\mathbb{P}(E^c) \approx \mathrm{e}^{-\sum_{j=1}^{n-1}\frac{j}{365}}.$$

Furthermore, it is not hard to prove that $\sum_{j=1}^{k} j = \frac{k(k+1)}{2}$, so we get

$$\mathbb{P}(E^c) \approx \mathrm{e}^{-\frac{1}{365}\frac{n(n-1)}{2}} = \mathrm{e}^{-\frac{n(n-1)}{730}}.$$

So now it remains to solve $\mathrm{e}^{-\frac{n(n-1)}{730}} = \frac{1}{2}$ for n. We do this the old-fashioned way:

$$\frac{1}{2} = \mathrm{e}^{-\frac{n(n-1)}{730}} \quad \Rightarrow \quad n^2 - n + 730\log\frac{1}{2} = 0 \quad \Rightarrow \quad n = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 2920\log\frac{1}{2}}.$$

It turns out that $1 - 2920\log\frac{1}{2} \approx 2025 = 45^2$ so

$$n \approx \frac{1}{2} + \frac{1}{2}\sqrt{45^2} = \frac{1}{2} + \frac{45}{2} = 23.$$

[15] This is the product analogue of the capital-sigma notation for summation, that is,

$$\prod_{i=0}^{3}(1+i) = (1+0)(1+1)(1+2)(1+3) = 24.$$

[16] We are going to use the natural logarithm (with base number $\mathrm{e} = 2.71828\ldots$), but this identity holds for any base number.

[17] This follows from the Taylor expansion of $\log(1-x)$ around $x = 0$, which gives

$$\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n} = -x - \frac{x^2}{2} - \ldots$$

So it turns out that the probability that two people have the same birthday is already 50% when there are just 23 persons in the room![18] (And when there are 50 persons in the room, $\mathbb{P}(E) \approx 0.97$.) $\Diamond$

[18] That a number as small as 23 should be the answer may seem a bit surprising at first sight, but it is certainly no paradox. Among 23 persons there are $\binom{23}{2} = 253$ pairs, and each pair has probability $\frac{1}{365}$ of having the same birthday.

4. Conditional probabilities

In the previous section we discovered how to set up a probability function $\mathbb{P}$ for a given experiment: we should first determine the sample space, and then determine the probabilities for each possible outcome in that sample space.[19] If we know these things, then we can calculate the probability of any given event.

So imagine that we have done all that work, and we are ready to start our experiment, but then we discover that actually some partial information about outcome is available. With this information can rule out some of the outcomes in our sample space. That implies that the sample space that we chose when we set up the probability is actually too big, and so our probabilities are all wrong. We need to start over.

Fortunately, there is a pretty easy way to figure out this new probability function, because we can use the sample space, the probability function and the information. Consider the following example:

Example 9. We flip a coin twice. Our sample space is $S = \{\text{HH, HT, TH, TT}\}$. The event that we get heads twice is $E = \{\text{HH}\}$, and the probability of E is

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{1}{4}.$$

Now we are informed that the first toss is heads. As a result, we can rule out the outcomes TT and TH. Our new sample space is thus $T = \{\text{HH, HT}\}$ and for our new probability function we are going to write $\mathbb{Q}$. That is, information reduces the sample space. Given the information T, the probability of getting heads twice is

$$\mathbb{Q}(E) = \frac{|E|}{|T|} = \frac{1}{2}.$$

Note that T is an event, that is $T \subset S$. Also note that we can rewrite for any event $F \subset T$,

$$\mathbb{P}(F) = \frac{|F|}{|S|} = \frac{|F|}{|T|}\frac{|T|}{|S|} = \mathbb{Q}(F)\mathbb{P}(T),$$

so that

$$\mathbb{Q}(F) = \frac{\mathbb{P}(F)}{\mathbb{P}(T)}.$$

△

The above example shows that given a probability function $\mathbb{P}$ and some information T, we can define a new probability function $\mathbb{Q}$ that takes the information T into account. The probability function $\mathbb{Q}$ is called a the conditional probability function. See Figure **??** for a sketch.

We can generalise the construction of Example 9 so that it works for any probability function and any information T:

[19] There are cases where the sample space is so large that we cannot do this. We will discuss this case in much detail in upcoming sections. For now, assume that our sample space is not too large.

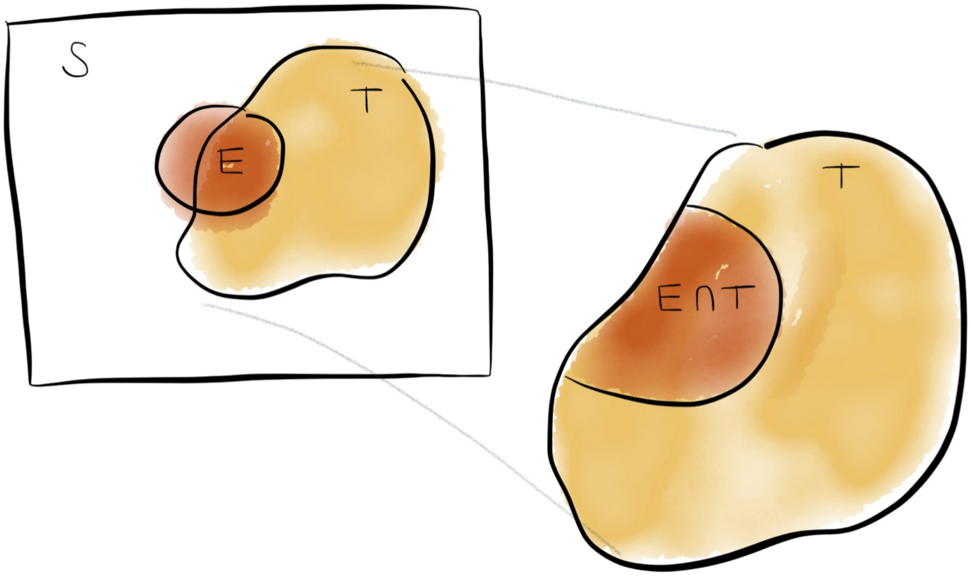

FIGURE 5. A schematic drawing of conditioning E on the information T in terms of Venn diagrams.

DEFINITION 6 [Conditional probability]. *If $\mathbb{P}(T) > 0$, then we define the probability of E given information T as*

$$\mathbb{P}(E \mid T) = \frac{\mathbb{P}(E \cap T)}{\mathbb{P}(T)}.$$

We say that $\mathbb{P}(E \mid T)$ is the 'conditional probability of E given that T occurred'.

QUESTION 17. *We roll 2 dice. Recall from Example 2 that the probability that the sum of two dice rolls is 8 is $5/36$. Now it is revealed that the outcome of the roll of the second die is 3. Now what is the probability that the sum of the two dice rolls is 8?*

Solution. Our sample space for two dice rolls is $S = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$. Each outcome in S is equally probable and $|S| = 6$. The event that we want to calculate the probability of is $E = \{(2,6),(3,5),\dots,(6,2)\}$. Our information is $T = \{(1,3),(2,3),\dots,(6,3)\}$. Note that $E \cap T = \{(5,3)\}$ and that $|T| = 6$. Therefore,

$$\mathbb{P}(E \mid T) = \frac{\mathbb{P}(E \cap T)}{\mathbb{P}(T)} = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{6}.$$

So, given the information T, the probability that we rolled a sum of 8 increases to $1/6$. ◊

QUESTION 18. *Toss two coins again. What is the probability that the first toss comes up heads if we know that at least one coin comes up heads?*

Solution. Our sample space for two coin tosses is $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. All outcomes in S are equally likely. Our event is $E = \{\text{HH}, \text{HT}\}$. Our information is $T = \{\text{HH}, \text{HT}, \text{TH}\}$. The conditional probability of E given T is

$$\mathbb{P}(E \mid T) = \frac{\mathbb{P}(E \cap T)}{\mathbb{P}(T)} = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3}.$$

◊

4.1. **Conditional probabilities are probabilities.** It is important to remember that any conditional probability is a full-fledged probability function. That is, for any information T such that $\mathbb{P}(T) > 0$ (i.e., for any information that isn't impossible), the conditional probability function $\mathbb{P}(\,\cdot \mid T)$ satisfies the three axioms of probability theory:

Given a sample space S, and events $E, T \subset S$ such that $\mathbb{P}(T) > 0$,

(#1)

$$0 \le \mathbb{P}(E \mid T) \le 1.$$

(#2)

$$\mathbb{P}(S \mid T) = 1.$$

(#3) For any sequence of mutually exclusive events $E_1, E_2, \ldots,$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i \,\middle|\, T\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i \mid T).$$

Conditional probabilities can even come in handy when we have no information, because sometimes conditional probabilities are easier to calculate than normal probabilities. In particular, the following principle comes in handy very often:

PRINCIPLE 12.

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

Proof: The identity follows directly from rewriting the definition of a conditional probability.

QUESTION 19. *Draw two cards from a deck of 52. What is the probability of drawing two aces?*

Solution. Write E_1 for the event that the first card is an ace, and E_2 for the event that the second one is an ace. We want to calculate $\mathbb{P}(E_1 \cap E_2)$. We can use Principle 12:

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2 \mid E_1) = \frac{4}{52} \cdot \frac{3}{51}.$$

Note that this calculation is much easier than the counting techniques we used before. ◊

Question 20. *The probability that a randomly chosen snake at the zoo is venomous is 15%. The probability that an Australian snake is venomous is 80%. About 10% of the snakes at the zoo is Australian. What percentage of the snakes at the zoo is venomous but not Australian?*

Solution. Write V for the event that a snake is venomous. Write A for the event that a snake is Australian. Observe that $\mathbb{P}(V) = 0.15$, that $\mathbb{P}(A) = 0.1$ and that $\mathbb{P}(V \mid A) = 0.8$. We want to calculate $\mathbb{P}(V \cap A^c)$. Note that

$$\mathbb{P}(V) = \mathbb{P}(V \cap A) + \mathbb{P}(V \cap A^c),$$

so

$$\mathbb{P}(V \cap A^c) = \mathbb{P}(V) - \mathbb{P}(V \cap A).$$

We use Principle 12 to write $\mathbb{P}(V \cap A) = \mathbb{P}(V \mid A)\mathbb{P}(A)$:

$$\mathbb{P}(V \cap A^c) = \mathbb{P}(V) - \mathbb{P}(V \mid A)\mathbb{P}(A) = 0.15 - 0.8 \times 0.1 = 0.07.$$

So 7% of the non-Australian snakes at the zoo is venomous. ◊

We can generalise Principle 12 to more than two events:

Principle 13 [The multiplication rule].

$$\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) = \mathbb{P}(E_1) \times \mathbb{P}(E_2 \mid E_1) \times \mathbb{P}(E_3 \mid E_1 \cap E_2) \times \cdots \times \mathbb{P}(E_n \mid E_1 \cap \cdots \cap E_{n-1})$$

Proof. Apply Definition 6 to each term on the right-hand side:

$$\frac{\mathbb{P}(E_1)}{1} \times \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} \times \frac{\mathbb{P}(E_1 \cap E_2 \cap E_3)}{\mathbb{P}(E_1 \cap E_2)} \times \cdots \times \frac{\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n)}{\mathbb{P}(E_1 \cap \cdots \cap E_{n-1})}.$$

This is a telescoping product, so everything is canceled except for the first denominator and the last numerator. □

Question 21. *Recall Question 15 about the probability that nobody draws their own name in preparation of a secret Santa party. We calculated that the probability that nobody in a group on n friends draws their own name is*

$$P_n = \sum_{r=0}^{n} \frac{(-1)^r}{r!} \approx \frac{1}{e}.$$

What is the probability that exactly k friends draw their own name?

Solution. There are $\binom{n}{k}$ ways of choosing k friends. First we are going to calculate the probability that a fixed group of k friends draws their own name, and the others don't. Write E for the event that these k friends all draw their own name, and write F for the event that all of the $n - k$ other friends don't draw their own name. So we want to calculate the probability of the event $E \cap F$. We use Principle 12:

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F \mid E).$$

Now write G_i, $i = 1, \dots, k$ for the event that friend i draws his or her own name. Then we can write $E = G_1 \cap \cdots \cap G_k$ so we can use Principle 13 to calculate the probability of E:

$$\begin{aligned}\mathbb{P}(E) = \mathbb{P}(G_1 \cap \cdots \cap G_k) &= \mathbb{P}(G_1)\mathbb{P}(G_2 \mid G_1) \cdots \mathbb{P}(G_k \mid G_1 \cap \cdots \cap G_{k-1}) \\ &= \frac{1}{n}\frac{1}{n-1} \cdots \frac{1}{n-k+1} \\ &= \frac{(n-k)!}{n!}.\end{aligned}$$

The event E implies that k friends draw their own name. If event E occurs, this implies that the other $n-k$ friends will be choosing from among their own $n-k$ names. Therefore,

$$\mathbb{P}(F \mid E) = P_{n-k} = \sum_{r=0}^{n-k} \frac{(-1)^r}{r!}.$$

It follows that

$$\mathbb{P}(E \cap F) = \frac{(n-k)!}{n!} P_{n-k}$$

Now recall that we fixed the group of k friends, so we need to multiply $\mathbb{P}(E \cap F)$ by the number of ways in which we can do this for the final answer, i.e.,

$$\mathbb{P}(\text{exactly } k \text{ friends draw own name}) = \frac{n!}{k!(n-k)!} \frac{(n-k)!}{n!} P_{n-k} = \frac{P_{n-k}}{k!} \approx \frac{1}{ek!}.$$

◊

4.2. **Bayes' rule.** It is a simple consequence of set arithmetic that we can write for any two events E, F,

$$E = (E \cap F) \cup (E \cap F^c).$$

Applying this identity, Axiom #3, and Principle 12 we can derive the following useful rule:

Principle 14 [Total probability rule].
(a) For events E, F,

$$\mathbb{P}(E) = \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c).$$

(b) For a sample space S and events E and $F_1, \dots, F_n$ such that $F_i \cap F_j = \varnothing$ for $i \neq j$ and $\bigcup_{i=1}^n F_i = S$,

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i).$$

See figure 6 for an illustration of the total probability rule.

We have seen how to use added information to modify our probability function. It would be useful if we could reverse this procedure. The following question illustrates why:

Question 22. *Professional cyclists are tested for the performance enhancing drug EPO at the end of a race. The test gives a false positive* 1% *of the time, i.e., the test is sometimes positive even if the cyclist has not used EPO. Moreover, the test gives a false negative* 2% *of the time,*

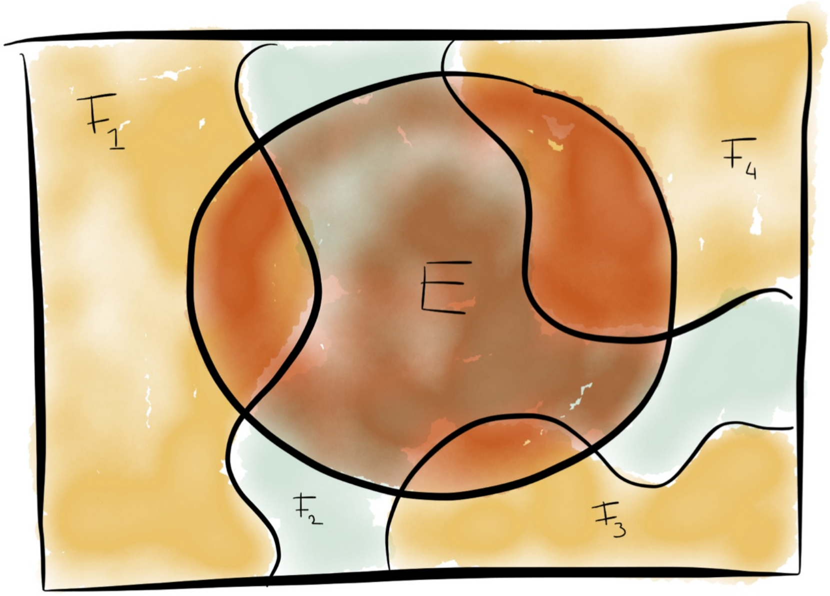

FIGURE 6. A schematic drawing of the total probability rule. The events $F_1, \ldots, F_4$ partition the state space S. By evaluating the probability of E with respect to each of the F_i individually and then adding these probabilities together (according to their own probabilities), we get the same as we would get if we evaluate the probability of E with respect to S.

i.e., the test is negative even if the cyclist has used EPO. Suppose that 3% *of all professional cyclists uses EPO. What is the probability that a cyclist who tests positive has used EPO?*

Solution. Write E for the event that a cyclist uses EPO and P for the event that the test is positive. We are asked to calculate $\mathbb{P}(E|P)$ and we are given that

$$\mathbb{P}(E) = 0.03, \qquad \mathbb{P}(P|E^c) = 0.01 \qquad \text{and } \mathbb{P}(P^c|E) = 0.02.$$

We can rewrite

$$\mathbb{P}(E|P) = \frac{\mathbb{P}(E \cap P)}{\mathbb{P}(P)} = \frac{\mathbb{P}(P|E)\mathbb{P}(E)}{\mathbb{P}(P)} \tag{3}$$

From the data we can derive that $\mathbb{P}(P|E) = 1 - \mathbb{P}(P^c|E) = 0.98$. Moreover, by the total probability rule,

$$\mathbb{P}(P) = \mathbb{P}(P|E)\mathbb{P}(E) + \mathbb{P}(P|E^c)\mathbb{P}(E^c) = 0.98 \times 0.03 + 0.01 \times 0.97 = 0.0391.$$

We insert these numbers into (3) to obtain

$$\mathbb{P}(E|P) = \frac{0.98 \times 0.03}{0.0391} = 0.75,$$

So even with a very accurate test, there is still 25% chance that the cyclist was innocent.[20]

◊

Equation (3) is an incredibly useful equation. It is, in fact, one of the cornerstones of statistics. The formula has a name:

PRINCIPLE 15 [Bayes' rule].

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F)}$$

QUESTION 23. *Suppose there are three cards in a hat: one card is red on both sides, one card is black on both sides, and one card is red on one side and black on the other side. A card is drawn from the hat and placed on the table. The top of the card is red. What is the probability that the bottom of the card is black?*

Solution. Write R for the event that the top is red, and write Rr for the event that both sides are red, Bb that both sides are black, Rb that one side is red, the other black. We want to calculate $\mathbb{P}(Rb|R)$. We apply Bayes' rule:

$$\mathbb{P}(Rb|R) = \frac{\mathbb{P}(R|Rb)\mathbb{P}(Rb)}{\mathbb{P}(R)}.$$

Note that $\mathbb{P}(Rb) = \mathbb{P}(Rr) = \mathbb{P}(Bb) = \frac{1}{3}$. Moreover, by the total probability rule,

$$\begin{aligned}\mathbb{P}(R) &= \mathbb{P}(R|Rr)\mathbb{P}(Rr) + \mathbb{P}(R|Rb)\mathbb{P}(Rb) + \mathbb{P}(R|Bb)\mathbb{P}(Bb)\\ &= 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Therefore,

$$\mathbb{P}(Rb|R) = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

◊

4.3. **Independence.** In the previous sections we saw that information can sometimes drastically change the probability of an event. But sometimes, having information does not change the outcome of any event. For instance, the information that it is snowing on the north pole will not increase your chances of winning the lottery. Having information that does not affect the probability function may seem useless, but in fact this is not so. Realising that knowledge of one event does not change the probability of another event can be incredibly useful information in itself:

[20] Before you start to think that your favourite cyclist may be innocent, keep in mind that this calculation depends strongly on the assumption that only 3% of all cyclists used EPO. If we assume that 30% uses, then the probability that the cyclist is an EPO if the test is positive goes up to about 98%.

Definition 7 [Independent events]. *The events E and F are* independent *if*

$$\mathbb{P}(E|F) = \mathbb{P}(E) \qquad \textit{and} \qquad \mathbb{P}(F|E) = \mathbb{P}(F).$$

If the events do not satisfy the above equalities, then E and F are dependent.

Note that by the definition of conditional probabilities, the left-hand equation implies the right-hand one, and vice versa, so either of the two equations suffices as a definition.

Independent events are much easier to calculate with than dependent events. In particular

Principle 16. *If E and F are independent events, then*

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F).$$

Example 10. Successive coin tosses are independent. Consider the sample space of two coin tosses $S = \{\text{HH, HT, TH, TT}\}$, then

$$\mathbb{P}(\text{1st toss is heads}) = \mathbb{P}(\{HH, HT\}) = \frac{1}{2} \quad \text{and} \quad \mathbb{P}(\text{2nd toss is tails}) = \mathbb{P}(\{HT, TT\}) = \frac{1}{2},$$

and

$$\mathbb{P}(\{\text{1st toss is heads}\} \cap \{\text{2nd toss is tails}\}) = \mathbb{P}(\{HT\}) = \frac{1}{4}.$$

△

Question 24. *Draw a card from a deck. If E is the event that the card is an ace, and F is the event that is a spade, are E and F independent?*

Solution. Calculate

$$\mathbb{P}(E) = \frac{1}{13}, \qquad , \mathbb{P}(F) = \frac{13}{52}, \qquad \text{and} \qquad \mathbb{P}(E \cap F) = \frac{1}{52}.$$

Note that $\mathbb{P}(E) \times \mathbb{P}(F) = \mathbb{P}(E \cap F)$, so E and F are indeed independent. ◊

Question 25. *For events A and B we know that*

$$\mathbb{P}(A) = 0.4, \qquad \mathbb{P}(B) = 0.3, \qquad \textit{and} \qquad \mathbb{P}(A \cup B) = 0.58.$$

Are A and B independent?

Solution. The events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. From the data it follows that $\mathbb{P}(A)\mathbb{P}(B) = 0.4 \times 0.3 = 0.12$. We can use Principle 10 to determine $\mathbb{P}(A \cap B)$:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

so

$$\mathbb{P}(A \cap B) = 0.4 + 0.3 - 0.58 = 0.12.$$

It follows that A and B are indeed independent. ◊

If two events are independent, then the events are also independent of each other's complements:

PRINCIPLE 17. *If E and F are independent events, then*

$$\mathbb{P}(E \cap F^c) = \mathbb{P}(E)\mathbb{P}(F^c).$$

Proof. Since $E = (E \cap F)\dot{\cup}(E \cap F^c)$ it follows that

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c)\\ &= \mathbb{P}(E)\mathbb{P}(F) + \mathbb{P}(E \cap F^c).\end{aligned}$$

Rearranging gives

$$\begin{aligned}\mathbb{P}(E \cap F^c) &= \mathbb{P}(E) - \mathbb{P}(E)\mathbb{P}(F)\\ &= \mathbb{P}(E)(1 - \mathbb{P}(F))\\ &= \mathbb{P}(E)\mathbb{P}(F^c).\end{aligned}$$

□

If two events are independent, then their complements are also independent:

PRINCIPLE 18. *If E and F are independent events, then*

$$\mathbb{P}(E^c \cap F^c) = \mathbb{P}(E^c)\mathbb{P}(F^c).$$

Proof. Homework.

QUESTION 26. *If E is independent of F and E is independent of G, does it follow that F is independent of G?*

Solution. The answer is 'no'. We prove this by giving a counter example.[21] As we have already seen, if E is independent of F, then it is also independent of F^c. Suppose that $0 < \mathbb{P}(F) < 1$, then it is not hard to see that F and F^c are not independent:

$$\mathbb{P}(F \cap F^c) = \mathbb{P}(\varnothing) = 0, \qquad \text{while} \qquad \mathbb{P}(F)\mathbb{P}(F^c) = \mathbb{P}(F)(1 - \mathbb{P}(F)) > 0$$

so in particular,

$$\mathbb{P}(F \cap F^c) \neq \mathbb{P}(F)\mathbb{P}(F^c).$$

◊

In general it should be noted that any two mutually exclusive events that are not empty sets cannot be independent.

[21] There exist many more counter examples to this statement, see for instance Example 4e on page 76 of Ross.

Question 27. *A sequence of n times the same experiment is performed. Each experiment can end in success, with probability p, or in failure, with probability $1-p$. The outcomes of the experiments are independent.*

(a) What is the probability that at least 1 success occurs?

(b) What is the probability that the first success occurs on the rth trial?

(c) What is the probability that exactly k successes occur?

Solution. Write E_i for the event that the ith experiment is a success.

(a) We want to calculate

$$\begin{aligned}\mathbb{P}(\text{at least 1 success}) &= 1-\mathbb{P}(\text{no success})\\ &= 1-\mathbb{P}(E_1^c\cap E_2^c\cap\cdots\cap E_n^c)\\ &= 1-\mathbb{P}(E_1^c)\mathbb{P}(E_2^c)\cdots\mathbb{P}(E_n^c)\\ &= 1-(1-p)^n.\end{aligned}$$

(b) The event that the rth trial is the first success implies that the first $r-1$ trials are a failure. Therefore,

$$\begin{aligned}\mathbb{P}(\text{1st success on } r\text{th trial}) &= \mathbb{P}(E_1^c\cap E_2^c\cap\cdots\cap E_{r-1}^c\cap E_r)\\ &= \mathbb{P}(E_1^c)\cdots\mathbb{P}(E_{r-1}^c)\mathbb{P}(E_r)\\ &= (1-p)^{r-1}p.\end{aligned}$$

(c) There are $\binom{n}{k}$ sequences in which k successes and $n-k$ failures can occur. Therefore,

$$\mathbb{P}(\text{exactly } k \text{ successes}) = \binom{n}{k}p^k(1-p)^{n-k}.$$

◊

The following is one of the oldest probability problems out there.[22] Like all old probability problems, this one is about gambling.

Question 28 [Gambler's ruin]. *You and a friend are playing a game of coin flip. You flip a coin, and if it comes up heads, you win a dollar from your friend. But if it comes up tails, you lose a dollar to your friend. You keep playing the game until one of you goes broke. You start with r dollars, your friend with $n-r$ dollars (so there are n dollars in total). What is the probability that you go broke?*

Solution. Each coin flip has the same probability of coming up heads (i.e., 1/2), so the game is fair.[23] Write A_r for the event that you lose all your money if you start with r dollars. We

[22] The great Dutch scientist Christiaan Huygens studied it first in 1657. Huygens was one of the most brilliant scientists ever. Besides pretty much inventing probability theory, he also invented the pendulum clock, discovered the rings of Saturn, and was the first to develop the theory of light as waves.

[23] Ross gives the same calculation for an unfair game.

want to calculate $p_r := \mathbb{P}(A_r)$. We can use the total probability rule to split the probability on the event B that your friend wins the first game:

$$p_r = \mathbb{P}(A_r) = \mathbb{P}(B)\mathbb{P}(A_r|B) + \mathbb{P}(B^c)\mathbb{P}(A_r|B^c).$$

Since the game is fair, $\mathbb{P}(B) = \mathbb{P}(B^c) = \frac{1}{2}$. Moreover, the coin flips are independent, so the probability of going broke after you lose the first game is the same as the probability that you go broke if you start with $r - 1$ dollars, i.e.,

$$\mathbb{P}(A_r|B) = \mathbb{P}(A_{r-1}) = p_{r-1}.$$

Similarly,

$$\mathbb{P}(A_r|B^c) = \mathbb{P}(A_{r+1}) = p_{r+1}.$$

As a result,

$$p_r = \frac{1}{2}p_{r+1} + \frac{1}{2}p_{r-1}.$$

We can rewrite

$$p_{r+1} - p_r = p_r - p_{r-1}.$$

It is obvious that $p_n = 0$ so we get the following system of equations:

$$\begin{aligned}
p_n - p_{n-1} &= & 0 - p_{n-1} &= -p_{n-1} \\
p_{n-1} - p_{n-2} &= & p_n - p_{n-1} &= -p_{n-1} \\
p_{n-2} - p_{n-3} &= & p_{n-1} - p_{n-1} &= -p_{n-1} \\
\vdots & & \vdots & \quad \vdots \\
p_2 - p_1 &= & p_3 - p_2 &= -p_{n-1} \\
p_1 - p_0 &= & p_2 - p_1 &= -p_{n-1}.
\end{aligned}$$

It follows that each increment (i.e., the difference between p_i and p_{i-1}) is equal, so

$$p_r = (n - r)p_{n-1}.$$

It is also obvious that $p_0 = 1$, so $1 = (n - 0)p_{n-1}$, which implies that

$$p_{n-1} = \frac{1}{n}.$$

So the probability that you go broke if you start with r dollars is

$$p_r = \frac{n - r}{n}.$$

◊

Question 29. *Suppose you are hiring a secretary and n people apply for the job. You don't want to waste too much time with the application procedure, so you decide to hire a candidate you like on the spot. Your aim is to hire the very best candidate. Your strategy is to interview r applicants without hiring any, and then to hire the next applicant that is better than all the first r applicants. How big should r be to maximise your chance of getting the best candidate?*

Solution. Assume that no two candidates are equally qualified, and that the order in which they do their interview is random. Write

$$\begin{aligned} A &= \{\text{the best applicant is selected}\}, \\ B_i &= \{\text{applicant } i \text{ is the best}\}, \\ C_i &= \{\text{applicant } i \text{ is selected}\}. \end{aligned}$$

We can write A in terms of B_i and C_i using some set arithmetic:

$$A = \bigcup_{i=1}^{n} (B_i \cap C_i) .$$

Note that B_i and B_j are mutually exclusive for $i \neq j$, and similarly, C_i and C_j are mutually exclusive, so by Axiom #3,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i \cap C_i) .$$

Now we apply the definition of conditional probability:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(C_i|B_i)\, \mathbb{P}(B_i).$$

The best candidate could be any of the n candidates, so

$$\mathbb{P}(B_i) = \frac{1}{n} \qquad \text{for all } i = 1, \dots n.$$

Our strategy dictates that we are not going to hire candidate i if $i \leq r$, even if that candidate is the best, so

$$\mathbb{P}(C_i|B_i) = 0 \qquad \text{for all } i = 1, \dots, r.$$

This means that

$$\begin{aligned} \mathbb{P}(A) &= \sum_{i=1}^{r} 0 \times \frac{1}{n} + \sum_{i=r+1}^{n} \mathbb{P}(C_i|B_i)\frac{1}{n} \\ &= \frac{1}{n} \sum_{i=r+1}^{n} \mathbb{P}(C_i|B_i). \end{aligned}$$

Now, for $i > r$, the event that you select candidate i, given that i is the best, implies that the best candidate among the first $i-1$ candidates was actually among the first r candidates. This must be the case, because if you interviewed a candidate that was better than the first r candidates before you interviewed candidate i, than that candidate would have been hired. Since the event is that candidate i gets hired, this must not happen. Write D_{i-1} for the event that the best candidate among the first $i-1$ candidates is among the first r candidates. Our hiring strategy dictates that, conditioned on the event B_i we have $C_i = D_{i-1}$ for $i > r$. Moreover, the event D_{i-1} only depends on the *relative order* of the first $i-1$ candidates. Because we are only interested in the ranking of these $i-1$ candidates among themselves, this is *independent* of how they compare to candidates $i, \dots, n$, so D_{i-1} is also independent the event B_i, i.e.,

$$\mathbb{P}(C_i|B_i) = \mathbb{P}(D_{i-1}|B_i) = \mathbb{P}(D_{i-1}).$$

We can calculate $\mathbb{P}(D_{i-1})$ by counting:

$$\mathbb{P}(D_{i-1}) = \frac{\text{\# places for best candidate among first } r}{\text{\# places for best candidate among first } i-1} = \frac{r}{i-1}.$$

We plug this into our equation for $\mathbb{P}(A)$:

$$\mathbb{P}(A) = \frac{1}{n}\sum_{i=r+1}^{n}\frac{r}{i-1} = \frac{r}{n}\sum_{j=r}^{n-1}\frac{1}{j}.$$

Using a computer, we can find the optimal value for r:

n	1	2	3	4	5	6	7	8	9
optimal r	0	0	1	1	2	2	2	3	3
$\mathbb{P}(A)$	1	0.5	0.5	0.458	0.433	0.428	0.414	0.410	0.406

Finally, we approximate the sum by an integral to determine the optimal r and $\mathbb{P}(A)$ for large n:

$$\begin{aligned}\mathbb{P}(A) &= \frac{r}{n}\sum_{j=r}^{n-1}\frac{1}{j} \approx \frac{r}{n}\int_r^n \frac{1}{x}\mathrm{d}x = \frac{r}{n}\left[\log x\right|_r^n \\ &= \frac{r}{n}\left(\log(n) - \log(r)\right) = -\frac{r}{n}\log\left(\frac{r}{n}\right).\end{aligned}$$

If we write $t = \frac{r}{n}$, then it follows that

$$\mathbb{P}(A) \approx -t\log t.$$

We want to maximise $\mathbb{P}(A)$, so we need to find the value of $t \in [0,1]$ that maximises $-t\log t$, i.e., we want to solve

$$-\frac{\mathrm{d}}{\mathrm{d}t}t\log t = 0 \qquad \text{and} \qquad -\frac{\mathrm{d}^2}{\mathrm{d}t^2}t\log t < 0. \tag{4}$$

Recall from calculus that

$$-\frac{\mathrm{d}}{\mathrm{d}t}t\log t = -\log t - 1 \qquad \text{and} \qquad -\frac{\mathrm{d}^2}{\mathrm{d}t^2}t\log t = -\frac{1}{t}.$$

You can check that the equations (4) are solved for $t = \frac{1}{e}$. So, if we choose $r = \frac{n}{\mathrm{e}}$, then we get that the probability that we select the best candidate is

$$\mathbb{P}(A) \approx -\frac{1}{\mathrm{e}}\log\frac{1}{\mathrm{e}} = \frac{1}{\mathrm{e}} = 0.368\dots$$

(Using some advanced mathematics you can actually prove that the strategy of passing on r candidates and then picking the best one after that is actually the best possible strategy.) ◊

4.4. **Simpson's paradox.** The following data is from a medical paper comparing the effectiveness of treatments for kidney stones.[24] Two types of treatment were studied: treatment A is surgical removal of the kidney stones, treatment B is percutaneous nephrolithotomy. The data gives the following rates of effectiveness:

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

Something weird is going on. If we look at the total effectiveness of each treatment, then treatment B seems most effective, but if we look at the effectiveness for both small and large stones separately, then treatment A seems most effective in both cases. Which treatment should we recommend?

First of all, we should note that this is not actually a paradox. When we compare these success rates, we are comparing two different probability functions: the sample spaces of the probabilities of having success with treatment A and treatment B, respectively, are not the same. If we look at the effectiveness for treatment A, we sample from a population where 87 patients have small stones, while 263 have large stones, while the distribution of the population for treatment B this ratio is pretty much reversed:

	Treatment group A	Treatment group B
Small stones	87	270
Large stones	263	80
Both	350	350

It turns out that treatment B is less invasive than treatment A, but it is much harder to perform on patients with large kidney stones, so doctors prefer treatment A for large stones and treatment B for small ones. This causes the difference in sample spaces. Both treatments, however, work better for small kidney stones than for large ones. So the reason that treatment B seems more effective on the whole, is that it gets applied to the more easy cases, while the more effective treatment A gets applied to mostly hard cases.

To conclude, if you only care about success (and not invasiveness of the procedure) then treatment A is your best choice.

[24] C. R. Charig, D. R. Webb, S. R. Payne, J. E. Wickham (29 March 1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". *Br Med J (Clin Res Ed)* 292 (6524): 879–882.

5. Random variables

It often happens that the outcome of an experiment is a number, i.e., for each outcome in the sample space S a number X is observed. We call such a number a *random variable*.[25] Studying a problem in terms of random variables gives us a new way of understanding probabilities, because we can treat random variables as *functions* $X : S \to \mathbb{R}$.

Example 11 (Bernoulli random variable). An experiment has only two outcomes: F (failure) and F^c (success), with probabilities $\mathbb{P}(F) = 1-p$ and $\mathbb{P}(F^c) = p$. We define the random variable X as

$$X := \begin{cases} 0 & \text{if the outcome is } F, \\ 1 & \text{if the outcome if } F^c. \end{cases}$$

Now we can write the probabilities of our original problem as

$$\mathbb{P}(F) = \mathbb{P}(X = 0) \qquad \text{and} \qquad \mathbb{P}(F^c) = \mathbb{P}(X = 1).$$

This is one of the simplest random variables that we can define, but it is a very important one nonetheless: it is called the Bernoulli random variable with parameter p. To indicate that random variable X is a Bernoulli random variable with parameter p, we write

$$X \sim \operatorname{Ber}(p)$$

The Bernoulli random variable does not appear to be a particularly useful function. But with it, we can define another function: we define the *probability mass function* of X as $f(x) := \mathbb{P}(X = x)$, i.e.,[26]

$$f(x) := \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1, \\ 0 & \text{for other values of } x. \end{cases}$$

△

We can generalise this example to a define a probability mass function for a given random variable:

Definition 8. *Given a random variable X, the probability mass function of X is defined as*

$$f(x) = \mathbb{P}(X = x) \qquad \textit{for all } x \in \mathbb{R}.$$

Example 12 (Binomial random variable). Repeat n times the Bernoulli experiment with parameter p, and now let

$$X := \#\text{ observed successes.}$$

[25] We will often write random variables as capital letters near the end of the alphabet, e.g., X, Y, or Z. We will reserve the lower case letters for real variables, e.g., x, y and z.

[26] The Ross book writes $p(x)$ for the probability mass function, but I think the letter p is getting a bit overused, so we will typically write $f(x)$.

We call X a *binomial random variable*. X can take the values $0, 1, 2, \dots, n$. To indicate that X is a binomial random variable with parameters n and p we write

$$X \sim \text{Bin}(n, p).$$

The probability mass function of X is given by

$$f(x) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n$$

(recall Example 27(c)). △

Example 13 (Geometric random variable). We repeat a Bernoulli experiment with parameter p until we see the first success, and let

$$X := \text{\# trials until first success.}$$

We call X a geometric random variable with parameter p. X can take the values $1, 2, 3, \dots$. To indicate that X is a geometric random variable with parameter p, we write

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

The probability mass function of X is given by

$$f(x) = \mathbb{P}(X = k) = (1-p)^{n-1} p$$

(recall Example 27(b)). △

In this section we will study discrete random variables. These are random variables that satisfy the following technical definition.

DEFINITION 9. *We say that a random variable X is* discrete *if X takes values on a set $\mathcal{I} \subset \mathbb{R}$ that contains only separated points. That is, $\mathcal{I}$ contains no intervals of numbers, so there exists no a, b such that the interval $[a, b] \subset \mathcal{I}$.*

Example 14 (Discrete random variables). Bernoulli random variables are discrete, because they take values on the set $\{0, 1\}$. Similarly, Binomial random variables are discrete, because they take values on the set $\{0, 1, \dots, n\}$. Geometric random variables take values on the infinite set of natural numbers $\mathbb{N} = \{1, 2, \dots\}$. Other examples of sets $\mathcal{I}$ on which a random variable is discrete are the infinite set of integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$, the set of non-negative numbers rounded to the nearest decimal $\{0, 0.1, 0.2, 0.3, \dots\}$, etc. △

In general, probability mass functions of discrete random variables have the following properties:

PRINCIPLE 19 [Properties of probability mass functions]. *If X is a discrete random variable that takes its values on a set $\mathcal{I} \subset \mathbb{R}$, and $f(x)$ is the probability mass function of X, then*

(a) $f(x) \geq 0$ *for all* $x \in \mathbb{R}$,

(b) $\sum_{x \in \mathcal{I}} f(x) = 1$.

Proof. (a) The probability mass function is defined as a probability for those values of x that correspond to possible values of X, so for those x the p.m.f. is non-negative. For all other values of x, the p.m.f. is defined to equal 0, so there it is also non-negative.

(b) A random variable has to associate a number with every outcome in the state space, so the event S (i.e., the entire sample space) is equivalent tot the event $\{X \in \mathcal{I}\}$ (i.e., the event that X takes some value). Moreover, the random variable associates exactly one value to each outcome in S, so the events $\{X = y\}$ and $\{X = z\}$ are mutually exclusive if $y \neq z$. This, together with Axiom #3 implies that

$$1 = \mathbb{P}(S) = \mathbb{P}(X \in \mathcal{I}) = \mathbb{P}\left(\bigcup_{x \in \mathcal{I}} \{X = x\}\right) = \sum_{x \in \mathcal{I}} \mathbb{P}(X = x) = \sum_{x \in \mathcal{I}} f(x).$$

□

Example 15. Let's do a quick check of Principle 19 for Bernoulli, Binomial, and Geometric random variables:

Bernoulli random variable: If $X \sim \text{Ber}(p)$, then Principle 19(a) is satisfied since it is clearly non-negative, and Principle 19(b) is satisfied because

$$\sum_{x \in \{0,1\}} f(x) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = (1 - p) + p = 1.$$

Binomial random variable: If $X \sim \text{Bin}(n, p)$, then Principle 19(a) is obviously satisfied, and Principle 19(b) is satisfied because by the binomial theorem,

$$\sum_{x \in \{0,1,\ldots,n\}} f(x) = \sum_{k=0}^{n} \mathbb{P}(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1^n = 1,$$

Geometric random variable: If $X \sim \text{Geo}(p)$, then again Principle 19(a) is obviously satisfied. To prove that Principle 19(b) is satisfied we use the *geometric identity* for $-1 < z < 1$

$$\sum_{m=0}^{\infty} z^m = \frac{1}{1 - z},$$

as follows:

$$\sum_{x \in \mathbb{N}} f(x) = \sum_{n=1}^{\infty} (1 - p)^{n-1} p = p \sum_{m=0}^{\infty} (1 - p)^m = p \times \frac{1}{1 - (1 - p)} = \frac{p}{p} = 1.$$

△

Using random variables can make calculations significantly easier:

Question 30. *An unfair coin comes up heads 60% of the time. What is the probability that in ten tosses, the coin comes up heads at least eight times?*

Solution. Let $X = \#$ heads. X counts the number of successes in a sequence of 8 independent trials that all have success probability 0.6 so it follows that $X \sim \operatorname{Bin}(n = 10, p = 0.6)$. We want to calculate $\mathbb{P}(X \geq 8)$:

$$\begin{aligned}
\mathbb{P}(X \geq 8) &= \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10) \\
&= \binom{10}{8} 0.6^8 (1 - 0.6)^2 + \binom{10}{9} 0.6^9 (1 - 0.6)^1 + \binom{10}{10} 0.6^{10} (1 - 0.6)^0 \\
&= 0.12093 + 0.04031 + 0.00605 \\
&= 0.16729.
\end{aligned}$$

◊

QUESTION 31. *You ask randomly chosen people about their birthday until you find someone with the same birthday as yourself. How many people do you have to ask, so that the probability of getting a match is more than 1/2?*

Solution. Let $X = \#$ people you ask until you get a match. Assume that the probability of asking one person and getting a match is $1/365$. Then it follows that $X \sim \operatorname{Geo}(p = 1/365)$. We want to determine r such that $\mathbb{P}(X < r) \geq \frac{1}{2}$, which of course implies that $\mathbb{P}(X \geq r) \leq \frac{1}{2}$:

$$\begin{aligned}
\mathbb{P}(X \geq r) &= \sum_{n=r}^{\infty} (1 - p)^{n-1} p \\
&= (1 - p)^{r-1} p \sum_{m=0}^{\infty} (1 - p)^m \\
&= (1 - p)^{r-1} p \times \frac{1}{1 - (1 - p)} \\
&= (1 - p)^{r-1}.
\end{aligned}$$

It remains to find the smallest r such that

$$(1 - p)^{r-1} \leq \frac{1}{2}.$$

Equating and taking the log on both sides gives

$$(r - 1) \log(1 - p) = \log \frac{1}{2} \qquad \Rightarrow \qquad r = \frac{\log \frac{1}{2}}{\log(1 - p)} + 1 = 252.7,$$

So we need to ask at least 253 people to have more than 50% chance that we get match.[27] ◊

The random variables that we have considered can be used to describe large families of problems. Sometimes, a problem arises that doesn't fit into such a class. In these cases, we can sometimes construct a random variable and p.m.f. from the data, as the following example illustrates:

[27] You can approximate the solution without a calculator if you remember that $\log(1 - x) \approx -x$ when x is small and that $-\log \frac{1}{2} = \log 2 \approx 0.7$.

Example 16. A pond contains five fishes. You go fishing. The probability that you catch 0 fish is 11%, that you catch 1 fish is 31%, 2 fish is 24%, 3 fish is 19%, 4 fish is 8%, and 5 fish is 7%.

We can define a random variable X and p.m.f. $f(x)$ as follows:

$$X = \#\text{ fish you catch} \qquad \text{and} \qquad f(x) = \begin{cases} 0.11 & \text{if } x = 0 \\ 0.31 & \text{if } x = 1 \\ 0.24 & \text{if } x = 2 \\ 0.19 & \text{if } x = 3 \\ 0.08 & \text{if } x = 4 \\ 0.07 & \text{if } x = 5 \\ 0 & \text{otherwise.} \end{cases}$$

△

Here is another example of an important random variable:

Example 17 (Negative binomial). Repeat the Bernoulli experiment with success probability p until you see r times a success. Let

$$Y = \#\text{ trials until the } r\text{th success.}$$

We can calculate the probability mass function of Y (using the 'wall trick' described in Question 9):

$$f(k) = \mathbb{P}(Y = r + k) = \binom{r+k-1}{r-1} p^r (1-p)^k, \qquad \text{for } k = 0, 1, 2, \ldots$$

We write $Y \sim \text{NBin}(r, p)$. Checking Principle 19(b) is not particularly easy, so we will not do that here. △

The following is a classic example[28] of a problem that can be solved with random variables:

QUESTION 32 [Problem of points]. *Two evenly matched teams compete to see who will be the first to win n games. The winner gets prize money. The play gets interrupted before either team wins n games, so it is decided that the prize money will be divided among the two teams. At the time of interruption, Team A has won i games, Team B has won j games. Given this information, what is the most fair division of the prize money?*

Solution. The most fair division is according to the probability that the team would have won if the play had continued, that is, Team A deserves a fraction

$$\mathbb{P}(\text{Team A wins} | \text{ Team A won } i \text{ games, Team B won } j \text{ games}) \tag{5}$$

of the prize money.

[28] It was independently solved by Blaise Pascal and Pierre de Fermat in 1654.

Let's start by solving this problem in the case that $n = 7$, $i = 2$, and $j = 4$. For Team A to win the prize, they have to win 5 games before Team B wins 3 games. Let

$$Y = \#\text{ games until Team A wins 5 games},$$

Observe that Y is a negative binomial random variable with parameters $r = 5$ and $p = 1/2$.

It follows that if Team A is to win, then Y can only take the values 5, 6, or 7 (if Y was any larger, Team B would have won 3 or more games, and Y can't be smaller than 5). This means that the probability (5) is equal to

$$\begin{aligned}\mathbb{P}(Y = 5 \text{ or } 6 \text{ or } 7) &= \mathbb{P}(Y = 5) + \mathbb{P}(Y = 6) + \mathbb{P}(Y = 7)\\ &= \binom{5-1}{5-1}p^5(1-p)^0 + \binom{5+1-1}{5-1}p^5(1-p)^1 + \binom{5+2-1}{5-1}p^5(1-p)^2\\ &= \left(\frac{1}{2}\right)^5\left(1 + 5\frac{1}{2} + \frac{6\cdot 5}{2}\left(\frac{1}{2}\right)^2\right)\\ &= \left(\frac{1}{2}\right)^5\left(\frac{4+10+15}{2^2}\right)\\ &= \left(\frac{1}{2}\right)^7 29 = \frac{29}{128} = 0.227.\end{aligned}$$

We can calculate the general solution in the same way: write $s = n - i$ and $t = n - j$, then the probability in (5) becomes

$$\sum_{m=s}^{s+t-1}\binom{m-1}{s-1}\frac{1}{2^m}.$$

◊

Example 18 (Relation between geometric and negative binomial random variables). By definition, the geometric random variable is a special case of a negative binomial random variable, namely the case where $r = 1$. Observe that moreover, if Y is a $\mathrm{NBin}(r, p)$ random variable, and if $X_1, \ldots, X_r$ are r independent $\mathrm{Geo}(p)$ r.v.'s, then

$$\begin{aligned}Y =&\# \text{ trials until } r\text{th success}\\ =&\# \text{ trials until 1st success } + \# \text{ additional trials until 2nd success } +\\ &+ \cdots + \# \text{ additional trials until } r\text{th success}\\ =&X_1 + X_2 + \cdots + X_r,\end{aligned}$$

that is, a negative binomial random variable with parameters r and p is the sum of r independent geometric random variables with parameter p. △

5.1. **The expected value of a random variable.** If we think about a random variable X as the outcome of an experiment *before* we perform the experiment, than we can ask ourselves what we *expect* to see. There are multiple ways of thinking about what it means to expect to see something. We might for instance say that we expect to see the outcome that has the highest probability of occurring. We might also say that we expect to see the outcome that has a value M such that the probability that the real outcome is larger than M is about the same as the probability that its smaller than M (e.g., both probabilities are about 50%). Both of these interpretations have their own value in the study of probabilities. The first interpretation is known as the *maximum likelihood of* X, the second is the *median of* X.

There is, however, a third way of interpreting the expectation of X, namely, to interpret it at the *average of X weighted with the p.m.f.* It is this interpretation that we will call the *expectation of X:*

Definition 10 [The expectation of a random variable]. *The* expectation *or* expected value *of a discrete random variable X that takes values on the set $\mathcal{I}$ is given by*

$$\mathbb{E}[X] := \sum_{x\in\mathcal{I}} x\mathbb{P}(X=x) = \sum_{x\in\mathcal{I}} xf(x).$$

With this definition of the expectation we are essentially describing the outcome of X that we will see if we average over a large number of repeated experiments.

Let's examine what the expectation of a random variable means with a couple of examples:

Example 19 (Roulette). Bet $1 on black. Let

$$X = \begin{cases} +1\$ & \text{if you win,} \\ -1\$ & \text{if you lose,} \end{cases} \qquad \text{and} \qquad f(x) = \begin{cases} \frac{18}{38} & \text{if } x = +1, \\ \frac{20}{38} & \text{if } x = -1. \end{cases}$$

We use the definition of the expectation of X to determine

$$\mathbb{E}[X] = (+1\$)\frac{18}{38} + (-1\$)\frac{20}{38} = -\frac{2}{38} = -0.0526\$,$$

so *in the long run* you should expect to lose about 5 cents for every dollar that you bet on black. △

Example 20 (The expected value of a die). Roll a die. Let X denote the outcome of the die. The expected value of X is given by

$$\mathbb{E}[X] = \sum_{n=1}^{6} n\frac{1}{6} = \frac{1}{6}\sum_{n=1}^{6} n = \frac{1}{6}\cdot\frac{6\cdot(6+1)}{2} = \frac{7}{2} = 3.5,$$

so the expected value of a die (as averaged over the long run) is 3.5. △

Observe that you cannot actually roll a value of 3.5 with a die, so the expected value of a die roll does not correspond to an actual outcome of a die roll. More generally, it holds for any random variable X that takes values on the set $\mathcal{I}$ that $\mathbb{E}[X]$ does not need to take its value on $\mathcal{I}$. In particular, $\mathbb{E}[X]$ is a function from $\mathcal{I}$ to $\mathbb{R}$, i.e., $\mathbb{E}[X] : \mathcal{I} \to \mathbb{R}$.

Example 21 (The expectation of a geometric r.v.). Let $X \sim \mathrm{Geo}(p)$, then, since

$$\frac{\mathrm{d}}{\mathrm{d}p}(1-p)^n = -n(1-p)^{n-1}, \qquad \text{and} \qquad \sum_{n=0}^{\infty} z^n = \frac{1}{1-z},$$

we can calculate

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=1}^{\infty} n\mathbb{P}(X=n) = \sum_{n=1}^{\infty} n(1-p)^{n-1}p \\ &= p\sum_{n=1}^{\infty} n(1-p)^{n-1} = p\frac{\mathrm{d}}{\mathrm{d}p}\sum_{n=0}^{\infty} -(1-p)^n \\ &= -p\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{1-(1-p)}\right) = -p\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{p}\right) \\ &= -p\left(-\frac{1}{p^2}\right) = \frac{1}{p},\end{aligned}$$

So $\mathbb{E}[X] = 1/p$ for a geometric random variable. To see that this makes sense think of the following question: how often should I expect to roll a die before I roll a 6? △

Example 22 (The expectation of a binomial r.v.). Let $X \sim \mathrm{Bin}(n,p)$, and recall that

$$f(k) = \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}.$$

Using moreover that $\frac{k}{k!} = \frac{1}{(k-1)!}$ and that $n-k = (n-1)-(k-1)$, we can calculate

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} kf(k) \\ &= 0 + \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{k}{k!}\frac{n(n-1)}{(n-k)!}pp^{k-1}(1-p)^{n-k} \\ &= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!}p^{k-1}(1-p)^{(n-1)-(k-1)} \\ &= np\sum_{s=0}^{n-1}\binom{n-1}{s}p^s(1-p)^{(n-1)-s} \\ &= np(p+(1-p))^{n-1} = np,\end{aligned}$$

where in the third-to-last step we relabelled the summation with $s = k-1$ and in the second-to-last step we used the binomial theorem. △

Sometimes the most useful random variable is the simplest one:

Example 23 (The indicator function). We say that the random variable $\mathbb{1}_A$ is the *indicator function* of the event A if

$$\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A^c \text{ occurs.} \end{cases}$$

Observe that the p.m.f. of $\mathbb{1}_A$ is given by $f(1) = \mathbb{P}(A)$ and $f(0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. It follows that the expectation of $\mathbb{1}_A$ is given by

$$\mathbb{E}[\mathbb{1}_A] = 1 \times \mathbb{P}(A) + 0 \times (1 - \mathbb{P}(A)) = \mathbb{P}(A),$$

that is, $\mathbb{1}_A$ lets us use the tricks and calculations that we will discover for expectations to probabilities. This will turn out to be incredibly useful. △

5.2. **The expectation of a function of a random variable.** Random variables often have a physical interpretation, and often we want to use that interpretation in further calculations. Consider for instance a machine that cuts circular slices of silica from a rod, to be used for the production of computer chips. The number of chips that we can put on a slice depends on the surface area of the disk. We measure the radius of the rod up to the nearest micron:[29] this is our random variable R. Because our measurement is in microns, R is a discrete random variable. We can use the standard formula for surface area of a circle to calculate the area of the disc A from our measurement of the radius R, i.e.,

$$A = \pi R^2.$$

It follows that A is also a discrete random variable. If we want to calculate $\mathbb{E}[A]$, it stands to reason that we should know the p.m.f. of A. What this p.m.f. is obviously depends on the p.m.f. of R. And if the p.m.f. of R is particularly difficult function, then it may be very difficult to find the p.m.f. of A. Fortunately, if we only care about $\mathbb{E}[A]$, we don't have to go through all the trouble of determining the p.m.f. of A. Instead, we can just use the p.m.f. of R and the function of A in terms of R to calculate $\mathbb{E}[A]$. The tool that we use to do this the following theorem:

THEOREM 2 [The law of the unconscious statistician]. *If X is a discrete random variable that takes values on $\mathcal{I}$ with p.m.f. $f(x)$, and if $g(x)$ is a real-valued function, then*

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{I}} g(x) f(x).$$

Before we prove the this theorem, let's see how it works in practice:

[29] A micron is a millionth of a meter.

Example 24. Roll a fair die, write X for the outcome. Let $g(x) = (x-3)^2$. Applying the law of the unconscious statistician, we can calculate

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{x=1}^{6} g(x)f(x) = \sum_{x=1}^{6} \frac{(x-3)^2}{6} \\ &= \frac{1}{6}\big((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 + (6-3)^2\big) \\ &= \frac{1}{6}(4+1+0+1+4+9) = \frac{19}{6}.\end{aligned}$$

Of course, X and $g(x)$ are not so difficult, so we can figure out the p.m.f. of $g(X)$ and calculate $\mathbb{E}[g(X)]$ in the traditional way also:

$$g(X) = \begin{cases} 0 & \text{when we roll 3,} \\ 1 & \text{when we roll 2 or 4,} \\ 4 & \text{when we roll 1 or 5,} \\ 9 & \text{when we roll 6 ,} \end{cases}$$

so it follows that

$$\mathbb{E}[g(x)] = 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} + 4 \cdot \frac{2}{6} + 9 \cdot \frac{1}{6} = \frac{19}{6}.$$

So at least the theorem works for this particular case. △

Now we prove the law of the unconscious statistician:

Proof of Theorem 2. Let $Y = g(X)$, then

$$\mathbb{E}[Y] = \sum_{y \in g(\mathcal{I})} y\,\mathbb{P}(Y = y),$$

(where we write $g(\mathcal{I})$ for the set $\{g(x) : x \in \mathcal{I}\}$), and

$$\begin{aligned}\mathbb{P}(Y = y) &= \mathbb{P}(X = \text{ a value } x \text{ such that } y = g(x)) \\ &= \sum_{x \in \mathcal{I} : g(x) = y} \mathbb{P}(X = x)\end{aligned}$$

(think of $Y = X^2$, then $\mathbb{P}(Y = y) = \mathbb{P}(X = \sqrt{y}) + \mathbb{P}(X = -\sqrt{y})$).

We can use the above equations to rearrange the sums

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in g(\mathcal{I})} y \sum_{x \in \mathcal{I} : g(x) = y} \mathbb{P}(X = x) \\ &= \sum_{y \in g(\mathcal{I})} \sum_{x \in \mathcal{I} : g(x) = y} g(x)\mathbb{P}(X = x).\end{aligned}$$

Now we use that the double sum effectively is a sum over $x \in \mathcal{I}$, i.e.,

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in g(\mathcal{I})} \sum_{x \in \mathcal{I} : g(x) = y} g(x)\mathbb{P}(X = x) \\ &= \sum_{x \in \mathcal{I}} g(x)\mathbb{P}(X = x),\end{aligned}$$

and that completes the proof. □

Example 25 (The moments of X). Let $g(x) = x^k$ for some integer k. We call $\mathbb{E}[X^k]$ the *kth moment of* X. It follows that

$$\begin{aligned}\text{1st moment:} &\qquad \mathbb{E}[X] = \sum_{x \in \mathcal{I}} x f(x) =: \mu, \\ \text{2nd moment:} &\qquad \mathbb{E}[X^2] = \sum_{x \in \mathcal{I}} x^2 f(x).\end{aligned}$$

(We often write μ (the Greek letter 'mu') for the 'mean of X'.) △

Using the first and second moment, we can calculate an important quantity:

Definition 11 [The variance of a random variable]. *Let X be a discrete random variable. We define the* variance of X *as*

$$\sigma^2 := \operatorname{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Moreover, we define the standard deviation of X *as* $\sigma = \sqrt{\operatorname{Var}(X)}$.

Roughly speaking, the standard deviation σ (the Greek letter 'sigma') of a random variable X describes the size of a typical fluctuation in X, i.e.,

- when σ is small compared to μ, then X is typically close to μ,
- when σ is large compared to μ, then X is typically far from μ.

We can give a quantitative statement of this observation:

Principle 20 [Chebychev's inequality]. *Given a random variable X that takes values on $\mathcal{I}$ and any $\varepsilon > 0$,*

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. By the definition of σ^2,

$$\begin{aligned}\sigma^2 = \mathrm{Var}(X) &= \sum_{x\in\mathcal{I}}(x-\mu)^2 f(x) \\ &\geq \sum_{x\in\mathcal{I}\,:\,(x-\mu)^2\geq\varepsilon}(x-\mu)^2 f(x) && \text{[sum over fewer values]} \\ &\geq \sum_{x\in\mathcal{I}\,:\,(x-\mu)^2\geq\varepsilon}\varepsilon^2 f(x) && \text{[use that } (x-\mu)\geq\varepsilon\text{]} \\ &= \varepsilon^2\sum_{x\in\mathcal{I}\,:\,(x-\mu)^2\geq\varepsilon} f(x) && \text{[rearrange]} \\ &= \varepsilon^2\mathbb{P}((X-\mu)^2\geq\varepsilon^2) && \text{[sum over } x\text{]} \\ &= \varepsilon^2\mathbb{P}(|X-\mu|\geq\varepsilon),\end{aligned}$$

so $\varepsilon^2\mathbb{P}(|X-\mu|\geq\varepsilon)\leq\sigma^2$. Rearranging completes the proof. □

QUESTION 33. *Acme Widgets produces an average of 1000 widgets a day with a standard deviation of 20 widgets. Find an interval where you can be 90% sure that the daily production will lie in this interval.*

Solution. Write X for the number of widgets produced on a given day. We don't know the distribution of X, but we do know that $\mu = 1000$ and $\sigma = 20$. We can restate the question as follows: Find a number ε so that on 9 days out of 10, the number of widgets produced is not less than $1000 - \varepsilon$ and not more than $1000 + \varepsilon$. That is, we want to calculate the value of ε such that $\mathbb{P}(|X-\mu|\geq\varepsilon)\leq 0.1$. We can use Chebychev's inequality:

$$0.1 = \mathbb{P}(|X-\mu|\geq\varepsilon)\leq\frac{\sigma^2}{\varepsilon^2}\leq\frac{400}{\varepsilon^2}.$$

It follows that we should take $\varepsilon^2 = 4000$, and therefore $\varepsilon = 63.2$. We can thus be (more than) 90% confident that the daily production of widgets will be within 936 and 1064 widgets. ◊

Relating the standard deviation to fluctuations of the random variable. Chebychev's inequality tells us that

$$\mathbb{P}(|X-\mu|\leq\varepsilon)\geq 1-\frac{\sigma^2}{\varepsilon^2},$$

so if we take $\varepsilon = k\sigma$, then

$$\mathbb{P}(|X-\mu|\leq k\sigma)\geq 1-\frac{\sigma^2}{k^2\sigma^2} = 1-\frac{1}{k^2}.$$

Setting $k = 2$ gives

$$\mathbb{P}(|X-\mu|\leq 2\sigma)\geq 1-\frac{1}{2^2} = 0.75,$$

so X is within two standard deviations of μ at least 75% of the time.

5.3. **Properties of expectation and variance.** We can determine the following useful properties of the expectation of a random variable:

Principle 21 [Properties of expectation]. *Let X be a random variable that takes its values on the set $\mathcal{I}$.*

(a) *Linearity of expectation: for real numbers a, b,*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

(b) *If X and Y are random variables:*

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

We will prove (b) later.

Proof of (a). By Theorem 2 and Principle 19(b),

$$\mathbb{E}[aX + b] = \sum_{x\in\mathcal{I}}(ax + b)f(x) = a\sum_{x\in\mathcal{I}} xf(x) + b\sum_{x\in\mathcal{I}} f(x) = a\mathbb{E}[X] + b\cdot 1. \qquad \square$$

We can determine the following properties for the variance:

Principle 22 [Properties of variance]. *Let X be a random variable.*

(a) *Quadratic behaviour of variance: for real numbers a, b,*

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

(b) *Alternative formula for the variance:*

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

(c) *If X and Y are independent random variables,*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Keep in mind that once again, property (c) *only* holds (in generality) if X and Y are independent! We will prove property (c) later.

Proof of (a). By the definition of the variance and Principle 21(a),

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

Proof of (b). Write $\mu = \mathbb{E}[X]$, then, by the definition of the variance and Principle 21(a)

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2]\\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2\\ &= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \qquad \square\end{aligned}$$

Example 26 (The variance of a Bernoulli r.v.). Recall that

$$X = \begin{cases} 1 & \text{with prob. } p, \\ 0 & \text{with prob. } 1 - p, \end{cases}$$

so

$$\mathbb{E}[X] = 0 \cdot (1 - p) + 1 \cdot p = p,$$

and

$$\begin{aligned} \operatorname{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - p^2 \\ &= (0^2 \cdot (1 - p) + 1^2 \cdot p) - p^2 \\ &= p - p^2 = p(1 - p). \end{aligned}$$

△

Example 27 (The variance of a binomial r.v.). Let $X \sim \operatorname{Bin}(n, p)$. We have already calculated that $\mathbb{E}[X] = np$. The variance is given by

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - n^2 p^2.$$

We can try and calculate the second moment of X, $\mathbb{E}[X^2]$, directly using

$$\mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1 - p)^{n-k},$$

but this is not an easy calculation.[30]

The calculation is much easier than that. Recall that the Binomial random variable X counts the number of successes in n independent trials, where the success probability for each trial is p. Therefore, another way of looking at X is as a sum of *n independent* Bernoulli random variables. That is, let $Y_i \sim \operatorname{Ber}(p)$ for $i = 1, \dots, n$, then $X = Y_1 + \cdots + Y_n$, and we can use Principle 22(c) to determine

$$\begin{aligned} \operatorname{Var}(X) &= \operatorname{Var}(Y_1 + \cdots + Y_n) \\ &= \operatorname{Var}(Y_1) + \cdots + \operatorname{Var}(Y_n) \\ &= p(1 - p) + \cdots + p(1 - p) \\ &= np(1 - p). \end{aligned}$$

△

5.4. **The Poisson random variable describes rare events.** The Vancouver metro area is home to some 2.25 million people. On any given day, none of these people are very likely to call 911, but some do. Suppose we want to describe the number of people that call 911 in Vancouver on a given day. If we assume that everybody is equally probable to do so (say with probability 0.0001), and that calls are more-or-less independent, then we can describe this number as a Binomial random variable $X \sim \operatorname{Bin}(n = 2.25 \cdot 10^6, p = 0.0001)$.

[30] Ross, in Section 4.6.1 takes this approach.

The expected number of people that calls 911 on a given day is $\mathbb{E}[X] = np = 225$. The variance is very close to the expectation: $\mathrm{Var}(X) = np(1-p) = 224.99$.

Because calls are rare, it does not matter that we don't know the precise number of people in Vancouver. Likewise, since the number of people in Vancouver is very large, it does not matter that we don't know the precise probability that someone calls 911. Maybe it is therefore enough if we just know the mean number of calls to 911 per day in Vancouver.

As it turns out, it is enough to know the mean. There exists a very good approximation for binomial random variables with n large, p small, and np moderate. This approximation is known as the *Poisson random variable.*[31] Because scenarios that look like rare events that can be described by binomials with large n, small p, and moderate np are common, the Poisson random variable takes a prominent place in probability theory.

Before we derive the approximation, let's see what a Poisson random variable looks like:

Example 28 (The Poisson random variable). We say that a random variable X that takes values on $0, 1, 2, \dots$ is a Poisson r.v. with parameter λ (the Greek letter 'lambda') if, for some $\lambda > 0$,

$$\mathbb{P}(X = n) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}, \qquad \text{for } n = 0, 1, 2, \dots$$

We write $X \sim \mathrm{Poi}(\lambda)$.

The above probabilities satisfy the properties of a p.m.f. as given by Principle 19:

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

(here we used the definition of e^x). △

Approximation of a binomial by a Poisson random variable. Suppose that $X \sim \mathrm{Bin}(n, p)$ with n large, p small, and np moderate (say more than 5). We will write $\lambda = np$ (in anticipation of the formula of the Poisson r.v. that will come out at the end of our calculation). Then it follows that $p = \frac{\lambda}{n}$, and we can calculate

$$\begin{aligned}
\mathbb{P}(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \left[\frac{\lambda^k}{k!}\right] \times \left[\left(1 - \frac{\lambda}{n}\right)^n\right] \times \left[\frac{1}{\left(1 - \frac{\lambda}{n}\right)^k} \frac{1}{n^k} \frac{n!}{(n-k)!}\right].
\end{aligned}$$

We will leave the first term as it is, and approximate the other two terms separately, starting with the third term: We will show that the third term is close to 1 when n is large. Start by

[31] After Siméon Denis Poisson, 1781–1840.

observing that

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

so

$$\begin{aligned}
\frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\frac{1}{n^k}\frac{n!}{(n-k)!} &= \frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\cdots\cdot\frac{(n-k+1)}{n} \\
&= \frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\cdot 1\cdot\left(1-\frac{1}{n}\right)\cdot\cdots\cdot\left(1-\frac{k-1}{n}\right) \\
&= \frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\prod_{r=0}^{k-1}\left(1-\frac{r}{n}\right) \\
&= \prod_{r=0}^{k-1}\frac{1-\frac{r}{n}}{1-\frac{\lambda}{n}}.
\end{aligned}$$

Now we use the fact that $x = \mathrm{e}^{\log(x)}$ and that $\log(xy) = \log(x) + \log(y)$, and proceed[32]

$$\begin{aligned}
\prod_{r=0}^{k-1}\frac{1-\frac{r}{n}}{1-\frac{\lambda}{n}} &= \mathrm{e}^{\log\left(\prod_{r=0}^{k-1}\frac{1-\frac{r}{n}}{1-\frac{\lambda}{n}}\right)} \\
&= \mathrm{e}^{\sum_{r=0}^{k-1}\log\left(1-\frac{r}{n}\right)-\sum_{r=0}^{k-1}\log\left(1-\frac{\lambda}{n}\right)} \\
&\approx \mathrm{e}^{-\sum_{r=0}^{k-1}\frac{r}{n}+\sum_{r=0}^{k-1}\frac{\lambda}{n}} \\
&= \mathrm{e}^{-\frac{k(k-1)}{n}+\frac{\lambda k}{n}} \\
&\approx \mathrm{e}^{0} = 1,
\end{aligned}$$

where, for the first approximation we again used that $\log(1-x) \approx -x$ when x is small. In the last step we used the approximation that $\frac{k(k-1)}{n} \approx 0$ and that $\frac{\lambda k}{n} \approx 0$, and so determined that the third term on the right-hand side of (5.4) is close to 1 when n is large.

Now we have to bound the second term on the right-hand side of (5.4). We will use that

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

To see this, we again take the exponent of logarithm and approximate:

$$\left(1-\frac{\lambda}{n}\right)^n = \mathrm{e}^{n\log\left(1-\frac{\lambda}{n}\right)} \approx \mathrm{e}^{-n\frac{\lambda}{n}} = \mathrm{e}^{-\lambda}.$$

From the analysis of the second and third terms of (5.4), it follows that when n is large,

$$\mathbb{P}(X=k) \approx \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},$$

which is the p.m.f. of a Poisson random variable with parameter λ.

[32] We will essentially use the same analysis as in the treatment of the birthday paradox, question 16.

Example 29 (The mean and variance of a Poisson r.v.). Let $X \sim \mathrm{Poi}(\lambda)$. Then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=0}^{\infty} n\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\\ &= \lambda\mathrm{e}^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{(n-1)!}\\ &= \lambda\mathrm{e}^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}\\ &= \lambda\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = \lambda,\end{aligned}$$

so $\mathbb{E}[X] = \lambda$.

Now we calculate the variance:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \lambda^2.$$

We write

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

Calculating $\mathbb{E}[X(X-1)]$ is similar to calculating $\mathbb{E}[X]$:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{n=0}^{\infty} n(n-1)\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\\ &= \lambda^2\mathrm{e}^{-\lambda}\sum_{n=2}^{\infty}\frac{\lambda^{n-2}}{(n-2)!}\\ &= \lambda^2\mathrm{e}^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}\\ &= \lambda^2\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = \lambda^2.\end{aligned}$$

So we can conclude that

$$\mathrm{Var}(X) = \mathbb{E}[X(X-1)] + \lambda - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda,$$

so $\mathrm{Var}(X) = \lambda$ also. △

Some examples of Poisson random variables.

Example 30 (Radioactive decay). Uranium 238 has a half-life of approximately 4 billion years ($\approx 10^{17}$ seconds). That is, if we have a single Uranium-238 atom, we have to wait approximately 4 billion years to have a 50-50 chance of seeing it decay. There are about 2.5×10^{18} atoms in a milligram of Uranium 238. Therefore, we are looking at a situation with a large number of rare events, so we can model the observed number of decays of a 1 milligram sample of Uranium 238 in a second by a Poisson random variable. To determine the parameter λ we can measure the average number of decays per second (we don't necessarily need to know the number of atoms or the decay probability for this). Suppose that it turns out that $\lambda = 1.4$. We can calculate the probability that we see at most 3 decays in a second:

$$\begin{aligned}\mathbb{P}(\text{at most 3 decays observed}) &= \mathbb{P}(X \le 3)\\ &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3)\\ &= \mathrm{e}^{-1.4}\left(\frac{1.4^0}{0!} + \frac{1.4^1}{1!} + \frac{1.4^2}{2!} + \frac{1.4^3}{3!}\right)\\ &= 0.946.\end{aligned}$$

△

Example 31 (Some more examples of things that can be described with a Poisson r.v.).

(a) The number of typos on a page of text.
(b) The number of flaws in a length of wire.
(c) The number of customers at a restaurant on a given day.
(d) The number of jumps in a stock price during a day.
(e) The number of photons arriving at a telescope.
(f) The number of shark attacks per year.
(g) The number of earthquakes per year.
(h) The number of mutations on a chromosome.

△

Example 32 (The birthday paradox revisited – *Not treated in class*). In question 16 we calculated the probability that in a room with n people, no two people share the same birthday. If n is large (but not too large) then the number of people who share a birthday is almost a Poisson random variable: We can write X for the random variable for the number of pairs of people with the same birthday. In total there are $\binom{n}{2}$ pairs of such people, so we can view X as the sum of $\binom{n}{2}$ success-failure trials, where the trial is "pick 2 persons and check if their birthdays are the same". Write $E_{i,j}$ for the event that person i and person j have the same birthday. Observe that these events are *not* independent, because if for instance $E_{1,2}$ happens, and $E_{2,3}$ happens, then $E_{1,3}$ must also happen. But when n is large, the trials will be 'almost' independent, so we can do an (uncontrolled) approximation of a Poisson r.v. with

$$\lambda = (\#\text{ trials}) \times (\text{success prob.}) = \binom{n}{2} \cdot \frac{1}{365}.$$

We can then calculate

$$\mathbb{P}(\text{no match}) = \mathbb{P}(X = 0) = \frac{\lambda^0}{0!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\binom{n}{2}\frac{1}{365}} = \mathrm{e}^{-\frac{n(n-1)}{730}},$$

which is the same approximation as we get in question 16.

We can moreover approximate the probability that we get exactly one match:

$$\mathbb{P}(\text{exactly one match}) \approx \mathbb{P}(X = 1) = \mathrm{e}^{-\lambda}\frac{\lambda}{1!} = \binom{n}{2}\frac{1}{365}\mathrm{e}^{-\binom{n}{2}\frac{1}{365}}.$$

We can even approximate the probability of a triple match by letting the random variable Y describe the number of triple matches, and noting that there are now $\binom{n}{3}$ trials with success probability $\left(\frac{1}{365}\right)^2$, i.e.,

$$\mathbb{P}(\text{no triple match}) \approx \mathbb{P}(Y = 0) = e^{-\binom{n}{3}\left(\frac{1}{365}\right)^2}.$$

(If we choose $n = 90$, we get $\mathbb{P}(Y = 0) \approx 0.41$.) △

Finally, some sketches of four common discrete distributions:

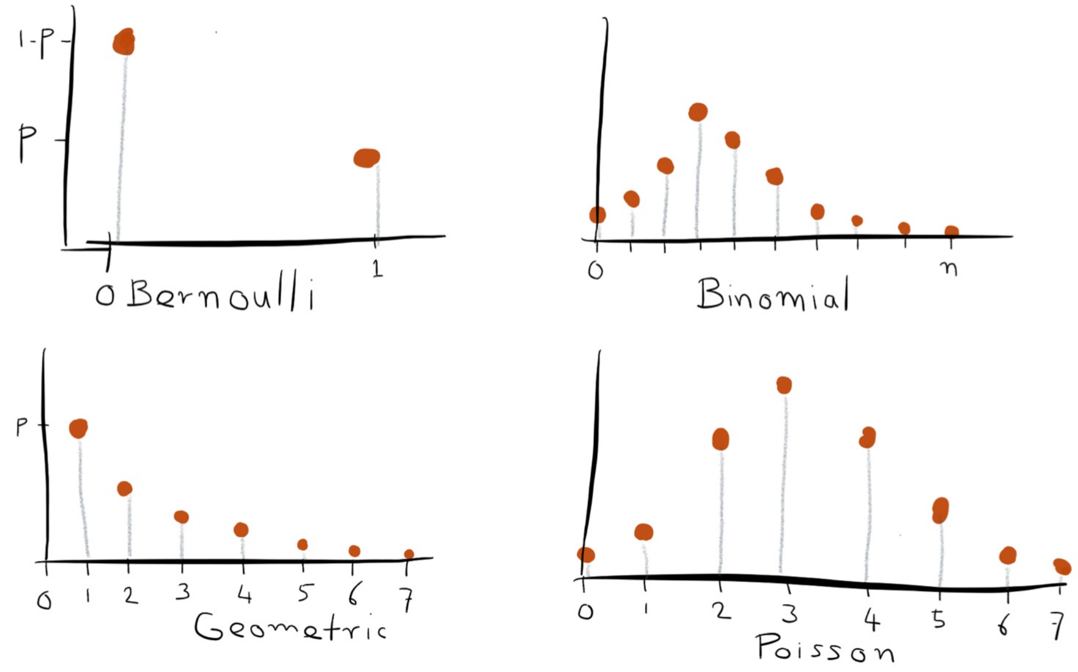

FIGURE 7. Some sketches of typical shapes of the p.m.f.'s of four common discrete random variables.

6. Continuous random variables

So far, we have only discussed random variables that take their values on discrete sets. Such random variables are useful when we want to count whole things, such as the number of successes, or when a measurement is rounded off at a given decimal. But sometimes, discrete random variables do not give a satisfactory description, for instance when we are interested in quantities like the time until a certain event happens, or in the outcome of a very precise measurement. For such cases we can expand our definition of a random variable in the following way:

Definition 12 [Continuous random variables]. *We say that X is a* continuous random variable *if there exists a non-negative function f on the set of real numbers (i.e., the set of all x such that $x \in (-\infty, \infty)$) that has the property that for any set B of real numbers,*

$$\mathbb{P}(X \in B) = \int_B f(x) \mathrm{d}x.$$

We call the function f the probability density function *(or p.d.f.) of the random variable X.*

The essential difference between the definitions of discrete and continuous random variables is that we use summation to evaluate discrete random variables, and we use integration to evaluate continuous random variables. This makes discrete and continuous r.v.'s somewhat different, but on the whole the two are very similar.

Principle 23 [Properties of continuous random variables]. *Let X be a continuous random variable with p.d.f. f, then X satisfies the following properties:*

(a) *The random variable X must take some value in the reals:*

$$1 = \mathbb{P}(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x) \mathrm{d}x.$$

(b) *For sets that are intervals, i.e., $B = [a, b]$,*

$$\mathbb{P}(X \in B) = \mathbb{P}(a \le X \le b) = \int_a^b f(x) \mathrm{d}x.$$

(c) *Isolated points have zero probability:*

$$\mathbb{P}(X = a) = 0,$$

and moreover,

$$\mathbb{P}(X \le a) = \mathbb{P}(X < a).$$

From property (c) it is clear that a probability density function does *not* describe the probability of seeing a single value. Rather, $f(x)$ describes how likely it is that X takes a value *near* x.

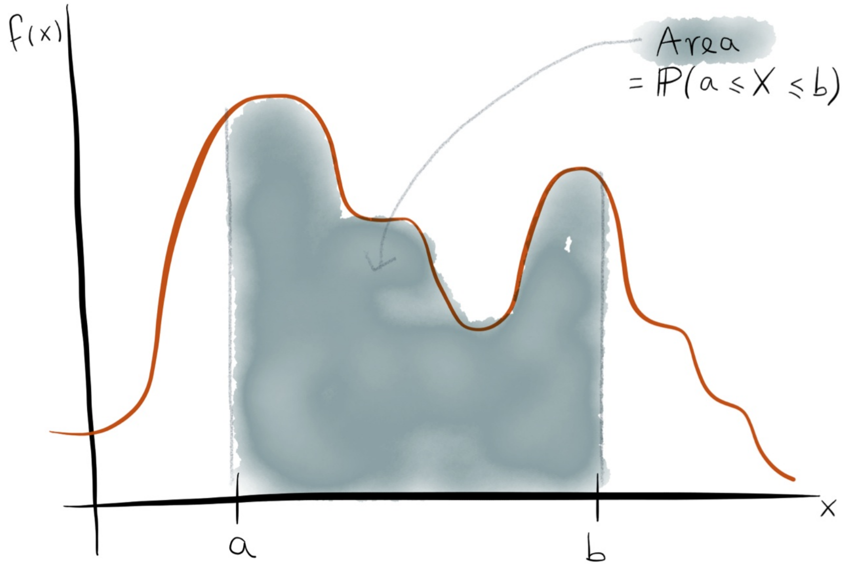

FIGURE 8. A schematic drawing of the $\mathbb{P}(a \leq X \leq b)$ in terms of the area between a and b underneath $f(x)$.

Proof. (a) and (b) follow directly from the definition of a continuous random variable. (See figure 8 for a sketch of property (b).)The proof of (c) is elementary: let F be the antiderivative[33] (a.k.a. primitive) of f, then

$$\mathbb{P}(X = a) = \mathbb{P}(X \in [a, a]) = \int_a^a f(x)\mathrm{d}x = F(a) - F(a) = 0$$

and so

$$\mathbb{P}(X \leq a) = \mathbb{P}(X < a) + \mathbb{P}(X = a) = \mathbb{P}(X < a) + 0.$$

□

Since we calculate probabilities by taking the integral of a p.d.f. f, it makes sense to study the antiderivative as well. It turns out that the antiderivative of a p.d.f. has a clear probabilistic interpretation:

DEFINITION 13 [The cumulative distribution function]. *Given a continuous random variable X with p.d.f. f, we define the* cumulative distribution function *(or c.d.f.) F of X as the function*

$$F(b) = \mathbb{P}(X \leq b) = \int_{-\infty}^{b} f(x)\mathrm{d}x.$$

[33] I.e., the function F such that $\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$.

Note that we can also define an analogous function for discrete random variables.

Principle 24 [Properties of the c.d.f.]. *Let X be a continuous random variable with c.d.f. F, then F has the following properties:*

(a) $F(-\infty) = 0$ *and* $F(\infty) = 1$.
(b) $F(x)$ *is non-decreasing in* x.
(c) $\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$.
(d) $\frac{\mathrm{d}}{\mathrm{d}b}F(b) = f(b)$.

Proof. (a) $F(-\infty) = \mathbb{P}(X \leq -\infty) = 0$ since X only takes values in $(-\infty, \infty)$. Similarly, $F(\infty) = \mathbb{P}(X < \infty) = 1$ holds.

(b) Let a and b be real numbers such that $a < b$, then

$$F(a) = \mathbb{P}(X \leq a) \leq \mathbb{P}(X \leq b) = F(b).$$

(c) By mutual exclusivity,

$$\mathbb{P}(X \leq b) = \mathbb{P}(\{X < a\} \cup \{a \leq X \leq b\}) = \mathbb{P}(X \leq a) + \mathbb{P}(a \leq X \leq b),$$

so

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a).$$

(d) By the definition of $F(b)$,

$$\frac{\mathrm{d}}{\mathrm{d}b}F(b) = \frac{\mathrm{d}}{\mathrm{d}b}\int_{-\infty}^{b} f(x)\mathrm{d}x = f(b).$$

□

Let's look at some examples of continuous random variables:

Example 33. Let X be a continuous r.v. with p.d.f.

$$f(x) = \begin{cases} \mathrm{e}^{-x} & \text{when } x \geq 0, \\ 0 & \text{when } x < 0. \end{cases}$$

We can check that this is indeed a p.d.f.:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_{-\infty}^{0} 0\,\mathrm{d}x + \int_{0}^{\infty} \mathrm{e}^{-x}\mathrm{d}x = 0 + [-\mathrm{e}^{-x}|_0^{\infty} = -(0-1) = 1.$$

We can also calculate the c.d.f. of X: if $b \leq 0$, then $F(b) = 0$, and if $b > 0$, then

$$F(b) = \int_{-\infty}^{b} f(x)\mathrm{d}x = 0 + \int_{0}^{b} \mathrm{e}^{-x}\mathrm{d}x = [-\mathrm{e}^{-x}|_0^{b} = 1 - \mathrm{e}^{-b}.$$

△

Example 34. Let X be a continuous r.v. with p.d.f.

$$f(x) = \begin{cases} 2x & \text{when } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We can again check that this is a p.d.f.:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_0^1 2x\mathrm{d}x = \left[2 \cdot \frac{x^2}{2}\right|_0^1 = 2\left(\frac{1}{2} - 0\right) = 1.$$

We can calculate the probability that $X > 1/2$:

$$\mathbb{P}(X \ge 1/2) = \int_{1/2}^{1} 2x\mathrm{d}x = 2\left[\frac{x^2}{2}\right|_{1/2}^{1} = 2\left(\frac{1}{2} - \frac{(1/2)^2}{2}\right) = 1 - \left(\frac{1}{2}\right)^2 = \frac{3}{4}.$$

We can calculate the probability that $X > 1/2$ given the information that we know that $X > 1/4$:

$$\mathbb{P}(X > 1/2|X > 1/4) = \frac{\mathbb{P}(X > 1/2)}{\mathbb{P}(X > 1/4)} = \frac{3/4}{\int_{1/4}^{1} 2x\mathrm{d}x} = \frac{3/4}{1 - (1/4)^2} = \frac{3/4}{15/16} = \frac{12}{15} = \frac{4}{5}.$$

We can calculate the conditioning the other way around too (although this is pretty obvious):

$$\mathbb{P}(X > 1/4|X > 1/2) = \frac{\mathbb{P}(X > 1/2)}{\mathbb{P}(X > 1/2)} = 1.$$

Finally, we can calculate the c.d.f.: For $x < 0, F(x) = 0$. For $0 < x < 1$,

$$F(x) = \int_0^x 2t\mathrm{d}t = x^2.$$

And for $x > 1$, $F(x) = 1$. △

We can calculate the expectation of a continuous random variable, just like a discrete random variable, but with summation replaced with integration:

Definition 14 [The expectation of a continuous random variable]. *For a continuous random variable X with p.d.f. f, we define the expectation of X as*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x.$$

Example 35. Let X be a continuous r.v. with p.d.f.

$$f(x) = \begin{cases} 2x & \text{when } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

then we can calculate

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_{0}^{1} 2x^2 = \left[2\frac{x^3}{3}\right|_0^1 = \frac{2}{3}.$$

△

PRINCIPLE 25 [Properties of continuous r.v.'s]. *Let X be a continuous random variable with p.d.f. f, then the following properties that we determined for discrete random variables also hold for X:*

(a) *The law of the unconscious statistician for continuous r.v.'s: for any function $g(x)$,*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x.$$

(b) *The variance of a continuous random variable:*

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x = \mathbb{E}[X^2] - \mu^2.$$

(c) *Linearity of expectation, quadratic behaviour of variance:*

$$\mathbb{E}[aX+b] = a\mathbb{E}[X] + b \qquad \textit{and} \qquad \mathrm{Var}(aX+b) = a^2\mathrm{Var}(X).$$

(d) *Chebychev's inequality:*

$$\mathbb{P}(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

It is often useful to think of $\mathbb{E}[X]$ as the x-coordinate of the 'centre of mass' of the shape between the graph of $f(x)$ and the x-axis.

Proof. For (a), (b), and (c), see Ross, Section 5.2. For proof of (d), see Ross Section 8.2.

6.1. **Examples of continuous random variables.** In this section we discuss some important examples of continuous random variables.

6.1.1. *The uniform random variable.* A random variable X is said to be *uniformly distributed* over the interval $[a,b]$ if the p.d.f. of X is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{when } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

We write $X \sim \mathrm{Unif}[a,b]$.

Checking that X is normalised is easy: the shape of $f(x)$ is a rectangle with sides $(b-a)$ and $\frac{1}{b-a}$, so the area underneath $f(x)$ is simply the area of the rectangle:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = (b-a) \times \frac{1}{b-a} = 1.$$

It also follows that for $a \le c < d \le b$,

$$\mathbb{P}(c \le X \le d) = \int_c^d \frac{1}{b-a}\mathrm{d}x = \frac{d-c}{b-a}.$$

If we think of $\mathbb{E}[X]$ as the x-coordinate of the centre of mass of $f(x)$, then since $f(x)$ has a rectangular shape, $\mathbb{E}[X]$ should lie in the middle of the interval $[a,b]$, so $\mathbb{E}[X] = (a+b)/2$. We can also just calculate the mean of X:

$$\mathbb{E}[X] = \int_a^b x\frac{1}{b-a}\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^2}{2}\right|_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{1}{2}\frac{(b-a)(b+a)}{b-a} = \frac{a+b}{2}.$$

And we can calculate the variance of X: direct computation works fine, but it is a bit tricky to get a nice equation that way.[34] A better way of calculating the variance of X is by introducing the random variable

$$Y = \frac{X-\mu}{(b-a)}.$$

The random variable Y is a 'shifted and squeezed' version of X, where the shift and squeeze are chosen such that the p.d.f. of Y is simply

$$g(x) = \begin{cases} 1 & \text{when } -\frac{1}{2} \le x \le \frac{1}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

so $Y \sim \text{Unif}[-1/2, 1/2]$. As a result, $\mathbb{E}[Y] = \frac{-1/2+1/2}{2} = 0$. Since $X = (b-a)Y + \mu$, it follows from the quadratic behaviour of the variance that

$$\text{Var}(X) = (b-a)^2\text{Var}(Y).$$

[34] I.e.,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \frac{a^2+2ab+b^2}{4}.$$

By the law of the unconscious statistician,

$$\mathbb{E}[X^2] = \int_a^b x^2\frac{1}{b-a}\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^3}{3}\right|_a^b = \frac{1}{b-a}\frac{b^3-a^3}{3}.$$

We can simplify this a bit more, since $b^3 - a^3 = (b-a)(b^2+ab+a^2)$, so we get

$$\mathbb{E}[X^2] = \frac{b^2+ab+a^2}{3}.$$

Plugging this into the formula for the variance we get

$$\text{Var}(X) = \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{a^2-ab+b^2}{12} = \frac{(b-a)^2}{12}.$$

Moreover, $\mathrm{Var}(Y)$ is easy to calculate:

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \int_{-1/2}^{1/2} x^2 \mathrm{d}x - 0^2 = \left[\frac{x^3}{3}\right|_{-1/2}^{1/2} = \left(\frac{1}{3\cdot 2^3} - \left(-\frac{1}{3\cdot 2^3}\right)\right) = \frac{2}{3\cdot 8} = \frac{1}{12}.$$

So it follows that

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

Question 34. *The failure of a component interrupts work at an assembly line. The time X (in hours) until the replacement component is installed is distributed as $X \sim$ Unif$[1,5]$. Find the probability that replacement takes more than two hours.*

Solution. The time X has p.d.f.

$$f(x) = \begin{cases} \frac{1}{5-1} = \frac{1}{4} & \text{when } 1 \le x \le 5, \\ 0 & \text{otherwise.} \end{cases}$$

We are asked to calculate $\mathbb{P}(X \ge 2)$:

$$\mathbb{P}(X \ge 2) = \mathbb{P}(X \in [2,\infty)) = \int_2^\infty f(x)\mathrm{d}x = \int_2^5 \frac{1}{4}\mathrm{d}x = \left[\frac{x}{4}\right|_2^5 = \frac{5}{4} - \frac{2}{4} = \frac{3}{4}.$$

◊

6.1.2. *The exponential random variable.* A random variable is said to be *exponentially distributed* with parameter λ if X has p.d.f.

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{when } x \ge 0, \\ 0 & \text{when } x < 0. \end{cases}$$

We write $X \sim \mathrm{Exp}(\lambda)$. See figure **??** for some sketches of p.d.f.'s for exponential random variables.

We can check that X is properly normalised:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_0^\infty \lambda \mathrm{e}^{-\lambda x}\mathrm{d}x = \left[-\mathrm{e}^{-\lambda x}\right|_0^\infty = -0 - (-1) = 1.$$

We can calculate the c.d.f. of X: for $x < 0$ it is easy to see that $F(x) = 0$ and for $x \ge 0$,

$$F(x) = \mathbb{P}(X \le x) = \int_0^x \lambda \mathrm{e}^{-\lambda t}\mathrm{d}t = \left[-\mathrm{e}^{-\lambda t}\right|_0^x = 1 - \mathrm{e}^{-\lambda x}.$$

We can calculate the mean of X:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \int_0^\infty x\lambda \mathrm{e}^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda}\int_0^\infty y\mathrm{e}^{-y}\mathrm{d}y,$$

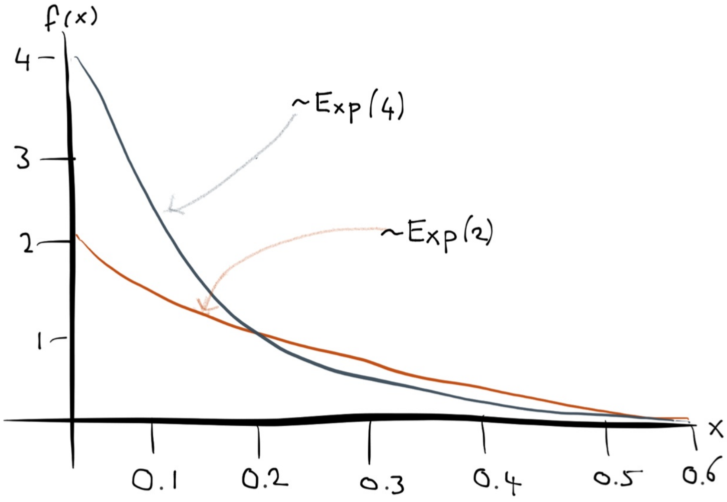

FIGURE 9. A schematic drawing of the p.d.f.'s of an Exp(2) and an Exp(2) random variable.

where in the last step we substituted[35] $y = \lambda x$.

Rather than solving just this integral, we solve the more general case, which is known as the *Gamma function:*

$$\Gamma(n) = \int_0^\infty y^{n-1} \mathrm{e}^{-y} \mathrm{d}y.$$

We use integration by parts[36] with $u(y) = y^{n-1}$ and $\frac{\mathrm{d}}{\mathrm{d}y} v(y) = \mathrm{e}^{-y}$:

$$\int_0^\infty y^{n-1}\mathrm{e}^{-y}\mathrm{d}y = \left[-y^{n-1}\mathrm{e}^{-y}\right|_0^\infty - \int_0^\infty (n-1) y^{n-2}\left(-\mathrm{e}^{-y}\right)\mathrm{d}y = 0 + (n-1)\int_0^\infty y^{n-2}\mathrm{e}^{-y}\mathrm{d}y = (n-1)\Gamma(n-1).$$

[35] Recall the substitution rule for integrals from calculus: for a continuously differentiable function $g(t)$, we can substitute $x = g(t)$

$$\int_{g(a)}^{g(b)} f(x)\mathrm{d}x = \int_a^b f(g(x))\left(\frac{\mathrm{d}}{\mathrm{d}x} g(x)\right)\mathrm{d}x$$

so if $y = \lambda x$ then $\mathrm{d}y = \lambda \mathrm{d}x$ and the integration limits become $0/\lambda = 0$ and $\infty/\lambda = \infty$.

[36] Recall that integration by parts is a technique from calculus that uses the fact that

$$\int_a^b u(y)\left(\frac{\mathrm{d}}{\mathrm{d}y} v(y)\right)\mathrm{d}y = [u(y)v(y)|_a^b - \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}y} u(y)\right) v(y)\mathrm{d}y.$$

So in particular, we have the equation $\Gamma(n) = (n-1)\Gamma(n-1)$. Observe that

$$\Gamma(1) = \int_0^\infty \mathrm{e}^{-y}\mathrm{d}y = \left[-\mathrm{e}^{-y}\right|_0^\infty = 0 - (-1) = 1,$$

It follows that

$$\begin{aligned}
\Gamma(2) &= 1 \cdot \Gamma(1) = 1,\\
\Gamma(3) &= 2 \cdot \Gamma(2) = 2 \cdot 1 = 2!,\\
\Gamma(4) &= 3 \cdot \Gamma(3) = 3 \cdot 2 \cdot 1 = 3!,\\
&\vdots\\
\Gamma(n) &= (n-1) \cdot (n-2) \cdots 2 \cdot 1 = (n-1)!,
\end{aligned}$$

so the Gamma function, when evaluated at the integers gives the factorials.

Applying this property of the Gamma function to our current problem, we get

$$\mathbb{E}[X] = \frac{1}{\lambda}\Gamma(2) = \frac{1}{\lambda}.$$

Using the Gamma function we can also easily calculate the nth moment of X:

$$\mathbb{E}[X^n] = \int_0^\infty x^n \lambda \mathrm{e}^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda^n}\Gamma(n+1) = \frac{n!}{\lambda}.$$

Now we can calculate the variance of X:

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

Example 36 (Lack of memory of the exponential distribution). Exponential random variables can be viewed as the continuous analogue of geometric random variables. In particular, the exponential random variable has the *lack-of-memory property:* Let $X \sim \operatorname{Exp}(\lambda)$, then

$$\begin{aligned}
\mathbb{P}(X > s+t | X > s) &= \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{1 - \mathbb{P}(X \le s+t)}{1 - \mathbb{P}(X \le s)}\\
&= \frac{1 - F(s+t)}{1 - F(s)} = \frac{1 - (1 - \mathrm{e}^{-\lambda(s+t)})}{1 - (1 - \mathrm{e}^{-\lambda s})}\\
&= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = 1 - F(t)\\
&= \mathbb{P}(X > t).
\end{aligned}$$

We can moreover show that *any* continuous random variable that has the lack-of-memory property must have an exponential distribution.

Finally, a warning. A common mistake is to take the lack-of-memory property of Z to mean that

$$\mathbb{P}(Z > s | Z > t) = \mathbb{P}(Z > s).$$

This is not the case! If the above equality holds, that means that $\{Z > s\}$ and $\{Z > t\}$ are *independent*, which is entirely different. △

QUESTION 35. *It is a curious fact is that the lifetimes of lobsters are exponentially distributed random variables with a mean lifetime (in captivity) of 60 years. A zoo has two lobsters, one is 1 year old, the other is 100 years old. What is the probability that the 1 year old lobster dies before the 100 year old lobster dies?*

Solution. Write X_1 for the lifetime of the young lobster, and X_2 for the lifetime of the old lobster. Since the lifetimes are exponentially distributed, $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \frac{1}{\lambda}$ so $\lambda = \frac{1}{60}$. We are asked to calculate

$$\mathbb{P}(X_1 - 1 > X_2 - 100 | \{X_1 > 1\} \cap \{X_2 > 100\}).$$

This is a bit tricky, it is easier to calculate the probability that either lobster survives for t more years and compare these probabilities: by the lack of memory of exponential random variables, the probability that the 1 year old lobster lives another t years is

$$\mathbb{P}(X_1 > 1 + t | X_1 > 1) = \mathbb{P}(X_1 > t) = \mathrm{e}^{-t/60}.$$

But similarly, the probability that the old lobster survives t more years is

$$\mathbb{P}(X_2 > 100 + t | X_2 > 100) = \mathbb{P}(X_2 > t) = \mathrm{e}^{-t/60},$$

so at any given time, the probability that the 1 year old lobster dies is equal to the probability that the 100 year old lobster dies! We conclude that the probability that the young lobster outlives the old one is 50%. ◊

Example 37 (Some more examples of random variables that have exponential distribution). In general, we can assume that the exponential distribution is valid for any random variable that has the lack-of-memory property. Sometimes it is obvious to see that this property holds, other times it is not. Some examples of real-life random variables that have exponential distribution are

(a) The time until a single radioactive atom decays.
(b) The time between two customers at a store.
(c) The time between two consecutive queries to a database.
(d) The time until default on a loan payment.
(e) The distance between mutations on a strand of DNA.
(f) The distance between two typos in a text.
(g) The time between two shark attacks.[37]

[37] If some of these examples seem similar to the examples for Poisson r.v.s, that is because they are. There is a deep relation between exponential and Poisson random variables: if the times between consecutive events are distributed as exponential random variables, then the number of events in a given interval is distributed as a Poisson r.v. We will not explore this relation further in this course, but the relations between exponentials and Poissons is captured in the "Poisson process." See page 144 of Ross for a short treatment and Section 9.1 for a slightly longer overview.

△

Question 36. *The amount of beets (in metric tonnes) that a sugar refinery can process in a day is distributed as an Exp$(1/4)$ random variable. How many beets should the refinery have in stock on a given day so that the probability of running out is just 5%?*

Solution. Let X denote the amount of beets (in tonnes) processed in a day. Then $X \sim \mathrm{Exp}(1/4)$. Note that $\mathbb{E}[X] = 4$ tonnes. We want to calculate the value of a such that

$$\mathbb{P}(X > a) = 0.05.$$

We use the c.d.f. of X:

$$\mathbb{P}(X > a) = 1 - \mathbb{P}(X \le a) = 1 - F(a) = 1 - (1 - \mathrm{e}^{-a/4}) = \mathrm{e}^{-a/4}.$$

Now we take the log of both sides:

$$-\frac{a}{4} = \log(0.05) \qquad \Rightarrow \qquad a = 4\log(20) = 11.98.$$

◊

6.1.3. *The normal distribution.* The final example of a continuous random variable that we discuss is simply known as the *normal distribution*. The normal distribution got that name because statisticians are so much used to seeing it, that they say consider their data 'normal' if it has this distribution. We say that a random variable X is normally distributed with parameters μ and σ^2 (corresponding to the mean and variance of X) if X has the following p.d.f.:

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

If X is normally distributed with parameters μ and σ, then we write $X \sim \mathcal{N}(\mu, \sigma)$.

The shape of $f(x)$ is a bell-shaped curve that is symmetric around the point μ on the x-axis. (See figure 10.) For this reason, the normal distribution is also often referred to as the 'bell curve'.[38]

The normal distribution is normal for a reason. Mathematicians know this reason as the 'central limit theorem'. In short, the central limit theorem states that almost any averaged sum of random variables of almost any distribution will have a normal distribution. This fact is so deep and so useful that it has takes the central place in probability theory, hence the name. The central limit theorem will be the final goal of this course as well, but we still have a way to go before we can study it. One step is of course to understand random variables with a normal distribution.

[38] A third common name for the normal distribution is the 'Gauss distribution' (or simply the 'Gaussian'), in honour of Karl Friedrich Gauss, who extensively studied the normal distribution. Gauss did not discover the normal distribution though, that honour goes to Abraham de Moivre. There are short bios of Gauss and de Moivre at the end of Section 5.4 of Ross that are worth reading.

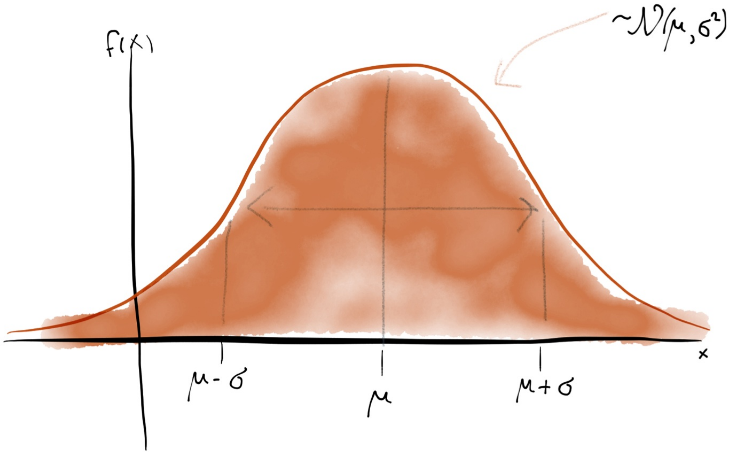

FIGURE 10. A schematic drawing the p.d.f. of a normally distributed random variable with mean μ and variance σ^2.

QUESTION 37. *Let $X \sim \mathcal{N}(\mu, \sigma)$. Calculate $\mathbb{P}(X \le a)$.*

Solution. Unless $a = -\infty, \mu$, or ∞, nobody knows how to do this![39] When $a = \mu$ we can use the symmetry of f around the point μ to determine that

$$\mathbb{P}(X \le \mu) = \mathbb{P}(X \ge \mu),$$

and since $\mathbb{P}(X \le \mu) + \mathbb{P}(X \ge \mu) = 1$ we get

$$\mathbb{P}(X \le \mu) = \frac{1}{2}.$$

◊

In the above calculation we assumed that $\mathbb{P}(X \in (-\infty, \infty)) = 1$, but we have yet to determine that, so lets. To prove that $f(x)$ is a normalised p.d.f. we need to show that

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

We start by making the substitution $y = (x - \mu)/\sigma$ so that

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

so we need to show that

$$I := \int_{-\infty}^{\infty} e^{-y^2/2} dy = \sqrt{2\pi}.$$

[39] We can approximate the solution very accurately, but we will get to that later.

We can show that this is the case by using the following beautiful little proof due to Gauss: We take the square of I:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}\mathrm{d}y \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)/2}\mathrm{d}y\mathrm{d}x. \end{aligned}$$

Now comes the clever part: since the equation $x^2 + y^2 = r^2$ describes a circle of radius r, we can move to polar coordinates and turn this double integral into a single integral: we write $x = r\cos\theta$ and $y = r\sin\theta$, and recall that this coordinate transformation turns $\mathrm{d}y\mathrm{d}x$ into $r\mathrm{d}r\mathrm{d}\theta$, to get

$$\begin{aligned} I^2 &= \int_{0}^{2\pi}\int_{0}^{\infty} \mathrm{e}^{-r^2/2} r\mathrm{d}r\mathrm{d}\theta \\ &= 2\pi \int_{0}^{\infty} r\mathrm{e}^{-r^2/2}\mathrm{d}r. \end{aligned}$$

Now we substitute once more: $u = r^2/2$ so that $\mathrm{d}u = r\mathrm{d}r$,

$$I^2 = 2\pi \int_{0}^{\infty} \mathrm{e}^{-u}\mathrm{d}u = 2\pi\left[-\mathrm{e}^{-u}\right|_0^\infty = 2\pi.$$

So it follows that $I = \sqrt{2\pi}$, as required.

The best way to write the c.d.f. of $X \sim \mathcal{N}(\mu,\sigma)$ is simply as an integral:

$$F_X(b) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\infty}^{b} \mathrm{e}^{-(x-\mu)/2\sigma^2}\mathrm{d}x$$

(where we write the subscript X to remind ourselves that this is the c.d.f. of X). As mentioned in Question 37 we cannot calculate the c.d.f. of X explicitly. We can calculate $F_X(b)$ numerically (with a computer), but since $\mathcal{N}(\mu,\sigma)$ is a two-parameter distribution, it can be a hassle. Fortunately, we don't have to. All we need to know is the c.d.f. $F_Y(a)$ for the random variable $Y \sim \mathcal{N}(\mu = 0, \sigma = 1)$. Such a random variable Y is known as a *standard random variable*. The reason that we only need to know the c.d.f. of Y is as follows:

PRINCIPLE 26 [Linear transformations of normal random variables are normal too]. *If X is a normal random variable with $X \sim \mathcal{N}(\mu,\sigma)$ then, for any $a \neq 0$ and b, $Y = aX + b$ is distributed as $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.*

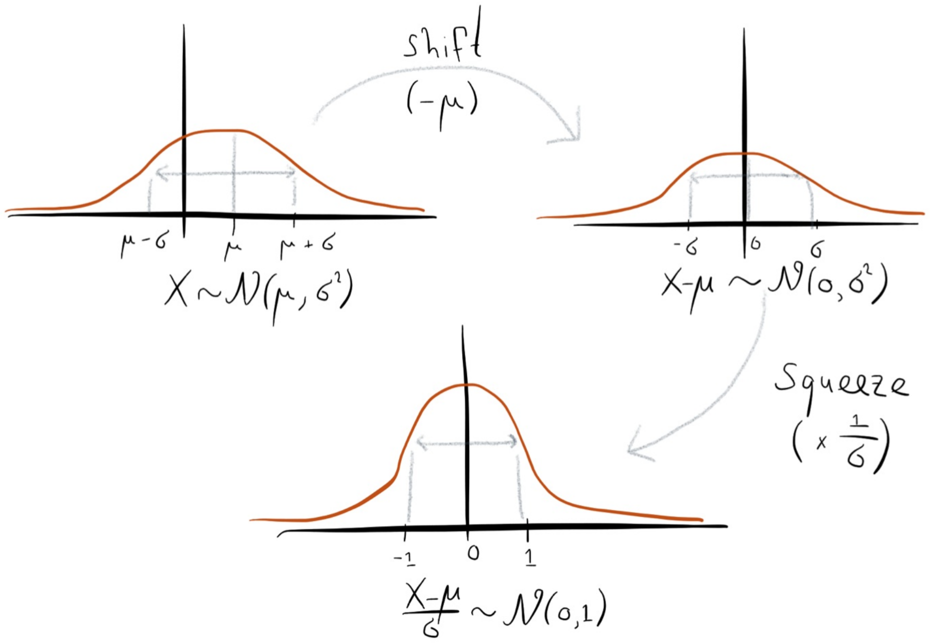
shift
$(-\mu)$
$\mu-\sigma$
μ
$\mu+\sigma$
$X\sim\mathcal{N}(\mu,\sigma^2)$
$-\sigma$
0
σ
$X-\mu\sim\mathcal{N}(0,\sigma^2)$
Squeeze
$(\times\frac{1}{\sigma})$
-1
0
1
$\frac{X-\mu}{\sigma}\sim\mathcal{N}(0,1)$

FIGURE 11. A schematic drawing of the standardisation of X via linear transformations.

Using this principle, we can *standardise* any X:

$$Y = \frac{X-\mu}{\sigma}.$$

Standardising a random variable can be viewed as first *shifting* X by $-\mu$, and by then *squeezing* (or stretching) $X-\mu$ by $\frac{1}{\sigma}$. See figure 11.

As we calculated in the homework, if $\mathbb{E}[X] = \mu$ and $\operatorname{Var}(X) = \sigma^2$, then $\mathbb{E}[Y] = 0$ and $\operatorname{Var}(Y) = 1$, so if $X \sim \mathcal{N}(\mu,\sigma)$ then $Y \sim \mathcal{N}(0,1)$. We can use this to F_X in terms of F_Y as follows:

$$F_X(b) = \mathbb{P}(X \le b) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = \mathbb{P}\left(Y \le \frac{b-\mu}{\sigma}\right) = F_Y\left(\frac{b-\mu}{\sigma}\right).$$

Proof of Principle 26. Since $Y = aX + b$, we have

$$F_Y(x) = \mathbb{P}(Y \le x) = \mathbb{P}(aX + b \le x) = \mathbb{P}\left(X \le \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right).$$

Since F_Y is the antiderivative of f_Y and F_X is the antiderivative of F_X it follows that

$$f_Y(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_Y(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X\left(\frac{x-b}{a}\right) = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}y}F_X(y),$$

where in the last line we applied the chain rule of differentiation with $y = \frac{x-b}{a}$. Now it follows that

$$f_Y(x) = \frac{1}{a}\frac{\mathrm{d}}{\mathrm{d}y}F_X(y) = \frac{1}{a}f_X(y) = \frac{1}{a}f_X\left(\frac{x-b}{a}\right).$$

Using the formula for f_X we get

$$f_Y(x) = \frac{1}{a}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(\frac{x-b}{a}-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}(a\sigma)}\mathrm{e}^{-\frac{(x-(a\mu+b))^2}{2(a\sigma)^2}},$$

so f_Y is the pdf of a normal random variable $Y \sim \mathcal{N}(a\mu + b, a\sigma)$ as claimed. □

With Principle 26 things simplify a lot. Someone has once gone through the trouble of calculating numerically, the values of the function

$$\Phi(z) := F_Z(z) \qquad \text{where } Z \sim \mathcal{N}(0,1),$$

and has made the neat table that appears on the next page.

With this table we can look up for instance the value of $\Phi(1.22)$ by looking at the number that is in the intersection of the row with value 1.2 and the column with value 0.02, i.e, $\Phi(1.22) = 0.8888$. See figure 12 for a sketch of $\Phi(z)$.

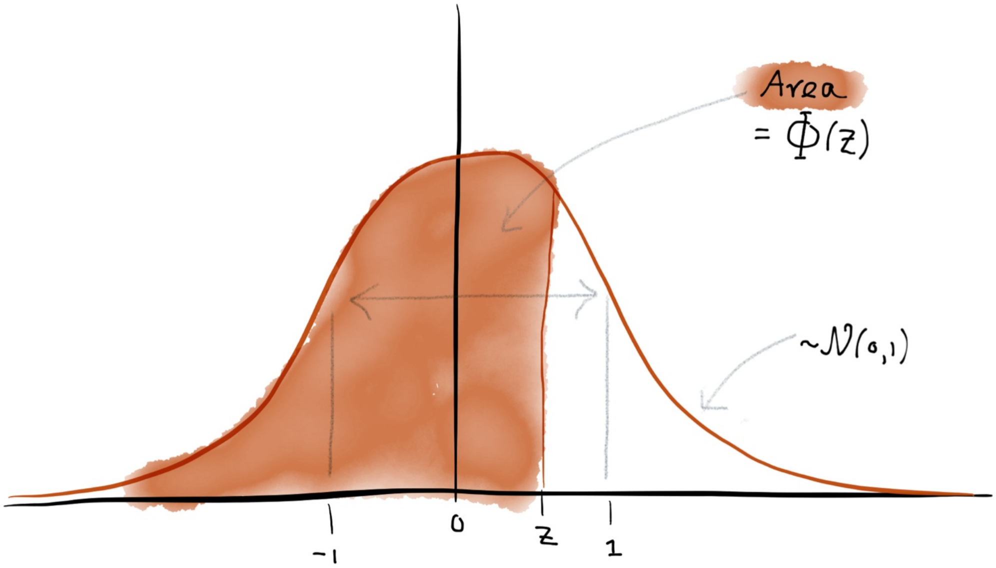

FIGURE 12. A schematic drawing of $\Phi(z)$ in terms of the area between $-\infty$ and z underneath $f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$.

Note that, since the standard normal distribution is symmetric around 0, it follows that

$$\Phi(-z) = \mathbb{P}(Z \le -z) = 1 - \mathbb{P}(Z > -z) = 1 - \mathbb{P}(Z < z) = 1 - \Phi(z).$$

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	0.8315	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

TABLE 1. Values of $\Phi(z)$ (the c.d.f. of an $\mathcal{N}(0,1)$ random variable.)

We can use this fact to look up the values for negative z as well. For instance, if we want to find $\Phi(-1.98)$ in the table, we use

$$\Phi(-1.98) = 1 - \Phi(1.98) = 1 - 0.9761 = 0.0239.$$

QUESTION 38. *The annual rainfall in Vancouver in centimetres is distributed as a normal random variable with mean $\mu = 110$ cm, and standard deviation $\sigma = 10$ cm.*

(a) *Find* $\mathbb{P}($*annual rainfall exceeds 135 cm*$)$.
(b) *Find* $\mathbb{P}($*annual rainfall is between 95 and 125 cm*$)$.
(c) *Find* $\mathbb{P}($*starting this year it will take over 10 years until the annual rainfall exceeds 135 cm*$)$.

Solution. Let $X \sim \mathcal{N}(\mu = 110, \sigma^2 = 10^2)$ denote the annual rainfall and let Z be the standardised random variable

$$Z = \frac{X-\mu}{\sigma} = \frac{X-110}{10} \sim \mathcal{N}(0,1).$$

(a) We calculate

$$\begin{aligned}\mathbb{P}(X > 135) &= \mathbb{P}\left(\frac{X-110}{10} > \frac{135-110}{10}\right)\\ &= \mathbb{P}(Z > 2.5) = 1 - \Phi(2.5) = 1 - 0.9938\\ &= 0.0062.\end{aligned}$$

(b) Similarly,

$$\begin{aligned}\mathbb{P}(95 \le X \le 125) &= \mathbb{P}\left(\frac{95-110}{10} \le \frac{X-110}{10} \le \frac{125-110}{10}\right)\\ &= \mathbb{P}(-1.5 \le Z \le 1.5) = \Phi(1.5) - \Phi(-1.5)\\ &= 2\Phi(1.5) - 1 = 2 \cdot 0.9332 - 1\\ &= 0.8664.\end{aligned}$$

(c) Let Y denote the 1st year with rainfall exceeding 135 cm. Then $Y \sim \text{Geo}(p = 0.0062)$ (by (a)), so

$$\mathbb{P}(Y > 10) = (1-p)^{10} = (0.9938)^{10} = 0.9397.$$

◊

QUESTION 39. *The annual return of a stock is normally distributed with mean 10% and standard deviation 12%. If we buy 100 shares at $60 each, what is the probability that after one year our net profit is at least $750?*

Solution. Let X = value of stock portfolio after one year. The return on our entire investment is given by

$$R = \frac{X - 6000}{6000}.$$

From the question we can deduce that $R \sim \mathcal{N}(\mu = 0.1, \sigma^2 = 0.12^2)$. We want to calculate

$$\begin{aligned}\mathbb{P}(X - 6000 \geq 750) &= \mathbb{P}\left(\frac{X - 6000}{6000} \geq \frac{750}{6000}\right) = \mathbb{P}(R \geq 0.125) \\ &= \mathbb{P}\left(\frac{R - 0.1}{0.12} \geq \frac{0.125 - 0.1}{0.12}\right) = \mathbb{P}(Z \geq 0.208) \\ &= 1 - \Phi(0.208) = 1 - 0.5832 \\ &= 0.4168.\end{aligned}$$

so with 42% chance we will make at least $750 on our investment. ◊

Question 40. *The scores of a test given to 100,000 students are normally distributed with mean 500 and standard deviation 100. What score will place a student in the top 10%?*

Solution. Let

$$X = \text{ score of randomly selected student } \sim \mathcal{N}(500, 100^2).$$

We seek s such that $\mathbb{P}(X \geq s) = 0.1$. We start by standardising:

$$0.1 = \mathbb{P}\left(\frac{X - \mu}{\sigma} \geq \frac{s - \mu}{\sigma}\right) = \mathbb{P}\left(Z \geq \frac{s}{100} - 5\right) = 1 - \Phi\left(\frac{s}{100} - 5\right)$$

So we want to solve

$$\Phi\left(\frac{s}{100} - 5\right) = 0.9.$$

In the table we can find the close approximation $\Phi(1.28) = 0.8997$, so it follows that

$$1.28 = \frac{s}{100} - 5 \qquad \Rightarrow \qquad s = 100(5 + 1.28) = 628,$$

so the student will be in the top 10% if the score is at least 628. ◊

7. Joint probability distributions

So far we have studied random variables either on their own, or as part of an ensemble of independent random variables. Now it is time to explore the situation where the random variables depend on one another. Such dependencies are very common, for instance with stock prices: viewed on their own, the Dow Jones and the Nasdaq behave as though they are randomly fluctuating, but if one takes a dive, then other index will likely do the same.

To study pairs of random variables jointly, we introduce the following notation:

Definition 15 [Joint probability mass function]. *For discrete random variables X and Y, the* joint probability mass function *of X and Y is given by*

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

Observe that $f(x, y)$ is a function that takes two variables as input and gives one number as it’s output. If the function $f(x, y)$ is continuous, it therefore describes a *surface*. See figure 13 for an example.

Question 41. *An urn contains 5 marbles, two red ones and 3 blue ones. We draw marbles from the urn without replacement until the two red marbles are found. Let*

X = # draws until 1st red marble and Y = # additional draws until 2nd red marble.

Find the joint p.m.f. of X and Y.

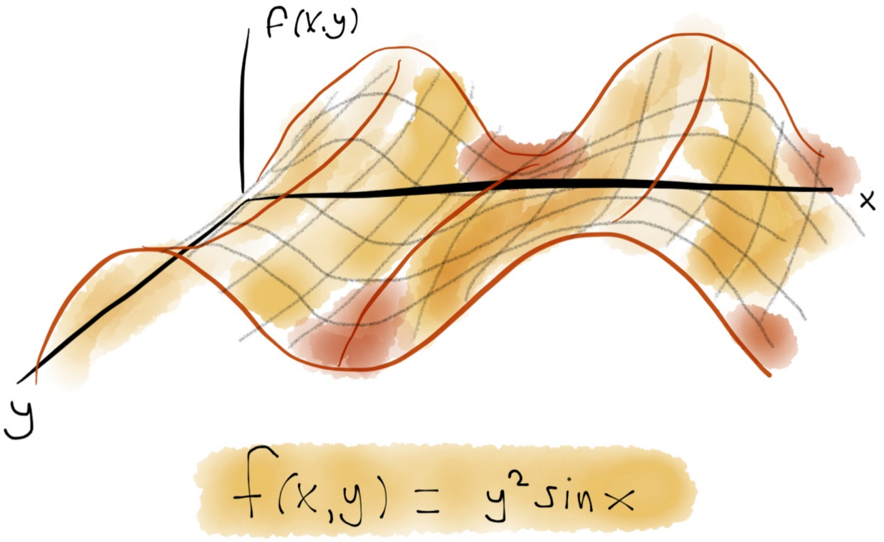

Figure 13. A sketch of a function of a continuous function on two variables.

Solution. Start by noting that $X, Y \geq 1$ and that $X + Y \leq 5$. We can calculate the individual probabilities and write the joint p.m.f. of X and Y as a table:

$X \setminus Y$	1	2	3	4
1	$\frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$	$\frac{2}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{10}$	$\frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{10}$	$\frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{1} = \frac{1}{10}$
2	$\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = \frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	0
3	$\frac{1}{10}$	$\frac{1}{10}$	0	0
4	$\frac{1}{10}$	0	0	0

Note that the entries of the table sum up to 1, and that we if we ignore the value of Y, then we can calculate the p.m.f. of X on its own:

$$f_X(x) = \mathbb{P}(X = x) = \sum_{y=1}^{4} \mathbb{P}(X = x, Y = y) = \sum_{y=1}^{4} f(x, y),$$

i.e.,

$$f_X(1) = \frac{4}{10}, \qquad f_X(2) = \frac{3}{10}, \qquad f_X(3) = \frac{2}{10}, \qquad \text{and} \qquad f_X(4) = \frac{1}{10}.$$

◊

We have a special name and notation for p.m.f.'s of single random variables that are derived from joint p.m.f.'s:

Definition 16 [Marginal probability mass functions]. *Let X be a discrete random variable that takes its values on the set $\mathcal{I}$ and let Y be a discrete random variable that takes its values on the set $\mathcal{J}$ with joint p.m.f. $f(x, y) = \mathbb{P}(X = x, Y = y)$, then we define the* marginal probability mass functions *of X and Y as*

$$f_X(x) := \mathbb{P}(X = x) = \sum_{y \in \mathcal{J}} f(x, y) \qquad \textit{and} \qquad f_Y(y) := \sum_{x \in \mathcal{I}} f(x, y).$$

We can of course give similar definitions for continuous random variables:

Definition 17 [Joint probability density function]. *For continuous random variables X and Y, the* joint probability density function *is given by $f(x,y)$ if for any set $A \subseteq \mathbb{R} \times \mathbb{R}$,*

$$\mathbb{P}((X,Y) \in A) = \iint_A f(x,y)\mathrm{d}x\mathrm{d}y.$$

In particular, for the rectangle $[a,b] \times [c,d] \subset \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}((X,Y) \in [a,b] \times [c,d]) = \int_c^d \int_a^b f(x,y)\mathrm{d}x\mathrm{d}y.$$

The marginal densities *$f_X(x)$ and $f_Y(y)$ of X and Y are given by*

$$f_X(x) := \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y \qquad \text{and} \qquad f_Y(y) := \int_{-\infty}^{\infty} f(x,y)\mathrm{d}x,$$

such that for $B \subset \mathbb{R}$,

$$\mathbb{P}(X \in B) = \mathbb{P}(X \in B, Y \in (-\infty,\infty)) = \int_B \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\mathrm{d}x = \int_B f_X(x)\mathrm{d}x,$$

and similarly,

$$\mathbb{P}(Y \in B) = \int_B f_Y(y)\mathrm{d}y.$$

Question 42. *The joint p.d.f. of X and Y is given by*

$$f(x,y) = \begin{cases} cxy & \text{for } 0 < x < 1 \text{ and } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) *Find c.*
(b) *Find $f_X(x)$.*
(c) *Find $\mathbb{P}(X > Y)$.*
(d) *Find $\mathbb{P}(Y > 1/2 | X < 1/2)$.*

Solution. (a) We want to find the value for c such that the integral is normalised:

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\mathrm{d}x\mathrm{d}y = \int_0^2 \int_0^1 cxy\mathrm{d}x\mathrm{d}y$$
$$= c\int_0^2 \left[\frac{x^2 y}{2}\right|_0^1 \mathrm{d}y = c\int_0^2 \frac{y}{2}\mathrm{d}y = c\left[\frac{y^2}{4}\right|_0^2 = c,$$

so it follows that $c = 1$.

(b) We use the definition of a marginal p.d.f.: for $0 < x < 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y = \int_0^2 xy\mathrm{d}y = \left[\frac{xy^2}{2}\right|_0^2 = 2x.$$

(c) Since $X > Y$, we want to integrate first y from 0 to x (the values for which Y is smaller than X), and then x from 0 to 1 (it helps if you draw a picture of the integration domain):

$$\begin{aligned}\mathbb{P}(X > Y) &= \iint_{x>y} f(x,y)\mathrm{d}y\mathrm{d}x = \int_0^1\int_0^x f(x,y)\mathrm{d}y\mathrm{d}x \\ &= \int_0^1\int_0^x xy\mathrm{d}y\mathrm{d}x = \int_0^1 \left[\frac{xy^2}{2}\right|_0^x \mathrm{d}x \\ &= \int_0^1 \frac{x^3}{2}\mathrm{d}x = \left[\frac{x^4}{8}\right|_0^1 = \frac{1}{8}.\end{aligned}$$

(d) Recall the definition of conditional probabilities:

$$\begin{aligned}\mathbb{P}(Y > 1/2 | X < 1/2) &= \frac{\mathbb{P}(Y > 1/2, X < 1/2)}{\mathbb{P}(X < 1/2)} = \frac{\int_{1/2}^{2}\int_0^{1/2} xy\mathrm{d}x\mathrm{d}y}{\int_0^{1/2} 2x\mathrm{d}x} \\ &= \frac{\int_{1/2}^{2} y\left[\frac{x^2}{2}\right|_0^{1/2}\mathrm{d}y}{\left[x^2\right|_0^{1/2}} = 4\int_{1/2}^{2}\frac{y}{8}\mathrm{d}y = 4\left[\frac{y^2}{16}\right|_{1/2}^{2} \\ &= 4\left(\frac{4}{16} - \frac{1/4}{16}\right) = 1 - \frac{1}{16} = \frac{15}{16}.\end{aligned}$$

◊

7.1. Independent random variables.

Recall from Section 4.3 that two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We can extend this notion of independence to joint probability distributions:

Definition 18. *Two random variables X and Y are with joint p.d.f. or joint p.m.f. $f(x,y)$ are independent if and only if*

$$f(x,y) = f_X(x)f_Y(y).$$

Question 43. *Are the random variables X and Y described in Question 41 independent?*

$X \setminus Y$	1	2	3	4
1	1/10	1/10	1/10	1/10
2	1/10	1/10	1/10	0
3	1/10	1/10	0	0
4	1/10	0	0	0

Solution. Recall that X and Y have the following joint p.m.f.:

and recall that the marginal distributions of X and Y are given by

$$f_X(1) = \frac{4}{10}, \qquad f_X(2) = \frac{3}{10}, \qquad f_X(3) = \frac{2}{10}, \qquad \text{and} \qquad f_X(4) = \frac{1}{10}.$$

and

$$f_Y(1) = \frac{4}{10}, \qquad f_Y(2) = \frac{3}{10}, \qquad f_Y(3) = \frac{2}{10}, \qquad \text{and} \qquad f_Y(4) = \frac{1}{10}.$$

Now we can check independence: suppose $X = 1$ and $Y = 1$, then we get

$$f(1,1) = \frac{1}{10}, \qquad \text{while} \qquad f_X(1)f_Y(1) = \frac{4}{10} \cdot \frac{4}{10} \neq \frac{1}{10},$$

so X and Y are not independent. ◊

Question 44. *An police car is patrolling an L kilometre long stretch of highway, travelling back and forth. An accident occurs somewhere on this stretch of highway. What is the distribution of the distance between the police car and the accident?*

Solution. We will start by assuming that both the location of the police car X and the location of the accident Y are uniformly distributed, i.e., $X \sim \text{Unif}[0, L]$ and $Y \sim \text{Unif}[0, L]$, so

$$f_X(x) = \begin{cases} \frac{1}{L} & \text{if } 0 < x < L, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad f_Y(y) = \begin{cases} \frac{1}{L} & \text{if } 0 < y < L, \\ 0 & \text{otherwise.} \end{cases}$$

We will also assume that the location of the police car and of the accident are independent, therefore

$$f(x, y) = \begin{cases} \frac{1}{L^2} & \text{if } 0 < x < L \text{ and } 0 < y < L, \\ 0 & \text{otherwise,} \end{cases}$$

We write Z for the random variable that describes the (absolute) distance between X and Y, i.e., $Z = |X - Y|$. We want to determine $f_Z(z)$.

First, we will try to find the c.d.f. of Z:

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \le z) = \mathbb{P}(|X - Y| \le z) \\ &= \mathbb{P}(-z \le X - Y \le z) = \mathbb{P}(X - z \le Y \le X + z). \end{aligned}$$

Let A denote the set of points $\{x, y\} \in \mathbb{R} \times \mathbb{R}$ such that $x - z \le y \le y + z$, then it follows that

$$F_z(z) = \mathbb{P}(X - z \le Y \le X + z) = \iint\limits_A f(x, y)\mathrm{d}x\mathrm{d}y.$$

Solving this integral with calculus is not easy, because the area of integration A is difficult to work with. Essentially, the area A is the area in the intersection of a square $L \times L$ and the area between the lines $y = x + a$ and $y = x - a$. (Try and draw the picture of this area.) Therefore, the area of A is equal to the area of the square with sides L minus the area of the two equilateral triangles with sides $L - a$, so

$$\begin{aligned} F_z(z) &= \iint\limits_A f(x, y)\mathrm{d}x\mathrm{d}y = \frac{1}{L^2} \times (\text{area of } A) \\ &= \frac{1}{L^2}\left(L^2 - 2\frac{1}{2}(L - z)^2\right) = \frac{1}{L^2}(2Lz - z^2) = \frac{2z}{L} - \frac{z^2}{L^2}. \end{aligned}$$

Now we can determine the p.d.f. of Z by differentiating $F_Z(z)$:

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) = \frac{2}{L} - \frac{2z}{L^2} \qquad (\text{ for } 0 \le z \le L).$$

◊

7.2. **The expectation of a joint distribution.** Recall the law of the unconscious statistician (LoUS):

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x \in \mathcal{I}} g(x)f(x) & \text{if } X \text{ is discrete,} \\ \int\limits_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases}$$

We can generalise again for joint distributions:

Principle 27 [LoUS for joint distributions]. *For any function $g(x, y)$,*

$$\mathbb{E}[g(X, Y)] = \sum_{x \in \mathcal{I}} \sum_{y \in \mathcal{J}} g(x, y)f(x, y) \qquad \textit{if } X \textit{ and } Y \textit{ are discrete,}$$

and

$$\mathbb{E}[g(X, Y)] = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(x, y)f(x, y)\mathrm{d}x\mathrm{d}y \qquad \textit{if } X \textit{ and } Y \textit{ are continuous.}$$

We will omit the proof.

With this generalised version of the LoUS, we can prove Principle 21(b) (back on page 51): recall

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

We can prove this with the function $g(x, y) = x + y$: for continuous random variables X and Y,

$$\begin{aligned}
\mathbb{E}[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\right)\mathrm{d}x + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f(x,y)\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} xf_X(x)\mathrm{d}x + \int_{-\infty}^{\infty} yf_Y(y)\mathrm{d}y \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

The proof for discrete random variables is similar.

Principle 28. *If X and Y are* independent *random variables and $g(x)$ and $h(x)$ are functions, then*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)],$$

and in particular,

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

Proof. Suppose X and Y are discrete (the proof for continuous r.v.'s is similar):

$$\begin{aligned}
\mathbb{E}[g(X)h(Y)] &= \sum_{x\in\mathcal{I}}\sum_{y\in\mathcal{J}} g(x)h(y)f(x,y) = \sum_{x\in\mathcal{I}}\sum_{y\in\mathcal{J}} g(x)h(y)f_X(x)f_Y(y) \\
&= \left(\sum_{x\in\mathcal{I}} g(x)f_X(x)\right)\left(\sum_{y\in\mathcal{J}} h(y)f_Y(y)\right) = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]. \qquad \square
\end{aligned}$$

7.3. **Covariance and correlation.** Random variables can have strong dependencies on one another.[40] We may want to quantify this dependence. To this end we define

Definition 19 [The covariance between two random variables]. *Let X and Y be random variables with $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$. The* covariance *between X and Y is defined as*

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

We will use the covariance mainly to calculate the variance of sums of random variables. To do this, we need to establish some properties of the covariance:

[40] Think again of stock prices.

PRINCIPLE 29 [Properties of the covariance]. *Let X and Y be random variables. Then the following hold:*

(a) *A simple formula for covariance:*

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

which implies that for independent *X and Y,*

$$\operatorname{Cov}(X,Y) = 0.$$

(Note that $\operatorname{Cov}(X,Y) = 0$ *does* not *also imply that X and Y are independent!)*

(b) *The covariance is* symmetric*:*

$$\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X).$$

(c) *The covariance between X and itself is the variance:*

$$\operatorname{Cov}(X,X) = \operatorname{Var}(X).$$

(d) *Linearity of covariance:*

$$\operatorname{Cov}(aX,Y) = a\operatorname{Cov}(X,Y).$$

(e) *Covariance of sums of random variables is* bilinear*:*

$$\operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\operatorname{Cov}(X_i,Y_j).$$

We omit the proofs here. For the proof of (a), see Ross, page 305. The proofs of (b), (c), and (d) follow directly from the definitions of covariance and variance. The proof of (e) can be found on page 306 of Ross.

With all these properties established we can determine the variance of sums of random variables:

PRINCIPLE 30 [The variance of sums of random variables]. *Let X and Y be random variables, then*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y),$$

and more generally

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i}\operatorname{Cov}(X_i,X_j).$$

Proof. We will prove it for two random variables. The general proof is in Ross on page 306. The proof simply involves the definitions of variance and covariance and a clever rearranging of the terms:

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2 + 2XY + Y^2] - \left(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\right) \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y). \quad \square\end{aligned}$$

Recall Principle 22(c) on page 51, which states that for *independent* random variables X and Y,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

With Principles 29(a) and 30 in hand we can finally prove this claim:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 0.$$

QUESTION 45. *A coin comes up heads with probability p. Flip the coin once. Let $X = 1$ if the coin came up heads, and let $X = 0$ if it came up tails. If the coin came up tails, flip it again. If it came up heads do nothing. Let $Y = 1$ if the coin is now heads, and let $Y = 0$ if the coin is now tails.*

(a) *Find $f(x,y)$ and its marginals.*
(b) *Find* $\mathrm{Cov}(X,Y)$.
(c) *Finc* $\mathrm{Var}(X+Y)$.

Solution. (a) We can write the joint p.m.f. of X and Y as a table:

$X \setminus Y$	0	1	$f_X(x)$
0	$(1-p)\cdot(1-p)$	$(1-p)\cdot p$	$(1-p)$
1	0	$p\cdot 1$	p
$f_Y(y)$	$(1-p)(1-p)$	$p + p(1-p)$	

(b) To calculate the covariance, we need to determine $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$, :

$$\begin{aligned}\mathbb{E}[X] &= 0\cdot f_X(0) + 1\cdot f_X(1) = p, \\ \mathbb{E}[Y] &= 0\cdot f_Y(0) + 1\cdot f_Y(1) = p + p(1-p), \\ \mathbb{E}[XY] &= 0\cdot 0\cdot f(0,0) + 1\cdot 0\cdot f(1,0) + 0\cdot 1\cdot f(0,1) + 1\cdot 1\cdot f(1,1) = f(1,1) = p.\end{aligned}$$

Therefore,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = p - p(p + p(1-p)) = p - 2p^2 + p^3.$$

(c) We need to calculate $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$. Marginally, both are Bernoulli random variables, i.e., $X \sim \mathrm{Ber}(p)$ and $Y \sim \mathrm{Ber}(p + p(1-p))$, so

$$\mathrm{Var}(X) = p(1-p) \qquad \text{and} \qquad \begin{aligned}\mathrm{Var}(Y) &= (p + p(1-p))(1-(p+p(1-p)))\\ &= 2p - 5p^2 + 4p^3 - p^4.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\\ &= p(1-p) + (2p - 5p^2 + 4p^3 - p^4) + 2(p + 2p^2 - p^3)\\ &= 5p - 2p^2 + 2p^3 - p^4.\end{aligned}$$

◊

The covariance of X and Y tells us something about their codependence, but it is not exactly clear from the formulation how strong this dependence is. If we want to understand this, we need to compare the covariance to the variances of X and Y:

DEFINITION 20 [The correlation coefficient]. *Given random variables X and Y, we define the* correlation coefficient *of X and Y as*

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Observe that if X is expressed in units $[a]$ and Y is expressed in units $[b]$ (e.g., X is in metres, and Y is in seconds) then $\mathrm{Var}(X)$ is expressed in units $[a^2]$, $\mathrm{Var}(Y)$ is expressed in units $[b^2]$, and $\mathrm{Cov}(X,Y)$ is expressed in units $[a \cdot b]$, so it follows that the correlation coefficient $\rho(X,Y)$ is expressed in units $[(a \cdot b)/\sqrt{a^2 \cdot b^2}] = [1]$, that is, $\rho(X,Y)$ is a unitless (or 'dimensionless') quantity.

Moreover, $\rho(X,Y)$ has the following property:

PRINCIPLE 31. *Let X and Y be random variables, then*

$$-1 \le \rho(X,Y) \le 1.$$

Proof. The proof uses the *Cauchy-Schwarz inequality:*[41]

$$\mathbb{E}[XY]^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

The proof of this inequality goes as follows: assume[42] that $\mathbb{E}[X^2] \neq 0$. Define $Z = tX + Y$. We have the following inequality

$$0 \le \mathbb{E}[Z^2] = \mathbb{E}[(tX+Y)^2] = t^2\mathbb{E}[X^2] + 2t\mathbb{E}[XY] + \mathbb{E}[Y^2] =: at^2 + bt + c.$$

[41] The Cauchy-Schwarz inequality holds more generally for *inner products:*

$$|\langle x, y\rangle|^2 \le \langle x, x\rangle \cdot \langle y, y\rangle,$$

so that for instance for vectors $x, y \in \mathbb{R}^n$, the inequality becomes $|\langle x,y\rangle|^2 \le |x|^2|y|^2$.

[42] $\mathbb{E}[X^2] = 0$ can only happen if $X = 0$ always, which is not an interesting random variable. Moreover, the Cauchy-Schwarz inequality holds trivially if this is the case.

The right hand side should always be positive (i.e., non-negative *and* non-zero) except possibly for the point $t = 0$, so we know that the discriminant of the quadratic function on the right should be non-positive, i.e.,

$$b^2 - 4ac \leq 0.$$

Therefore,

$$4\mathbb{E}[XY]^2 \leq 4\mathbb{E}[X^2]\mathbb{E}[Y^2].$$

This completes the proof of the Cauchy-Schwarz inequality.

Now we apply this inequality to the definition of the covariance:

$$\begin{aligned}|\mathrm{Cov}(X,Y)|^2 &= |\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]|^2\\ &\leq \mathbb{E}[(X-\mathbb{E}[X])^2]\mathbb{E}[(Y-\mathbb{E}[Y])^2]\\ &= \mathrm{Var}(X)\mathrm{Var}(Y).\end{aligned}$$

With this inequality, if follows from the definition of $\rho(X,Y)$ that

$$|\rho(X,Y)| \leq 1$$

and this completes the proof. □

7.4. **Sums of independent random variables.** If we have two independent random variables, X and Y, we can calculate the distribution of $X + Y$:

Principle 32 [The sums of independent random variables]. *Let X and Y be independent continuous random variables with p.d.f. $f_X(x)$ and $f_Y(y)$, respectively. The p.d.f. of $X + Y$ is given by*

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\mathrm{d}y.$$

The function f_{X+Y} is also known as the convolution *of f_X and f_Y.*

Proof. We start by determining the c.d.f. of $X + Y$:

$$\begin{aligned}F_{X+Y}(z) = \mathbb{P}(X+Y\leq z) &= \iint_{x+y\leq z} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f_X(x)\mathrm{d}x\right)f_Y(y)\mathrm{d}y\\ &= \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)\mathrm{d}y.\end{aligned}$$

Now we differentiate with respect to z:

$$\begin{aligned}
f_{X+Y}(z) &= \frac{\mathrm{d}}{\mathrm{d}z}\int_{-\infty}^{\infty} F_X(z-y)f_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}z}F_X(z-y)f_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\mathrm{d}y. \quad \square
\end{aligned}$$

Example 38 (The sum of two independent uniform random variables). Let $X, Y \sim \mathrm{Unif}[0,1]$. We can calculate the distribution of $X+Y$. Recall the p.d.f. of X and Y:

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

Therefore,

$$f_{X+Y}(z) = \int_0^1 f_X(z-y)\cdot 1\mathrm{d}y.$$

Since y is between 0 and 1 on the range of this integral, we get

$$f_X(z-y) = \begin{cases} 1 & \text{if } 0 \le z-y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The values y for which the function is non-zero depend on z: if $z \le 1$, then $0 \le z-y \le 1$ implies that $0 \le y \le z$, while if $1 \le z \le 2$, then $z-1 \le y \le z$, so

$$f_{X+Y}(z) = \begin{cases} \int_0^z 1\cdot 1\mathrm{d}y = z & \text{if } 0 \le z \le 1, \\ \int_{z-1}^1 1\cdot 1\mathrm{d}y = 2-z & \text{if } 1 \le z \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

△

Example 39 (Gamma random variables). Recall the gamma function (discussed on page 65):

$$\Gamma(n) = \int_0^{\infty} y^{n-1}\mathrm{e}^{-y}\mathrm{d}y.$$

The *gamma distribution* is a two-parameter family of probability distribution functions with parameters (t,θ) (where $t, \theta > 0$) and p.d.f.

$$f(x) = \frac{\theta\mathrm{e}^{-\theta x}(\theta x)^{t-1}}{\Gamma(t)} \qquad \text{when} \qquad 0 < x < \infty.$$

The gamma distribution is often used in the econometrics and actuarial science to model the time until a person dies, but it has numerous other applications as well.[43]

The property that we are interested in now, is the following fact: the family of gamma distributions with fixed θ is “closed under convolutions”. This means that if we take two random variables, $X \sim \text{Gamma}(s, \theta)$ and $Y \sim \text{Gamma}(t, \theta)$, then $X+Y \sim \text{Gamma}(s+t, \theta)$. We will now prove this fact:

By Principle 32,

$$\begin{aligned} f_{X+Y}(z) &= \int_0^z f_X(z-y) f_X(y) \mathrm{d}y \\ &= \frac{1}{\Gamma(s)\Gamma(t)} \int_0^z \theta \mathrm{e}^{-\theta(z-y)} \left(\theta(z-y)\right)^{s-1} \theta \mathrm{e}^{-\theta y} (\theta y)^{t-1} \mathrm{d}y \\ &= K\mathrm{e}^{-\theta z} \int_0^z (z-y)^{s-1} y^{t-1} \mathrm{d}y. \end{aligned}$$

Let’s analyse these steps: In the first step, we used the fact that the product $f_X(z-y)f_Y(y)$ is only non-zero when both y and $z-y$ are positive to limit our integration domain to $0 \leq y \leq z$. In the second step we simply applied the definition of the gamma distribution. In the third step, we moved all the factors that do not depend on y outside of the integral, and moreover, we write all the constants here combined as K. We do this because it turns out that it won’t matter to us what the precise value of this K is going to be, as we will see shortly.

We continue by substituting $x = y/z$:

$$f_{X+Y}(z) = K\mathrm{e}^{-\theta z} z^{s+t-1} \int_0^1 (1-x)^{s-1} x^{t-1} \mathrm{d}x.$$

The parameters s and t don’t need to be integers, so this integral can be very difficult, but that does not matter. All that matters is that the integral does not depend on z, so it will give us some constant value. Therefore there exists another constant C such that

$$f_{X+Y}(z) = C\mathrm{e}^{-\theta z} z^{s+t-1}.$$

The reason that we did not bother determining the value of C by solving the integral is as follows: by Principle 32 we already know that the above function is the p.d.f. of $X+Y$, so it must be a *normalised* function. Therefore, we can just determine the value of C by setting the integral equal to 1:

$$1 = \int_{-\infty}^{\infty} f_{X+Y}(z) \mathrm{d}z = C \int_0^{\infty} \mathrm{e}^{-\theta z} (\theta z)^{s+t-1} \mathrm{d}z.$$

[43] See Ross, Section 5.6.1 on page 203 for a brief discussion of the gamma distributions.

Now we substitute $y = \theta z$ to obtain

$$1 = \frac{C}{\theta^{s+t}} \int_0^\infty y^{s+t-1} e^{-y} dy = C \cdot \frac{\Gamma(s+t)}{\theta^{s+t}},$$

so it follows that $C = \theta^{s+t}/\Gamma(s+t)$ and therefore,

$$f_{X+Y}(z) = \frac{\theta e^{-\theta z}(\theta z)^{s+t-1}}{\Gamma(s+t)} \qquad \text{when} \qquad 0 < z < \infty,$$

so $X + Y \sim \text{Gamma}(s+t, \theta)$ as claimed. △

There is another family of distributions that is closed under convolutions:

Principle 33 [The sum of independent normals is normal]. *If X_i for $i = 1, \ldots, n$ are independent normal random variables such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then*

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Proof. Recall that $X \sim \mathcal{N}(\mu, \sigma^2)$ means that X has p.d.f.

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

We start with the special case where we have two random variables with the specific distributions $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{N}(0,1)$. To make our notation a bit easier, we define

$$c := \frac{1}{2\sigma^2} + \frac{1}{2} = \frac{1+\sigma^2}{2\sigma^2}.$$

We want to use Principle 32 so we need to determine $f_X(z-y)f_Y(y)$:

$$\begin{aligned} f_X(z-y)f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-y)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ &= \frac{1}{2\pi\sigma} e^{-\frac{z^2}{2\sigma^2}} e^{-c\left(y^2 - 2y\frac{z}{1+\sigma^2}\right)} \\ &= \frac{1}{2\pi\sigma} e^{-\frac{z^2}{2\sigma^2}} e^{\frac{z^2}{2\sigma^2(1+\sigma^2)}} e^{-c\left(y - \frac{z}{1+\sigma^2}\right)^2} \end{aligned}$$

In the third step we used a method known as *completing the square.*[44]

We can simplify this formula a bit: the exponents of the second and third factor can be written as

$$-\frac{z^2}{2\sigma^2} + \frac{z^2}{2\sigma^2(1+\sigma^2)} = -\frac{z^2(1+\sigma^2)}{2\sigma^2(1+\sigma^2)} + \frac{z^2}{2\sigma^2(1+\sigma^2)} = -\frac{z\sigma^2}{2\sigma^2(1+\sigma^2)} = -\frac{z^2}{2(1+\sigma^2)},$$

[44] In the exponent appears a term of the form $y^2 - by$. We rewrite this as $y^2 - by + b^2/4 - b^2/4$ and then we use the fact that $(y - b/2)^2 = y^2 - by + b^2/4$ to determine that $y^2 - by = (y - b/2)^2 - b^2/4$.

so that

$$f_X(z-y)f_Y(y) = \frac{1}{2\pi\sigma}\mathrm{e}^{-\frac{z^2}{2(1+\sigma^2)}}\mathrm{e}^{-c\left(y-\frac{z}{1+\sigma^2}\right)^2}.$$

We integrate over y to determine $f_{X+Y}(z)$:

$$\begin{aligned} f_{X+Y}(z) &= \frac{1}{2\pi\sigma}\mathrm{e}^{-\frac{z^2}{2(1+\sigma^2)}}\int_0^\infty \mathrm{e}^{-c\left(y-\frac{z}{1+\sigma^2}\right)^2}\mathrm{d}y \\ &= \frac{1}{2\pi\sigma}\mathrm{e}^{-\frac{z^2}{2(1+\sigma^2)}}\int_{-\infty}^\infty \mathrm{e}^{-cx^2}\mathrm{d}x \\ &= C\mathrm{e}^{-\frac{z^2}{2(1+\sigma^2)}}. \end{aligned}$$

In the second step we substituted $x = y - \frac{z}{2(1+\sigma^2)}$ and in the third step we introduced the constant C that does not depend on z. The value of C is unimportant since we know that $f_{X+Y}(z)$ is a p.d.f. and therefore normalised. Moreover, the non-constant part of the p.d.f. has the same form as that of a normally distributed random variable with mean 0 and variance $1+\sigma^2$, so we can conclude that $X+Y \sim \mathcal{N}(0, 1+\sigma^2)$.

Now we can determine that the same holds for two normally distributed random variables with arbitrary means and variances: let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and let $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then we can write

$$X_1 + X_2 = \sigma_2\left(\frac{X_1-\mu_1}{\sigma_2} + \frac{X_2-\mu_2}{\sigma_2}\right) + \mu_1 + \mu_2.$$

Since by Principle 26 normal random variables stay normally distributed under linear transformations, we have that

$$\frac{X_1-\mu_1}{\sigma_2} \sim \mathcal{N}(0, \sigma_1^2/\sigma_2^2) \qquad \text{and} \qquad \frac{X_2-\mu_2}{\sigma_2} \sim \mathcal{N}(0,1),$$

so we can apply our previous calculation to determine that

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1+\mu_2, \sigma_2^2(1+\sigma_1^2/\sigma_2^2)\right) = \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2).$$

Finally, extending this result to general values n is easy. We use the inductive method: we know that the principle holds when $n=2$, so all we need to do is show that if it holds for $n-1$, then it also holds for n. Assume that the principle holds for $n-1$, then

$$\sum_{i=1}^n X_i = \sum_{i=1}^{n-1} X_i + X_n.$$

It is our assumption that

$$\sum_{i=1}^{n-1} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n-1}\mu_i, \sum_{i=1}^{n-1}\sigma_i^2\right) \qquad \text{and} \qquad X_n \sim \mathcal{N}(\mu_n, \sigma_n^2),$$

so that we are adding two normally distributed random variables, and hence by our result for $n=2$, it also holds for general n. □

7.5. Conditional distributions.

Given a joint probability distribution, we can derive conditional probabilities as well.

Let's start by recalling what a conditional probability is: for events A and B, the conditional probability of A given B is

$$\mathbb{P}(A|B) = \begin{cases} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} & \text{if } \mathbb{P}(B) > 0, \\ 0 & \text{if } \mathbb{P}(B) = 0. \end{cases}.$$

We can extend this notion to joint probability distributions in a natural way. The extensions for discrete and continuous random variables are different, so we will treat them separately.

7.5.1. *Conditional distributions for discrete random variables.*

Definition 21 [The conditional p.m.f.]. *Given a joint p.m.f. for the discrete random variables X and Y, we define the conditional p.m.f. of X conditioned on the event $\{Y = y\}$ as*

$$f_{X|Y}(x|y) := \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(\{X = x\} \cap \{Y = y\})}{\mathbb{P}(Y = y)} = \frac{f(x,y)}{f_Y(y)}.$$

Question 46. *Calculate for the random variables X and Y described in Question 41 the following conditional probabilities:*

(a) $\mathbb{P}(X = 3|Y = 2)$
(b) $\mathbb{P}(X = 3|Y = 1)$
(c) $\mathbb{P}(Y = 2|X = 4)$

Solution. Recall that X and Y have the following joint p.m.f. and marginals:

$X \setminus Y$	1	2	3	4	$f_X(x)$
1	1/10	1/10	1/10	1/10	4/10
2	1/10	1/10	1/10	0	3/10
3	1/10	1/10	0	0	2/10
4	1/10	0	0	0	1/10
$f_Y(y)$	4/10	3/10	2/10	1/10	

(a) We use the definition:

$$\mathbb{P}(X = 3|Y = 2) = \frac{f(3,2)}{f_Y(2)} = \frac{1/10}{3/10} = \frac{1}{3}.$$

(b) Again, by the definition:

$$\mathbb{P}(X = 3|Y = 1) = \frac{f(2,2)}{f_Y(1)} = \frac{1/10}{4/10} = \frac{1}{4}.$$

(c) And again:

$$\mathbb{P}(Y = 2|X = 4) = \frac{f(4,2)}{f_X(4)} = \frac{0}{1/10} = 0.$$

◊

7.5.2. *Conditional distributions for continuous random variables.* Even though the probability of the event $\{Y = y\}$ is always 0 for continuous random variables, we can still define the conditional distributions of X given $\{Y = y\}$:

Definition 22 [The conditional p.d.f. and c.d.f.]. *Given a joint p.d.f. $f(x,y)$ for the continuous random variables X and Y, we define the conditional p.d.f. of X given $\{Y = y\}$ as*

$$f_{X|Y}(x|y) := \frac{f(x,y)}{f_Y(y)}.$$

Moreover, we define the conditional c.d.f. of X given Y as

$$F_{X|Y}(x|y) := \int_{-\infty}^{x} f_{X|Y}(x|y)\mathrm{d}x.$$

Note that both the conditional p.d.f. *and* the conditional c.d.f. can be functions of y. Nevertheless, the c.d.f. does describe a probability for any *fixed* y.

Question 47. *Recall from Question 42 on page 78 the following joint p.d.f. for random variables X and Y:*

$$f(x,y) = \begin{cases} xy & \text{for } 0 < x < 1 \text{ and } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(Y > 1/2|X = x)$.

Solution. Let's start by recalling the marginals of X and Y from Question 42: $f_X(x) = 2x$ and $f_Y(y) = \frac{y}{2}$. Therefore,

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \begin{cases} \frac{y}{2} & \text{for } 0 < x < 1 \text{ and } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Now we can determine $\mathbb{P}(Y > 1/2|X = x)$ by integrating:

$$\mathbb{P}(Y > 1/2|X = x) = \int_{1/2}^{\infty} f_{Y|X}(y|x)\mathrm{d}y = \int_{1/2}^{2} \frac{y}{2}\mathrm{d}y$$

$$= \left[\frac{y^2}{4}\right|_{1/2}^{2} = 1 - \frac{1}{16} = \frac{15}{16},$$

so in fact *any* conditioning on x gives us the same conditional probability distribution for Y. (Compare with Question 42(d).) ◊

More generally, we have the following

Principle 34 [Conditional distributions of independent r.v.'s]. *If X and Y are* independent *random variables (both either discrete or continuous), then*

$$f_{X|Y}(x|y) = f_X(x).$$

Proof. Simply apply the definitions of independence and conditional distributions:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x). \qquad \square$$

Question 48. *Let X and Y be continuous random variables with joint p.d.f.*

$$f(x,y) = \frac{\mathrm{e}^{-x/y}\mathrm{e}^{-y}}{y} \qquad \text{when } 0 < x < \infty \quad \text{and} \quad 0 < y < \infty.$$

(a) *Find $f_{X|Y}(x|y)$.*
(b) *Find $\mathbb{P}(X > 1|Y = y)$.*

Solution. (a) Since $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$ we should first determine the marginal of Y:

$$f_Y(y) = \int_0^\infty \frac{\mathrm{e}^{-x/y}\mathrm{e}^{-y}}{y}\mathrm{d}x = \frac{\mathrm{e}^{-y}}{y}\int_0^\infty \mathrm{e}^{-x/y}\mathrm{d}x = \frac{\mathrm{e}^{-y}}{y}\left[-y\mathrm{e}^{-x/y}\right|_0^\infty = \mathrm{e}^{-y}.$$

Now we can determine

$$f_{X|Y}(x|y) = \frac{\frac{\mathrm{e}^{-x/y}\mathrm{e}^{-y}}{y}}{\mathrm{e}^{-y}} = \frac{\mathrm{e}^{-x/y}}{y}.$$

(b) We integrate $f_{X|Y}(x|y)$ from 1 to ∞ for the answer:

$$\mathbb{P}(X > 1|Y = y) = \int_1^\infty \frac{\mathrm{e}^{-x/y}}{y}\mathrm{d}x = \left[-\mathrm{e}^{-x/y}\right|_1^\infty = \mathrm{e}^{-1/y}.$$

Note that this answer still depends on y, so X and Y are not independent. ◊

7.5.3. *Conditional expectation.* Just as with any probability distribution, we cal also determine the expectation of X conditioned on the event $\{Y = y\}$:

Definition 23 [Conditional expectation]. *Let X and Y be random variables with joint distribution $f(x,y)$, then the conditional expectation of X given y is defined as*

$$\mathbb{E}[X|Y = y] = \sum_{x\in\mathcal{I}} x f_{X|Y}(x|y) \qquad \text{if } X \text{ and } Y \text{ are discrete,}$$

and

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \qquad \text{if } X \text{ and } Y \text{ are continuous.}$$

Let's calculate the conditional expectation for the joint p.d.f. in Question 48:

Question 49. *Let X and Y be continuous random variables with joint p.d.f.*

$$f(x,y) = \frac{e^{-x/y}e^{-y}}{y} \qquad \textit{when } 0 < x < \infty \quad \textit{and} \quad 0 < y < \infty.$$

Find $\mathbb{E}[X|Y=y]$.

Solution. We already know the conditional density:

$$f_{X|Y}(x|y) = \frac{e^{-x/y}}{y}.$$

Therefore,

$$\mathbb{E}[X|Y=y] = \int_0^\infty \frac{x}{y}e^{-x/y}dx.$$

We can solve this integral via integration by parts, using $f(x) = x$ and $\frac{d}{dx}g(x) = \frac{e^{-x/y}}{y}$ (so that $g(x) = -e^{-x/y}$):

$$\begin{aligned}\mathbb{E}[X|Y=y] &= [f(x)g(x)|_0^\infty - \int_0^\infty \left(\frac{d}{dx}f(x)\right)g(x)dx \\ &= \left[-xe^{-x/y}\right|_0^\infty + \int_0^\infty e^{-x/y}dx \\ &= 0 + \left[-ye^{-x/y}\right|_0^\infty = y.\end{aligned}$$

so conditionally on the event that $\{Y=y\}$ the expected value of X also becomes y. ◊

In general, the expectation of X conditioned on $\{Y=y\}$ can be a function that depends on y. For any *fixed* y, however, $\mathbb{E}[X|Y=y]$ is a true expectation in the sense that is satisfies all the properties of an expectation. In essence, $\mathbb{E}[X|Y=y]$ is simply the expectation on the reduced sample space given by the information $\{Y=y\}$, as described in the beginning of Section 4.

7.6. **Using conditional expectations to compute unconditioned expectations.** Like conditional probabilities, so are conditional expectations powerful tool for computations. Unfortunately, conditional expectations are sometimes a bit difficult to work with, so some care is needed when we try and use this tool. Ross gives a number of nice examples of this method in Section 7.5.2, but we will restrict ourselves to a single example here.

The following principle supplies the tool:

Principle 35 [The expectation over conditional expectations is the ordinary expectation]. *For random variables X and Y, (a) For discrete random variables X and Y,*

$$\mathbb{E}[X] = \sum_{y \in \mathcal{J}} \mathbb{E}[X|Y = y]\mathbb{P}(Y = y).$$

(b) For continuous random variables X and Y,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] f_Y(y)\mathrm{d}y.$$

The general notation for this principle is

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]].$$

The proof for discrete random variables can be found on page 315 of Ross, so we omit that. The proof for the continuous case goes as follows:

Proof of (b). We simply apply the definitions of $\mathbb{E}[X|Y = y]$, $f_{X|Y}(x|y)$ and $f_X(x)$ (in that order) to the right-hand side:

$$\begin{aligned}\int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] f_Y(y)\mathrm{d}y &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y)\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} f_Y(y)\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\right)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x = \mathbb{E}[X]. \quad \square\end{aligned}$$

Question 50. *A miner has lost his way in a mine, in a room containting three doors. If the miner takes the first door, then the miner gets out of the mine in 3 hours. If the miner takes the second door, then he returns to the room after 5 hours. If the miner takes the third door, then he returns to the room in 7 hours. The miner is disoriented, so each time he gets to the room, he picks a door at random and follows the path. What is the expected time until the miner escapes?*

Solution. Let X denote the escape time and Y denote the door the miner chooses. We can determine $\mathbb{E}[X]$ with conditional expectations. Since $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$,

$$\mathbb{E}[X] = \mathbb{E}[X|Y = 1]\mathbb{P}(Y = 1) + \mathbb{E}[X|Y = 2]\mathbb{P}(Y = 2) + \mathbb{E}[X|Y = 3]\mathbb{P}(Y = 3).$$

We know from the question that $\mathbb{P}(Y = i) = 1/3$ for $i = 1, 2, 3$. Moreover, it is given that

$$\mathbb{E}[X|Y = 1] = 3.$$

To determine $\mathbb{E}[X|Y=2]$ we have to reason a bit: if the miner takes door 2, then he returns to the room after 5 hours. When he returns to the room, he picks a door at random again, so the expected remaining time to escape *after returning* to the room is also $\mathbb{E}[X]$, by the definition of X. Therefore,

$$\mathbb{E}[X|Y=2] = 5 + \mathbb{E}[X].$$

Similarly we can reason that

$$\mathbb{E}[X|Y=3] = 7 + \mathbb{E}[X].$$

We can now use that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ to solve the problem:

$$\mathbb{E}[X] = \frac{1}{3}\left(3 + (5 + \mathbb{E}[X]) + (7 + \mathbb{E}[X])\right) = 5 + \frac{2}{3}\mathbb{E}[X] \qquad \Rightarrow \qquad \mathbb{E}[X] = 15,$$

so it will take the miner an expected 15 hours to escape the mine. ◊

8. Moment generating functions

Let's start with the definition:

Definition 24 [The moment generating function]. *Given a random variable X, we define the* moment generating function of X *as*

$$M_X(t) := \mathbb{E}\left[e^{tX}\right],$$

so that in particular, if X is discrete, then

$$M_X(t) = \sum_{x\in\mathcal{I}} e^{tx}\mathbb{P}(X = x),$$

while if X is continuous with p.d.f. $f(x)$, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\mathrm{d}x.$$

So why should we bother with such a complicated looking thing? Because moment generating functions (or m.g.f.'s) are an extremely useful tool in probability theory. They are part of a larger framework, that of generating functions. Generating functions are an indispensable tool in both probability theory and discrete mathematics. The following quote from the excellent introductory book *Generatingfunctionology* by Herbert Wilf[45] sums up the basic idea of generating functions:

> A generating function is a clothesline on which we hang up a sequence of numbers for display. What that means is this: suppose we have a problem whose answer is a sequence of numbers, $a_0, a_1, a_2, \ldots$. We want to know what the sequence is. What kind of an answer might we expect?
> [...]
>
> Generating functions add another string to your bow. Although giving a simple formula for the members of the sequence may be out of the question, we might be able to give a simple formula for the sum of a power series, whose coefficients are the sequence that were looking for.

What Wilf is saying is that if we want to understand a sequence $\{a_n\}_{n=1}^{\infty}$ (or functions $f(x)$ for that matter), then we can do this by studying instead a *function that is described by this sequence (or function).* This way, we may be able to find a nice generating function of this sequence. We can then apply the tools that we have for functions to the generating function and learn something about the sequence. For instance, we know how to take the derivative of a function, and this will turn out to be a very useful aspect. Indeed, let's see what happens when we take the derivative of the moment generating function:

$$\frac{\mathrm{d}}{\mathrm{d}t}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}e^{tX}\right] = \mathbb{E}\left[Xe^{tX}\right].$$

[45] Herbert S. Wilf, *Generatingfunctionology*, 1990, Academic Press.

The equation on the right-hand side[46] may not seem much easier than the equation on the left-hand side, but it will be when we evaluate it at $t = 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\Big|_{t=0} = \mathbb{E}[X],$$

so the first derivative of $M_X(t)$ evaluated at $t = 0$ gives us the *first moment of* X.

We can continue differentiating:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[X\mathrm{e}^{tX}\right] = \mathbb{E}\left[X\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{tX}\right] = \mathbb{E}\left[X^2\mathrm{e}^{tX}\right].$$

Now, if we evaluate the second derivative of $M_X(t)$ at $t = 0$ we get

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t)\Big|_{t=0} = \mathbb{E}[X^2],$$

so the second derivative of $M_X(t)$ evaluated at $t = 0$ gives us the *second moment of* X.

We can of course keep on differentiating $M_X(t)$ as often as we like, and we will end up with the following

Principle 36 [The moment generating function of X describes the moments of X]. *Under some mild assumptions on* X *(see footnote 46), we have for all* $n \geq 1$,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t)\Big|_{t=0} = \mathbb{E}[X^n].$$

The true utility of the moment generating function comes from the fact that it allows us to determine the moments of a distribution via differentiation, rather than via integration, and it is often much easier to differentiate that to integrate. As an added bonus, we don't even need to understand the derivatives of $M_X(t)$ for all values of t, we just need to know it for one particular, easy value: when $t = 0$.

Let's examine the moment generating functions for a couple of common distributions.

Example 40 (The moment generating function of a binomial distribution). If $X \sim \mathrm{Bin}(n, p)$, then

$$\begin{aligned} M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \mathrm{e}^{tk} \\ &= \sum_{k=0}^{n} \binom{n}{k} (p\mathrm{e}^t)^k (1-p)^{n-k} \\ &= \left(p\mathrm{e}^t + (1-p)\right)^n, \end{aligned}$$

where the final equality is due to the binomial theorem (see page 8).

[46] We have cheated a little bit in the derivation: we have interchanged the order of expectation and differentiation. This is not always possible, but it is possible for almost any reasonable probability distribution. In particular, all the distributions that we have considered so far have this property. We will not check this for every distribution that we come across, but rather assume that this is the case in this course.

Now we can easily calculate the moments of X. Differentiating once (by making use of the chain rule with $u = pe^t + (1-p)$) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}M_X(t) = \frac{\mathrm{d}u}{\mathrm{d}t} \cdot \frac{\mathrm{d}}{\mathrm{d}u}u^n = pe^t \cdot nu^{n-1} = npe^t\left(pe^t - (1-p)\right)^{n-1},$$

so that

$$\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t}M_X(t)\Big|_{t=0} = np,$$

as expected. △

Example 41 (The m.g.f. of a Poisson random variable). If $X \sim \mathrm{Poi}(\lambda)$, then

$$\begin{aligned} M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] &= \sum_{n=0}^{\infty} \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!}\mathrm{e}^{tn} \\ &= \mathrm{e}^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda \mathrm{e}^t)^n}{n!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda \mathrm{e}^t} = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}. \end{aligned}$$

We can again determine the first derivative of $M_X(t)$ (using the chain rule with $u = \lambda(\mathrm{e}^t - 1)$)

$$\frac{\mathrm{d}}{\mathrm{d}t}M_X(t) = \frac{\mathrm{d}u}{\mathrm{d}t} \cdot \frac{\mathrm{d}}{\mathrm{d}u}\mathrm{e}^u = \lambda \mathrm{e}^t \cdot \mathrm{e}^{\lambda u} = \lambda \mathrm{e}^{\lambda(\mathrm{e}^t-1)+t},$$

and the second derivative (using the chain rule with $v = \lambda(\mathrm{e}^t - 1) + t$)

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}\lambda \mathrm{e}^{\lambda(\mathrm{e}^t-1)+t} = \frac{\mathrm{d}v}{\mathrm{d}t} \cdot \frac{\mathrm{d}}{\mathrm{d}v}\lambda \mathrm{e}^v = \lambda(\lambda \mathrm{e}^t + 1)\mathrm{e}^{\lambda(\mathrm{e}^t-1)+t}.$$

Evaluating the derivatives at $t = 0$ gives

$$\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t}M_X(t)\Big|_{t=0} = \lambda \mathrm{e}^{\lambda(1-1)+0} = \lambda$$

and

$$\mathbb{E}[X^2] = \frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t)\Big|_{t=0} = \lambda(\lambda+1)\mathrm{e}^{\lambda(1-1)+0} = \lambda(\lambda+1) = \lambda^2 + \lambda.$$

△

PRINCIPLE 37 [The m.g.f. of a standard normal random variable]. *If $Z \sim \mathcal{N}(0,1)$, then*

$$M_Z(t) = \mathrm{e}^{t^2/2}.$$

Proof.

$$\begin{aligned} M_Z(t) = \mathbb{E}[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{tx} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2+t^2/2} dx, \end{aligned}$$

where in the last step we again used the method of 'completing the square' to write

$$x^2 - 2tx = x^2 - 2tx + t^2 - t^2 = (x-t)^2 - t^2$$

(see footnote 44 on page 89). We continue, using the substitution $u = x - t$:

$$\begin{aligned} M_Z(t) &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du \\ &= e^{t^2/2}, \end{aligned}$$

where in the last step we used that the integral is equal to $\sqrt{2\pi}$ (see page 69). □

Example 42 (Additional properties of the m.g.f. of a standard normal / not treated in class). We can again differentiate (using the chain rule with $u = t^2/2$):

$$\frac{d}{dt} M_Z(t) = \frac{du}{dt} \cdot \frac{d}{du} e^u = t e^{t^2/2},$$

and we can differentiate again (now using the product rule with $u = e^{t^2/2}$ and $v = t$):

$$\frac{d^2}{d^2 t} M_Z(t) = v \frac{du}{dt} + u \frac{dv}{dt} = t^2 e^{t^2/2} + e^{t^2/2} = (t^2+1) e^{t^2/2}.$$

We can continue in this fashion (using always the product rule with $u = e^{t^2/2}$ and v equal to the other factor):

$$\begin{aligned} \frac{d^3}{d^3 t} M_Z(t) &= v \frac{du}{dt} + u \frac{dv}{dt} = (t^2+1) t e^{t^2/2} + 2t e^{t^2/2} = (t^3+3t) e^{t^2/2}, \\ \frac{d^4}{d^4 t} M_Z(t) &= v \frac{du}{dt} + u \frac{dv}{dt} = (t^3+3t) t e^{t^2/2} + (3t^2+3) e^{t^2/2} = (t^4+6t+3) e^{t^2/2}. \end{aligned}$$

We can thus determine the first four moments of Z:

$$\begin{aligned}\mathbb{E}[Z] &= \frac{\mathrm{d}}{\mathrm{d}t}M_Z(t)\Big|_{t=0} = 0,\\ \mathbb{E}[Z^2] &= \frac{\mathrm{d}^2}{\mathrm{d}^2t}M_Z(t)\Big|_{t=0} = 1,\\ \mathbb{E}[Z^3] &= \frac{\mathrm{d}^3}{\mathrm{d}^3t}M_Z(t)\Big|_{t=0} = 0,\\ \mathbb{E}[Z^4] &= \frac{\mathrm{d}^4}{\mathrm{d}^4t}M_Z(t)\Big|_{t=0} = 3.\end{aligned}$$

More generally we can determine that

$$\mathbb{E}[Z^n] = \begin{cases}(n-1)!! & \text{if } n \text{ is even,}\\ 0 & \text{if } n \text{ is odd,}\end{cases}$$

where $n!!$ denotes the *double factorial* which is the product of every *odd* number between n and 1. △

Principle 38. *If X and Y are random variables such that $Y = aX + b$, then*

$$M_Y(t) = \mathrm{e}^{tb}M_X(at).$$

Proof.

$$M_Y(t) = \mathbb{E}[\mathrm{e}^{tY}] = \mathbb{E}[\mathrm{e}^{t(aX+b)}] = \mathbb{E}[\mathrm{e}^{(ta)X}\mathrm{e}^{tb}] = \mathrm{e}^{tb}M_X(at). \quad \square$$

This principle has the following consequence:

Corollary 1 [The m.g.f. of a normal random variable]. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$M_X(t) = \mathrm{e}^{(t\sigma)^2/2+t\mu}.$$

Proof. This follows immediately from the principle since $X = \mu + \sigma Z$ if $Z \sim \mathcal{N}(0,1)$, and $M_Z(t) = \mathrm{e}^{t^2/2}$. □

8.1. **Moment generating functions of sums of random variables.** Another nice thing about the moment generating function approach is that it offers us an easy way of studying the sums of independent random variables:

Principle 39. *Let X and Y be independent random variables, then*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Proof. By the fact that $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ for independent X and Y (see Principle 28 on page 82),

$$M_{X+Y}(t) = \mathbb{E}[\mathrm{e}^{t(X+Y)}] = \mathbb{E}[\mathrm{e}^{tX}\mathrm{e}^{tX}] = \mathbb{E}[\mathrm{e}^{tX}]\mathbb{E}[\mathrm{e}^{tY}] = M_X(t)M_Y(t). \quad \square$$

Question 51. *Suppose $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ are independent. What is the distribution of $X + Y$?*

Solution. We use the m.g.f. approach:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = (pe^t + 1 - p)^n(pe^t + 1 - p)^m = (pe^t + 1 - p)^{n+m},$$

which is of course the m.g.f. of a random variable distributed as $\mathrm{Bin}(n+m, p)$. It is a bit too much to simply conclude from this that since the m.g.f. of $X+Y$ and of a $\mathrm{Bin}(n+m, p)$ are the same, it holds that $X+Y \sim \mathrm{Bin}(n+m, p)$ since we have not ruled out the possibility that two random variables with different distributions can have the same m.g.f. This can in fact not happen,[47] but we will not prove this (see also Lemma 1 below). ◊

Question 52 [Not treated in class]. *Suppose $X \sim Poi(\lambda)$ and $Y \sim Poi(\theta)$ are independent. What is the distribution of $X+Y$?*

Solution. We use the m.g.f. approach again:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)}e^{\theta(e^t-1)} = e^{(\lambda+\theta)(e^t-1)},$$

which is of course the m.g.f. of a random variable distributed as $\mathrm{Poi}(\lambda+\theta)$. ◊

Question 53 [Not treated in class]. *Suppose $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent. What is the distribution of $X+Y$?*

Solution. We use the m.g.f. approach again:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\sigma_1^2 t^2/2+\mu_1 t}e^{\sigma_2^2 t^2/2+\mu_2 t} = e^{(\sigma_1^2+\sigma_2^2)t^2/2+(\mu_1+\mu_2)t},$$

which is of course the m.g.f. of a random variable distributed as $\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. We already knew this, since it is the content of Principle 33, but observe how much easier this proof is. ◊

It should be noted here that the previous three examples were all for random variables that are closed under convolution, i.e., for random variables X and Y that have the property that $X+Y$ belongs to the same distributional family as X and Y. This is often not the case (for instance, it is not true for uniform random variables, as we proved in Example 38).

[47] Well, to be completely honest, it is possible for two random variables with different distributions to have the same moment generating function, but all known examples are fairly pathological random variables made up by mathematicians in an effort to find a counter example to the claim. Fortunately for us, all the random variables that we study can be uniquely characterized by their moment generating functions.

9. Limit theorems

In this section, we will prove and examine two of the main results of probability theory: the *weak law of large numbers* (WLLN) and the *central limit theorem* (CLT). Sheldon Ross takes a somewhat technical approach to the proofs. In my opinion, this make it harder to understand the underlying ideas, which are deep (especially the CLT is a profound result).[48] The big ideas are more important than the technical details (again, my opinion), so we are going to take a more heuristic approach to the proofs, starting with the WLLN:

9.1. The weak law of large numbers.

We start with an important notion:

Definition 25 [Convergence in probability]. *A sequence of random variables* $X_1, X_2, \ldots$ *converges in probability to a random variable* Y *if for every* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - Y| \geq \varepsilon) = 0.$$

Convergence in probability is a pretty strong statement to make about random variables, because it implies that as the sequence $X_1, X_2, \ldots$ progresses, the "error" between X_n and Y becomes smaller and smaller (although note that our definition does not require that the error is zero for any finite n.[49]

Example 43. You take up a bow and start shooting arrows at a target. You aim for the middle of the target, which is worth 10 points. Write X_n for the score of the nth shot. As you shoot more and more arrows, you become more experienced, and as a result, you hit the middle more frequently. In the mathematically idealized setting where you continue to shoot and improve ad infinitum, the sequence $X_1, X_2, \ldots$ will converge in probability to (the not particularly) random variable $Y = 10$. △

Example 44. A scientist measures the weight of a single bacterium whose true weight is given by a random variable Y. Because there is a statistical error each the scientist weighs the bacterium, the scientist weighs the bacterium many times in the hopes of *avering out* the statistical errors. She writes X_n for the average of the first n measurements. Again, in the mathematically idealized setting where the scientist continues measuring ad infinitum, the sequence $X_1, X_2, \ldots$ will converge in probability to Y. △

In this second example we are getting a bit ahead of ourselves. In fact, we only know this kind of averaging results because of the weak law of large numbers:

[48] This is not to say that you shouldn't also study Ross's approach: two viewpoints will teach you more than one.

[49] Soon we will learn of another form of convergence of random variables which is considerably weaker.

Theorem 3 [The weak law of large numbers]. *Suppose that $X_1, X_2, \dots$ is a sequence of* independent and identically distributed *random variables with finite means $\mathbb{E}[X_i] = \mu$ and* $\mathrm{Var}(X_i) = \sigma^2$. *Let*

$$\bar{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$$

denote the sample mean *of the first n random variables. Then, $\bar{X}_n$ converges in probability to μ, i.e., for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) = 0.$$

The WLLN states that the expectation of the average over many identically distributed random variables converges to its mean.[50] This essentially justifies our interpretation of the expectation as the long-run average.

The phrase "independent and identically distributed" is a common term in probability, so it is often abbreviated as "*i.i.d.*" We will follow this convention.

Although it is not necessary, we will assume that the variances of the X_i's and of Y are all finite. With this assumption, the proof of the weak law of large numbers is rather easy:

Proof. We determine $\mathbb{E}[\bar{X}_n]$ and $\mathrm{Var}(\bar{X}_n)$ and then we apply Chebychev's inequality. Starting with $\mathbb{E}[\bar{X}_n]$: using the fact that all X_i's are identically distributed,

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \mu.$$

Now we can determine $\mathrm{Var}(\bar{X}_n)$ (using the fact that the X_i's are i.i.d.)

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Observe that the mean of $\bar{X}_n$ is the same for all n, whereas the variance of X_n gets smaller and smaller as n increases.

Now we apply Chebychev's inequality (see Principle 20 on page 49):

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0. \quad \square$$

[50] As the name *weak* law of large numbers suggests, there also exists a *strong* law of large numbers (SLLN). The SLLN states the same as the WLLN, but with the limit *inside* the probability, i.e., it states that

$$\mathbb{P}\left(\lim_{n\to\infty}(X_1 + \cdots + X_n)/n \neq \mu\right) = 0.$$

The proof of the SLLN is significantly more difficult than the proof of the WLLN, and we have not developed all the required tools during the course, so we will not treat the SLLN in more detail.

9.2. The central limit theorem. The WLLN is a great result: it justifies our ideas about expectation as the long-run average, and it has it's applications in statistics besides this. But compared to the next result, the WLLN looks a bit less impressive. We are about to prove the most important result of probability theory bar none. In terms of magnitude of consequence and philosophical depth, you could compare de Laplace's central limit theorem to Darwin's theory of evolution, or Einstein's theory of relativity.[51] It's that big.

Before we get to the central limit theorem, we need some defintions and lemmas[52] We start with a definition:

DEFINITION 26 [Convergence in distribution]. *Suppose $X_1, X_2, \ldots$ are random variables with c.d.f.'s $F_{X_1}, F_{X_2}, \ldots$. Let Y be random variable with c.d.f. F_Y. We say that X_n* converges in distribution *to Y if*

$$\lim_{n\to\infty} F_{X_n}(z) = F_Y(z)$$

for every z where $F_Y(z)$ is continuous.

Observe that convergence in distribution is a significantly weaker notion that convergence in probability, because even though the c.d.f.'s may converge, the associated random variables may fluctuate.

Example 45. Let $Y_1, Y_2, \ldots$ be i.i.d. Unif$[0,1]$ random variables. Let $X_n = \min\{Y_1, \ldots, Y_n\}$. Since $\{X_n > z\}$ means that $\{Y_i > z\}$ for all $i = 1, \ldots, n$, and all Y_i are independent, it follows that

$$\begin{aligned} F_{X_n}(z) &= \mathbb{P}(X_n \le z) = 1 - \mathbb{P}(X_n > z) \\ &= 1 - \mathbb{P}(\{Y_1 > z\} \cap \{Y_2 > z\} \cap \cdots \cap \{Y_n > z\}) \\ &= 1 - \mathbb{P}(Y_1 > z)\mathbb{P}(Y_2 > z)\cdots\mathbb{P}(Y_n > z) \end{aligned}$$

Now since the Y_i are i.i.d., and since $\mathbb{P}(Y_i > z) = (1-z)$ for $0 \le z \le 1$,

$$F_{X_n}(z) = \begin{cases} 0 & \text{when } z < 0, \\ 1 - (1-z)^n & \text{when } 0 \le z \le 1, \\ 1 & \text{when } z > 1. \end{cases}$$

Taking the limit $n \to \infty$ and using that $\lim_{n\to\infty}(1-z)^n = 0$ for all $0 < z \le 1$ we get

$$\lim_{n\to\infty} F_{X_n}(z) = \begin{cases} 0 & \text{when } z < 0, \\ 1 & \text{when } z > 0. \end{cases}$$

This is the c.d.f. of the (not very) random variable $Y = 0$, so it follows that X_n converges in distribution to 0. △

[51] Einstein, in fact, did not receive his Nobel prize for his theory of relativity, but rather for his explanation of the photoelectric effect and for his explanation of the random motion of particles in a solution as proof that molecules exist. This latter result makes crucial use of the CLT.

[52] A lemma is a helper theorem.

We will use the following technical lemma as well:

Lemma 1 [A criterion for convergence in distribution]. *Suppose that $X_1, X_2, \dots$ are random variables with moment generating functions $M_{X_1}(t), M_{X_2}(t), \dots$ and suppose that Y is a random variable with m.g.f. $M_Y(t)$. If*

$$\lim_{n\to\infty} M_{X_n}(t) = M_Y(t) \qquad \text{for all } t \in \mathbb{R},$$

then X_n converges to Y in distribution.

We will not prove this lemma, because the proof is too technical and difficult. The lemma makes intuitive sense though: If we know the moment generating function, then we know an *infinite* number of moments, which is rather a large amount of information. And moreover, all those infinite moments come just from the shape of the m.g.f. at the point $t = 0$, but we know the m.g.f. for all $t \in \mathbb{R}$, so it is not hard to imagine that with some cleverness, we can figure out the entire distribution just by looking at the m.g.f.

Now let's recall a fundamental theorem from calculus:

Theorem 4 [Taylor's theorem]. *Let k be an integer, and let the function $f : \mathbb{R} \to \mathbb{R}$ be at least k times differentiable at the point $a \in \mathbb{R}$. Write $f^{(n)}(a)$ for the nth derivative of $f(x)$ evaluated at a. Then there exists a function $h_k(x)$ such that*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k,$$

and $\lim_{x\to a} h_k(x) = 0$.

(Exercise: check that Taylor's theorem holds for $f(x) = \mathrm{e}^x$ and $a = 0$.)

We will also use the following simple lemma:[53]

[53] Lemmas 1 and 2 have the following nice consequence: together they can be used to reprove that a binomial r.v. converges to a Poisson r.v. in distribution:

Recall that the m.g.f. of $X_n \sim \mathrm{Bin}(n, p)$ is given by $M_{X_n}(t) = (p\mathrm{e}^t + 1 - p)^n$. We write $p = \lambda/n$ and take the limit $n \to \infty$:

$$\lim_{n\to\infty} M_{X_n}(t) = \lim_{n\to\infty}\left(\frac{\lambda}{n}\mathrm{e}^t + 1 - \frac{\lambda}{n}\right)^n = \lim_{n\to\infty}\left(1 + \frac{\lambda(\mathrm{e}^t - 1)}{n}\right)^n = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)},$$

where in the last step we applied Lemma 2 with $x = \lambda(\mathrm{e}^t - 1)$. Note that the function on the right-hand side is the m.g.f. of a Poisson r.v., so by Lemma 1, $X_n \to \mathrm{Poi}(\lambda)$ in distribution.

Lemma 2. *For any $x \in \mathbb{R}$ we have*

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

Moreover, for any $x \in \mathbb{R}$, and any $h(x,n)$ such that for all x,

$$\lim_{n\to\infty} h(x,n) = 0,$$

we have

$$\lim_{n\to\infty}\left(1+\frac{x}{n}+\frac{h(x,n)x}{n}\right)^n = \mathrm{e}^x.$$

Here is a 'lazy proof' the first claim, the second follows the same approach.

Lazy proof. We start taking the exponent of the log of the left-hand side:

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

Now we use the approximation[54] $\log(1+x/n) \approx x/n$ to obtain

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^{\lim_{n\to\infty} n\frac{x}{n}} = \mathrm{e}^x. \quad \bigcirc$$

Now we are finally ready to state the central limit theorem:

[54] This is why it's a lazy proof. The approximation of course depends on the fact that the higher order terms $x^2/(2n^2), x^3/(3n^3)$, etc. of the expansion are small compared to x/n. Here is the actual proof (not treated in class):

Actual proof. We start with the same steps, but then we write $t = 1/n$, so that we can take the limit $t \to 0$ instead:

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^{\lim_{n\to\infty} n\log\left(1+\frac{x}{n}\right)} = \mathrm{e}^{\lim_{t\to 0}\frac{1}{t}\log(1+tx)}.$$

We can apply L'Hôpital's rule (which states $\lim_{x\to a}\frac{f(x)}{g(x)} = \lim_{x\to a}\frac{f'(x)}{g'(x)}$) to the exponent:

$$\lim_{t\to 0}\frac{1}{t}\log(1+tx) = \lim_{t\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\log(1+tx)}{\frac{\mathrm{d}}{\mathrm{d}t}t} = \lim_{t\to 0}\frac{\frac{x}{1+tx}}{1} = \lim_{t\to 0}\frac{x}{1+tx}.$$

Now we use that the limit of a ratio is equal to the ratios of the limits:

$$\lim_{t\to 0}\frac{x}{1+tx} = \frac{\lim_{t\to 0} x}{\lim_{t\to 0} 1+xt} = \frac{x}{1} = x.$$

Therefore,

$$\mathrm{e}^{\lim_{t\to 0}\frac{1}{t}\log(1+tx)} = \mathrm{e}^x,$$

and this concludes the proof. □

Theorem 5 [The central limit theorem]. *Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and* $\mathrm{Var}(X_i) = \sigma^2$. *Let $S_n := X_1 + \cdots + X_n$, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \longrightarrow \mathcal{N}(0,1) \qquad \textit{in distribution.}$$

In other words,

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x = \Phi(b) - \Phi(a).$$

The most important point to note about the CLT is that *any* distribution for X_i will yield the *same* limiting distribution for S_n (as long as μ and σ^2 are finite). That is to say, the X_i can be Bernoulli random variables, or exponential random variables, or even some kind of distribution that we've never heard of: it doesn't matter for S_n. Moreover, there are versions of the CLT (which we won't discuss during this course) that show that the convergence can even hold if the X_i are not independent (if the dependence is weak enough), or if the X_i's are not identically distributed (if the distributions are not too dissimilar).

Before we delve deeper into the consequences of the CLT, let's first prove it:

Proof. We start by writing

$$Y_i := \frac{X_i - \mu}{\sigma}$$

so that Y_i are i.i.d. with $\mathbb{E}[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$. Now we assume that $M_{Y_i}(t)$ exists for all t (this is not always true, but again, the precise proof is much more technical). We apply Taylor's theorem with $k = 2$ and $a = 0$ to $M_{Y_i}(t)$:

$$\begin{aligned} M_{Y_i}(t) &= M_{Y_i}(0) + M'_{Y_i}(0)t + \frac{1}{2}M''_{Y_i}(0)t^2 + h_2(t)t^2 \\ &= 1 + \mathbb{E}[Y_i]t + \frac{1}{2}\mathbb{E}[Y_i^2]t^2 + h_2(t)t^2 \\ &= 1 + 0 + \frac{1}{2}t^2 + h_2(t)t^2. \end{aligned}$$

Now we write

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

To prove the CLT, we need to show that $U_n \to \mathcal{N}(0,1)$ in distribution, so by Lemma 1 it is enough to show that for all $t \in \mathbb{R}$,

$$M_{U_n}(t) = \mathrm{e}^{t^2/2} \qquad \text{when} \qquad n \to \infty.$$

We know from Principle 38 that $M_{aX}(t) = M_X(at)$ and we know from Principle 39 that $M_{X+Y}(t) = M_X(t)M_Y(t)$ if X and Y are independent, so

$$M_{U_n}(t) = M_{\sum_{i=1}^n Y_i}(t/\sqrt{n}) = \left(M_{Y_1}(t/\sqrt{n})\right)^n = \left(1 + \frac{1}{2}\frac{t^2}{n} + h_2\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n}\right)^n.$$

All that remains is to take the limit as $n \to \infty$. To determine what this limit is we apply Lemma 2 with $x = \frac{t^2}{2}$ and with $h(x,n) = 2h_2(\sqrt{2x/n})$ (which satisfies $\lim_{n\to\infty} h(x,n) = 0$ by Taylor's theorem) to obtain

$$\lim_{n\to\infty} M_{U_n}(t) = \left(1 + \frac{t^2}{2n} + h_2\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n}\right)^n = e^{t^2/2}. \quad \square$$

We will conclude with some examples of the CLT:

Question 54. *An airline company knows that about* 5% *of the people who buy a ticket don't show up for the flight, so they figure that they may as well sell a few more tickets than there are seats in the plane. They decide that it is good enough to be* 99% *sure that there are more available seats than passengers for any given flight. How many tickets should the airline sell for a flight that seats 400 passengers?*

Solution. Suppose the airline sells n tickets. Let $X_i = 1$ if the ith ticket holder shows up, and let $X_i = 0$ if the ticket holder does not show up. Then it follows that $X_i \sim \text{Ber}(p = 0.95)$. We will assume that all the X_i's are independent (this is a dubious assumption, considering people often travel in groups, but ok). Let $S_n = \sum_{i=1}^n$ denote the number of people who show up for the flight, so $S_n \sim \text{Bin}(n, p = 0.95)$. We want to determine n such that

$$\mathbb{P}(S_n \le 400) \approx 0.99.$$

We know that $\mathbb{E}[S_n] = np = 0.95n$ and $\text{Var}(S_n) = np(1-p) = 0.0475n$, so that $\sigma \approx 0.22\sqrt{n}$. Now, by the CLT,

$$0.99 \approx \mathbb{P}(S_n \le 400) = \mathbb{P}\left(\frac{S_n - np}{\sigma\sqrt{n}} \le \frac{400 - np}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{400 - np}{\sigma\sqrt{n}}\right).$$

From the table for the standard normal c.d.f. on page 73 we can determine that $\Phi(2.33) \approx 0.99$, so we should solve

$$\frac{400 - np}{\sigma\sqrt{n}} = 2.33.$$

This can be solved to give us that $n = 410$, so the airline can quite confidently sell 10 extra tickets for the flight. ◊

Question 55 [Simple random walk on the integers]. *To model the random motion of a single particle (e.g. a molecule) in a solution, consider the following model: we let the number line* $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ *describe the* x*-coordinate of the particle. At time* $t = 0$ *a particle sits at the point* $x = 0$*. The particle sits there until time* $t = 1$*, when it jumps. The* *particle either jumps one step to the left, to* $x = -1$, *or it jumps to the right, to* $x = 1$. *Both jumps are equally likely. At time* $t = 2$ *the particle again jumps one step in a random direction, and so forth. What is the probability that at* $t = 1000$ *the particle is more than 100 steps removed from its starting point?*

Solution. We write X_i for the increment of the ith step, i.e., $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$. Then $\mu = \mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mu^2 = 1$. The position of the particle at time $t = 1000$ is given by S_{1000}, so we want to determine

$$\mathbb{P}(|S_{1000}| > 100).$$

By the CLT and since $\sqrt{1000} \approx 31.6$,

$$\begin{aligned}\mathbb{P}(|S_{1000}| > 100) = 2\mathbb{P}(S_{1000} > 100) &= 2\mathbb{P}\left(\frac{S_{1000} - n\mu}{\sigma\sqrt{1000}} \geq \frac{100 - n\mu}{\sigma\sqrt{1000}}\right)\\ &= 2\left(1 - \Phi\left(\frac{100}{31.6}\right)\right) = 2(1 - \Phi(3.16)) = 2(1 - 0.9993) = 0.0014.\end{aligned}$$

The probability that the particle travels more than three standard deviations from the mean is extremely unlikely. ◊

Some concluding remarks

The CLT is a logical ending point for a course that aims at introducing probability theory. We have seen how to define probabilities, how to determine the probability of a given event, how to take extra information into account, how to deal with the random outcomes of experiments, and what happens to large averages of such outcomes. We have used all kinds of tools, from complex counting arguments to the highly abstracted logic of moment generating functions.

In front of us now lies a vast number of applications and extensions of the theories that we have studied. Many of these extensions are somehow related to the model described in the final example. Simple random walk (or SRW) is what is known as a *stochastic process.* Stochastic processes are models that describe how a system develops randomly over time. The position of a randomly jumping particle is the prime example of such a process. There are many more such processes, and they have a wide range of applications:

- *Brownian motion* is the continuous variant of SRW. Instead of jumping to a new direction at each time step, a Brownian motion wiggles around without ever jumping. Brownian motion has many successful applications, for instance in physics and chemistry, because it accurately describes the motion of particles in a solution. Unfortunately, there are also less successful applications. Black and Scholes famously used Brownian motion to model stock market fluctuations. Stock markets do jump, however, and their inaccurate model is viewed as one of the main causes of the stock market crash in 2008.[55]
- *Branching processes* are processes that can be used to model asexual reproduction. An individual gives birth to a random number of children, who in turn give birth to their own random number of children, etc. We can use such models to predict for instance extinction, and with some modification, we can use them to model the spread of evolutionary traits through a population.
- *Random networks* can be used to model the growth and structure of large, real life networks, such as the facebook network or the internet. We can for instance test the 'six degrees of separation' hypothesis.[56] It turns out that in a network with N nodes, the maximal distance is often between $\log(\log(N))$ and $\log(N)$, so the 6DoS hypothesis seems plausible.
- *Queues* are mathematical models that describe how a system deals with serving customers or requests. A queue model describes the situation where customers (or requests) come in randomly, and service times are also random. Questions that we want to answer are: how many customers will be helped in a certain amount of time, and how many servers should we hire/install?

[55] Stock markets can be modelled much better by Lévy processes, which are Brownian motions with jumps.

[56] The 6DoS hypothesis states that any randomly chosen person on this planet is at most a friend-of-a-friend-of-a-friend-of-a-friend-of-a-friend-of-a-friend-of-a-friend of yours.

- *Card shuffling* is a problem that has been getting a lot of attention recently. The questions are simple: for a given type of shuffling, how long do we need to shuffle a deck of cards before it is properly randomised? And can we even tell whether a deck of cards has been shuffled properly? These questions are surprisingly hard to answer. Card shuffling may seem like a rather trivial problem, but these questions have strong ties to problems in computer science.

Besides these processes, there are many other fields where knowledge of probability theory is useful. The obvious one is of course statistics, which is the science of large sets of random numbers. Quantum theory is another field that heavily uses probability theory, now to describe the totally random nature of the smallest particles. But even in cases where there seems to be no randomness at all, probability can still be useful. For instance, the very serious mathematical study of prime numbers[57] uses a lot of probability theory. Somehow, even though the prime numbers are of course a completely deterministic set, their distribution on the number line so much resembles a random mess that the probabilistic arguments that we studied can be used to answer big questions about the primes. In the spring of 2013 Yitang Zhang used probabilistic methods (along with a lot of other high level math) to show that there are infinitely many prime numbers p and q such that $|p-q| \leq 70,000,000$. This may not seem like a big deal, but mathematicians it is: this problem had been open for a hundred and fifty years, and was generally considered to be one of the hardest problems in mathematics.[58]

I guess what I'm trying to say is, there are still plenty more applications of probability theory for you to discover, and I hope you will.

[57] Numbers only divisible by one and themselves.

[58] The twin primes conjecture states that there are infinitely many primes p and q such that $|p-q| = 2$. The proof of Zhang has a much higher bound, and the fact that it is $70,000,000$ is not particularly important (it has since been reduced to 576), the amazing fact is that we now have a proof for some finite number.