BSM Probability notes

Péter Bálint

December 14, 2020

Abstract

These notes discuss the material for the Probability course at Budapest Semesters in Mathematics. The course textbook, Ross: A first course in probability, is frequently referred.

1 Probability spaces

The mathematical framework for probability is the probability space, which is a triple $(\Omega, \mathcal{F}, \mathbb{P})$. Let us consider first the first ingredient, the *sample space* Ω, the collection of possible outcomes of our random experiment. Elements $\omega \in \Omega$ are called outcomes, while (certain) subsets $A \subset \Omega$ are called events.

Examples of sample spaces.

I Horse race with three horses, Apple, Banana and Coconut. An outcome describes the order of the horses at the end of the race.

$$\Omega_I = \{(a,b,c),(a,c,b),\dots(c,b,a)\}; \qquad |\Omega_I| = 3! = 6$$

where $|E|$ denotes the cardinality of a set. An example of an event is

$$A_I = \{\text{Apple wins}\} = \{(a,b,c),(a,c,b)\}; \qquad |A_I| = 2.$$

Note that we have not claimed anything about the "probability" of this event.

II Two fair dice are rolled, a white die and a yellow die. An outcome is an ordered pair of the values rolled on the two dice, thus

$$\Omega_{II} = \{(1,1),(1,2),\dots(6,6)\}; \qquad |\Omega_{II}| = 36.$$

Note that $(1,2)$ and $(2,1)$ are two different outcomes. Examples of events are

$$A_{II} = \{\text{The sum of the values is 5}\} = \{(1,4),(2,3),(3,2),(4,1))\},$$
$$B_{II} = \{\text{The sum of the values is 4}\} = \{(1,3),(2,2),(3,1)\},$$
$$C_{II} = \{\text{The same values are rolled on the two dice}\} = \{(1,1),(2,2),\dots(6,6)\}.$$

III Keep on flipping a fair coin until a Head occurs. Then

$$\Omega_{III} = \{(H), (T,H), (T,T,H), (T,T,T,H), ...\}$$

which has infinite cardinality. An example of an event is

$$A_{III} = \{\text{There is an even number of flips}\} = \{(T,H), (T,T,T,H), ...\}$$

IV Darts on a table which is a circular disc of radius 10 inches. The outcome corresponds to the point of impact of the dart. Then

$$\Omega_{IV} = \{(x,y) \in \mathbb{R}^2 \,|\, x^2 + y^2 \leq 100\}$$

and an example of an event is

$$A_{IV} = \{\text{Score 50 points}\} = \{(x,y) \in \mathbb{R}^2 \,|\, x^2 + y^2 \leq 1\}.$$

In this framework, the usual set theoretic operations and notations make sense. $E \cup F$ means that at least one of the two events occur, while $E \cap F$ means that both of them occur. In particular, in Example II above

$$B_{II} \cup C_{II} = \{(1,3), (3,1), (1,1), \ldots, (6,6)\},$$
$$B_{II} \cap C_{II} = \{(2,2)\}.$$

If $E \cap F = \emptyset$ then the two events are mutually exclusive, which is the case of A_{II} and C_{II} in Example II. $E^c = \Omega \setminus E$ is the complement of the event E. If $E \subset F$ then the event E implies the event F.

Usual laws of set theoretic operations apply, see the Ross book for a summary. It is useful to point out the de Morgan laws:

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

The second ingredient is the *sigma algebra* $\mathcal{F}$. It is the collection of subsets of Ω that are considered as events. $\mathcal{F}$ has to satisfy the following properties:

- $\Omega \in \mathcal{F}$,

- if $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$,

- if $E_1, E_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$.

In examples I, II and III above, $\mathcal{F}$ can be chosen as the collection of all subsets of Ω. In example IV (if, in accordance with intuition, probability is proportional to area) certain subsets of Ω have to be excluded from $\mathcal{F}$ for deeper measure theoretic reasons, yet, it is hard to construct such subsets.

Axioms of probability

The third ingredient is probability itself, $\mathbb{P} : \mathcal{F} \to \mathbb{R}$, which is thus a function that assigns a number to an event. It has to satisfy the following axioms.

First Axiom $\mathbb{P}(E) \geq 0; \qquad \forall E \in \mathcal{F};$

Second Axiom $\mathbb{P}(\Omega) = 1;$

Third Axiom Given a sequence of mutually exclusive events $E_1, E_2, \cdots \in \mathcal{F}$ (that is, $E_i \cap E_j = \emptyset$ whenever $i \neq j$) we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

These axioms are in accordance with our intuition about probability. Yet, it should be noted that the third axiom applies not only to finite, but also to countably infinite collections of mutually exclusive sets. In this case the sum on the RHS is a limit – thus sigma additivity is actually a continuity property of $\mathbb{P}$.

Here we discuss some direct consequences of the axioms. As whenever $A \cap B = \emptyset$ we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, we have

$$\mathbb{P}(A^c) = \mathbb{P}(\Omega) - \mathbb{P}(A) = 1 - \mathbb{P}(A), \quad \forall A \in \mathcal{F}, \quad \text{and in particular}$$
$$\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 0.$$

Note, however, that $\mathbb{P}(E) = 0$ does not necessarily imply $E = \emptyset$ – consider example *IV* (with probability proportional to area) and the event $E = \{(0, 0)\}$.

Inclusion–exclusion formula

Consider now $A \cap B \neq \emptyset$. Then, by the third axiom:

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

With a similar reasoning, for arbitrary $A, B, C \in \mathcal{F}$:

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

In words, to get the probability of the union, the probabilities of the events have to be added, then the probabilities of all possible pair intersections have to be subtracted, then the probabilities of all triple intersections added... This generalizes to an arbitrary number of sets: the probability of the union can be computed by adding up all odd-fold intersections and subtracting all even-fold intersections. This is expressed in a concise form in the following Proposition, which can be proved by induction on n.

Proposition 1.1 (**Inclusion–exclusion formula**). *Let $n \geq 1$, and $E_1, E_2, \ldots E_n \in \mathcal{F}$. Then*

$$\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathbb{P}(E_{i_1} \cap \cdots \cap E_{i_r}).$$

2 Finite sample spaces with equally likely outcomes

Let us fix an integer $N \geq 1$, and consider probability spaces such that

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\},$$
$$\mathbb{P}(\{\omega_1\}) = \mathbb{P}(\{\omega_2\}) = \cdots = \mathbb{P}(\{\omega_N\}).$$

Then, by the axioms of probability

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\{\omega_1\}) + \cdots + \mathbb{P}(\{\omega_N\}) = N\mathbb{P}(\{\omega_1\}), \quad \Longrightarrow \quad \mathbb{P}(\{\omega_i\}) = \frac{1}{N}, \ \forall i = 1, \ldots, N.$$

Furthermore, for any subset $A \subset \Omega$, ($|A| = K \leq N$), we have

$$\begin{aligned} A =& \{\omega_{j_1}, \omega_{j_2}, \ldots, \omega_{j_K}\}, \\ \mathbb{P}(A) =& \mathbb{P}(\{\omega_{j_1}\}) + \cdots + \mathbb{P}(\{\omega_{j_K}\}) = \frac{K}{N}, \\ =& \frac{|A|}{|\Omega|} = \frac{\text{number of favored outcomes}}{\text{total number of outcomes}}. \end{aligned}$$

This formula may be familiar from high school, however, it has to be applied carefully. Before making computations, decide what is your sample space. Several choices may be suitable for the same problem, nonetheless, once Ω has been chosen, use it consistently.

Example 2.1. *There are* 11 *balls in a urn,* 6 *red and* 5 *blue balls.* 3 *balls out of this* 11 *are drawn (without replacement). What is the probability that there are exactly* 2 *red balls and* 1 *blue ball among the* 3 *balls drawn?*

Solution #1. Label the balls such that $1, \ldots, 6$ are red, while $7, \ldots, 11$ are blue. Construct Ω by noting the labels of the balls at the three consecutive draws – that is, the order is taken into account. We have

$$\Omega = \{(i_1, i_2, i_3) \mid i_1, i_2, i_3 \in \{1, \ldots, 11\}, \ i_1 \neq i_2, i_1 \neq i_3, i_2 \neq i_3\}$$
$$|\Omega| = 11 \cdot 10 \cdot 9.$$

Also

$$E = \{\text{Exactly one blue ball among the three drawn}\} = E_1 \cup E_2 \cup E_3,$$

where the sets

$$E_k = \{\text{The } k\text{th draw is blue, the other two draws are red}\}; \ k = 1, 2, 3$$

are mutually exclusive. In particular

$$E_1 = \{(i_1, i_2, i_3) \in \Omega \mid i_1 \in \{7, \ldots, 11\}; \ i_2, i_3 \in \{1, \ldots, 6\}\};$$
$$|E_1| = 5 \cdot 6 \cdot 5,$$

and similar computations yield

$$|E_2| = |E_3| = 6 \cdot 5 \cdot 5.$$

We arrive at

$$\mathbb{P}(E) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) = \frac{|E_1| + |E_2| + |E_3|}{|\Omega|} = \frac{3 \cdot 6 \cdot 5 \cdot 5}{11 \cdot 10 \cdot 9} = \frac{5}{11}.$$

Solution #2. The same labeling is used as in the previous solution, however, this time the order of choice is not taken into account, only the collection of the three balls drawn is considered as an outcome. Thus

$$\Omega = \{U \subset \{1, \ldots, 11\} \mid |U| = 3\};$$
$$|\Omega| = \binom{11}{3},$$

while

$$E = \{U_1 \cup U_2 \in \Omega \mid |U_1| = 2, U_1 \subset \{1, \ldots, 6\};\ |U_2| = 1, U_2 \subset \{7, \ldots, 11\}\};$$
$$|E| = \binom{6}{2} \cdot 5.$$

We arrive at

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{\binom{6}{2} \cdot 5}{\binom{11}{3}} = \frac{5}{11}.$$

Example 2.2 (**The matching problem**). *Consider a a great party with N people involved. The following morning, people leave one-by-one, and take one of the cell phones at random. What is the chance that noone picks her/his own phone?*

The sample space Ω is the set of all permutations of the phones labeled $1, \ldots, N$. If the permutation is $\omega = (i_1, \ldots, i_N)$, then person $\#k$ gets the cellphone of person $\#i_k$, $k = 1, \ldots N$. Let

$$A_k = \{\text{person } \#k \text{ picks her own phone}\} = \{(i_1, \ldots, i_N) \in \Omega \,|\, i_k = k\},$$

then

$$B = \{\text{No matches}\} = A_1^c \cap A_2^c \cap \cdots \cap A_N^c = (A_1 \cup \cdots \cup A_n)^c,$$

and we use the inclusion-exclusion formula to compute $\mathbb{P}(A_1 \cup \cdots \cup A_n)$. Now

$$|A_k| = |A_1| = (N-1)! \quad \text{and thus}$$
$$\mathbb{P}(A_k) = \mathbb{P}(A_1) = \frac{(N-1)!}{N!},$$

as the labels can be freely permuted at the remaining $N-1$ positions. Similarly, for arbitrary fixed $1 \le i_1 < i_2 < \cdots <_r \le N$:

$$|A_{i_1} \cap \cdots \cap A_{i_r}| = (N-r)! \quad \text{and thus}$$
$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}) = \frac{(N-r)!}{N!}.$$

Using this along with the inclusion-exclusion formula:

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_N) &= \sum_{r=1}^{N} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le N} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}) \\ &= \sum_{r=1}^{N} (-1)^{r+1} \binom{N}{r} \frac{(N-r)!}{N!} = \sum_{r=1}^{N} \frac{(-1)^{r+1}}{r!}.\end{aligned}$$

Now

$$\mathbb{P}(B) = 1 - \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_N) = \sum_{r=0}^{N} \frac{(-1)^r}{r!}$$

which tends to e^{-1} as $N \to \infty$.

3 Conditional probability

Introductory question. Roll two fair dice, a white die and a yellow die. What is the chance that the sum of the two values rolled is 10? By the previous section the answer to this question is $\frac{3}{36} = \frac{1}{12}$.

Now imagine that the two dice are not simultaneously, but consecutively rolled. The first die has been rolled and turned up 6; the yellow die is yet to be rolled, what is the probability (given the information on the white die) that the sum of the values will be 10? This happens if an only if the yellow die turns up 4, that is, with $\frac{1}{6}$ chance.

An alternative approach to this: the information that

$$A = \{\text{The white die turns up 6}\}$$

occurs is given. What is the chance (conditioned on A occurring) that

$$B = \{\text{The sum of the values is 10}\}$$

occurs, too? Now

$$A = \{(6,1), \ldots, (6,6)\} \qquad |A| = 6 \qquad \mathbb{P}(A) = \frac{1}{6};$$
$$A \cap B = \{(6,4)\} \qquad |A \cap B| = 1 \qquad \mathbb{P}(A \cap B) = \frac{1}{36};$$
$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1}{6}.$$

Definition 3.1. *Fix an event A with $\mathbb{P}(A) > 0$, the conditional probability of any other event $E \in \mathcal{F}$ given A is*

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

The following proposition can be verified by direct inspection. It shows that conditional probability generalizes the notion of probability.

Proposition 3.2. *Fix A with $\mathbb{P}(A) > 0$. Then $\mathbb{P}(\cdot|A) : \mathcal{F} \to \mathbb{R}$; $\mathcal{F} \ni E \mapsto \mathbb{P}(E|A)$ satisfies the axioms of probability.*

This conditional probability can be computed using either the definition, or the reduced sample space, as in the introductory example. Nonetheless, it should be clearly formulated on which event the conditioning takes place, as demonstrated in the following example.

Example 3.3. *Andrew and Bob play for the college basketball team. They get two T shirts each in closed bags. Any T shirt can be either black or white with equal chance. Andrew says:"If I have a black T shirt, I will have this one on." Bob says:"I do not have any preference regarding the color of the T shirt." At the next training, Andrew shows up with a black T shirt on. What is the probability that his other T shirt is black, too? (And consider the same question for Bob.)*

For Andrew, initially, there are four options of equal probability

$$\Omega = \{(W, W), (W, B), (B, W), (B, B)\}$$

where the first and the second entry of the pair describes the color of the T shirt in his first and second bag, respectively. Now given what Andrew says, we condition on the reduced sample space

$$\{(W, B), (B, W), (B, B)\}$$

As he has a black T shirt on, there is only one out of these three options when his other T shirt is black, too. Hence the answer is $\frac{1}{3}$.

How about Bob? By common sense, the fact that he has a black T shirt, does not influence the color of his other T shirt. Accordingly, the expected answer to the question is $\frac{1}{2}$ in Bob's case. But why is it that the logic applied in Andrew's case does not work for Bob's case?

Let us introduce the following events:

$A_1 = \{\text{Andrew has a least one black T shirt}\}$ $\quad$ $B_1 = \{\text{Bob has a least one black T shirt}\}$

$A_2 = \{\text{Both of Andrew's T shirts are black}\}$ $\quad$ $B_2 = \{\text{Both of Bob's T shirts are black}\}$

$A_3 = \{\text{Andrew has a black T shirt on}\}$ $\quad$ $B_3 = \{\text{Bob has a black T shirt on}\}$

Note that $A_1 = A_3$, and in fact, we have computed $\mathbb{P}(A_2|A_1)$ above. However, $B_1 \supset B_3$, but $B_1 \neq B_3$. To distinguish B_1 from B_3, we have to work with an extended sample space.

Bob decides randomly which bag to open, say by flipping a fair coin; in the extended sample space below, all previous outcomes are doubled, and boldface refers to the bag that Bob has picked. Let

$$\Omega_{\text{Bob}} = \{(\mathbf{W}, W), (W, \mathbf{W}), (\mathbf{W}, B), (W, \mathbf{B}), (\mathbf{B}, W), (B, \mathbf{W}), (\mathbf{B}, B), (B, \mathbf{B})\}$$

while

$$B_3 = \{(W, \mathbf{B}), (\mathbf{B}, W), (\mathbf{B}, B), (B, \mathbf{B})\}; \qquad B_2 = \{(\mathbf{B}, B), (B, \mathbf{B})\}$$

so that $\mathbb{P}(B_2 \,|\, B_3) = \frac{2}{4} = \frac{1}{2}$, as anticipated.

Multiplication rule

It follows immediately from the definition of conditional probability that given two events A and B we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A),$$

which reflects a useful point of view; often, it is easier and more natural to compute $\mathbb{P}(A \cap B)$ this way than directly. By induction, for events $E_1, E_2, \ldots E_n$

$$\mathbb{P}(E_n \cap \cdots \cap E_1) = \mathbb{P}(E_n|E_{n-1} \cap \cdots \cap E_1) \cdot \mathbb{P}(E_{n-1}|E_{n-2} \cap \cdots \cap E_1) \cdot \cdots \cdot \mathbb{P}(E_2|E_1)\mathbb{P}(E_1).$$

This can be particularly handy if the events $E_1, \ldots, E_n$ (may) occur consecutively.

Example 3.4 (Pólya's urn model). *Initially, there are two balls in the urn, a blue and a red ball. At each round, a ball in the urn is picked at random. The color of the ball picked is checked, and then this ball and an additional ball of the same color is put back into the urn. This way, the number of balls in the urn keeps growing with the number of the draws.*

Question: What is the probability that the first three balls are blue, red, blue, in this order? We may use the multiplication rule to answer this question as follows. For $k \geq 1$, let B_k denote the event that the kth ball drawn is blue. Similarly, let R_k denote the event that the kth ball drawn is red. Then

$$\mathbb{P}(B_3 \cap R_2 \cap B_1) = \mathbb{P}(B_3|R_2 \cap B_1) \cdot \mathbb{P}(R_2|B_1) \cdot \mathbb{P}(B_1) = \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{12}.$$

4 Bayes formula

Definition 4.1. *The finite collection of events $A_1, \ldots, A_n$ is a partition of the sample space if*

$$A_i \cap A_j = \emptyset \text{ whenever } i \neq j; \qquad \text{and} \qquad A_1 \cup A_2 \cup \cdots \cup A_n = \Omega.$$

You may think of $A_1, \ldots, A_n$ as pizza slices. Given a partition, the probability of an arbitrary event B can be computed as

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

where we have used the axioms of probability and the multiplication rule. The above formula is called **the law of total probability**. Now, for any index $k = 1, \ldots n$ fixed, we have

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(B \cap A_k)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_k)\mathbb{P}(A_k)}{\sum\limits_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

which is called **Bayes formula**. It can be interpreted as follows. The A_i can be thought of as various, mutually exclusive possibilities, that have some a priori chances $\mathbb{P}(A_i)$. Then some information are gained, namely, that B has occurred. Given this information, the probabilities of the A_i have to be updated, and it is precisely Bayes formula that gives the new, a posteriori chances. Here are two examples to demonstrate this.

Example 4.2. *Initially, there is just one blue ball in a box. First, a fair die is rolled. Then, if the number rolled is odd, one red ball is put in the box; if the number rolled is* 2 *or* 4, 3 *red balls are put in the box; and if the number rolled is* 6, 5 *red balls are put into the box. After that, a ball is drawn from the box. When entering the room at the end of this process you are informed that the ball drawn is red. What do you think is the chance that* 6 *was rolled on the die?*

Introduce the following events:

$$\begin{aligned} A_1 &= \{\text{odd number rolled}\} = \{\text{one red ball and one blue ball in the box}\};\\ A_2 &= \{\text{2 or 4 rolled}\} = \{\text{3 red balls and one blue ball in the box}\};\\ A_3 &= \{\text{6 rolled}\} = \{\text{5 red balls and one blue ball in the box}\};\\ B &= \{\text{Eventually, a red ball is drawn}\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(A_1) &= \frac{1}{2} & \mathbb{P}(B|A_1) &= \frac{1}{2};\\ \mathbb{P}(A_2) &= \frac{1}{3} & \mathbb{P}(B|A_2) &= \frac{3}{4};\\ \mathbb{P}(A_3) &= \frac{1}{6} & \mathbb{P}(B|A_3) &= \frac{5}{6}; \end{aligned}$$

thus

$$\mathbb{P}(A_3|B) = \frac{\mathbb{P}(B|A_3)\mathbb{P}(A_3)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2) + \mathbb{P}(B|A_3)\mathbb{P}(A_3)} = \frac{\frac{5}{6} \cdot \frac{1}{6}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{3} + \frac{5}{6} \cdot \frac{1}{6}} = \frac{5}{23}.$$

In the next example there are just two events, denoted by A and $A^c (= \Omega \setminus A)$, which make the partition.

Example 4.3 (**Medical test**). *Consider a blood test for a disease with the following characteristics. The test is* 95% *effective, that is, it is positive with* 0.95 *chance if the person tested is ill. However, with* 1% *chance it is false positive, that is, with* 0.01 *chance positive if the person tested is healthy. Also, the disease is known to effect* 0.5% *– that is,* 0.005 *proportion – of the population. What is the chance that someone tested positive is indeed ill?*

Introduce the following events:

$$\begin{aligned} A &= \{\text{the person tested is ill}\}; \\ A^c &= \{\text{the person tested is healthy}\}; \\ B &= \{\text{the test is positive}\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(A) &= 0.005 & \mathbb{P}(B|A) &= 0.95; \\ \mathbb{P}(A^c) &= 0.995 & \mathbb{P}(B|A^c) &= 0.01; \end{aligned}$$

thus

$$\mathbb{P}(A|B) = \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.995 \cdot 0.001} \approx 0.323$$

which may be surprisingly small at first sight. Note, however, that the a priori chance (before the test) of the person being ill was 0.005, which is updated to 0.323 with one positive test. Still, 0.323 may not be convincing enough. We will discuss how to proceed, but let us introduce the notion of independence before that.

5 Independence

Definition 5.1. *Two events A and B are independent if*

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

Example 5.2. *Draw one card from an ordinary deck of* 52 *cards. Let*

$$\begin{aligned} A &= \{\textit{the card drawn is a spade}\}; \\ B &= \{\textit{the card drawn is an ace}\}; \\ C &= \{\textit{the card drawn is a heart}\}. \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{P}(A) = \mathbb{P}(C) = \frac{13}{52} = \frac{1}{4}; \\ &\mathbb{P}(B) = \frac{4}{52} = \frac{1}{13}; \\ &A \cap B = \{\text{the card drawn is the ace of spades}\}, \text{ hence} \\ &\mathbb{P}(A \cap B) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = \mathbb{P}(A) \cdot \mathbb{P}(B) \end{aligned}$$

which shows that A and B are independent. Similarly, B and C are independent. Note, however

$$A \cap C = \emptyset, \text{ hence}$$
$$\mathbb{P}(A \cap C) = 0 \neq \mathbb{P}(A) \cdot \mathbb{P}(C)$$

so A and C are *not* independent. This is a warning that "mutually exclusive" and "independent" are very different notions. In a sense, mutually exclusive events are as far from being independent as possible, as the following Lemma shows. It can be proved by direct inspection using the definitions and the multiplication rule.

Lemma 5.3. *Consider two events E and F such that $\mathbb{P}(E) > 0$. Then E and F are independent if and only if*

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

In words: the independence of E and F means that the occurrence of E does not shift, that is, does not provide additional information on the chances of the occurrence of F.

To proceed, let us revisit the example with the two fair dice by considering the following events.

Example 5.4. *Roll two fair dice, a white and a yellow die. Let*

$$A = \{\textit{the white die turns up } 6\};$$
$$B = \{\textit{the sum of the two values rolled is } 7\};$$
$$C = \{\textit{the yellow die turns up } 1\}.$$

Direct inspections shows

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{6};$$
$$A \cap B = A \cap C = B \cap C = A \cap B \cap C = \{(6,1)\}.$$

Then $\mathbb{P}(A \cap B) = \frac{1}{36} = \mathbb{P}(A) \cdot \mathbb{P}(B)$, the event A and B are independent. Similarly A and C are independent, and also, B and C are independent. That is, the events A, B and C are pairwise independent. Yet, if both A and B occur, we can be certain that C occurs, too. This is reflected in

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{36} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C).$$

Definition 5.5. *The events $A_1, A_2, \ldots, A_n$ are independent as a collection if for any $r = 1, \ldots, n$ and any indices $1 \leq i_1 < \cdots < i_r \leq n$ we have*

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdot \cdots \cdot \mathbb{P}(A_{i_r}).$$

In particular, in the above example, the evens A, B and C are pairwise independent but they are not independent as a collection. From now on, unless otherwise stated, by independence we mean independence as a collection.

Independent trials

Let $n \geq 1$ and $p \in (0,1)$ be fixed parameters. Consider an experiment, where one particular outcome, considered as "success" has probability p. Then this experiment is performed n times, and we are interested in the number of successes. An examples for that: a fair die is rolled 100 times, and a success is when the value 6 is rolled ($n = 100, p = 1/6$).

A mathematical model for that: to start, consider some sample space Ω_0 with an event $A \subset \Omega_0$ with $\mathbb{P}(A) = p$. Let the sample space of the trial sequence be the n-fold Cartesian product

$$\Omega = \Omega_0 \times \Omega_0 \times \cdots \times \Omega_0$$

and for any $i = 1, \ldots, n$ let

$$A_i = \Omega \times \cdots \times \Omega \times A \times \Omega \times \cdots \times \Omega$$

where the A is at the ith factor. Then $P(A_i) = p$ and the events $A_1, \ldots, A_n$ are independent (as a collection). Using this we arrive at the following formulas

$$\begin{aligned}
&\mathbb{P}(\text{all trials succeed}) = p^n;\\
&\mathbb{P}(\text{there is at least one success}) = 1 - (1-p)^n;\\
&\mathbb{P}(\text{there are exactly } k\text{successes}) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n).
\end{aligned}$$

Conditional independence

Definition 5.6. *The events B_1 and B_2 are conditionally independent given A if*

$$\mathbb{P}(B_1 \cap B_2 | A) = \mathbb{P}(B_1|A) \cdot \mathbb{P}(B_2|A)$$

Recall Example 4.3 with the following extension: after the first test, a second "independent" test is performed. Let

$$\begin{aligned}
B_1 &= \{\text{The first test is positive}\},\\
B_2 &= \{\text{The second test is positive}\}.
\end{aligned}$$

Then B_1 and B_2 are not independent; as if B_1 occurs, we are getting more suspicious that the person tested is actually ill, and the chances that B_2 occurs are higher than without any information whatsoever. Yet, B_1 and B_2 are independent *given the status of the person tested*. That is, B_1 and B_2 are conditionally independent given either A or A^c (where the notations of Example 4.3 are used).

A natural question that arises is what the chances of the person being ill are given that both tests have turned out to be positive.

$$\begin{aligned}\mathbb{P}(A|B_1 \cap B_2) &= \frac{\mathbb{P}(A \cap B_1 \cap B_2)}{\mathbb{P}(B_1 \cap B_2)} = \frac{\mathbb{P}(B_1 \cap B_2|A)\mathbb{P}(A)}{\mathbb{P}(B_1 \cap B_2|A)\mathbb{P}(A) + \mathbb{P}(B_1 \cap B_2|A^c)\mathbb{P}(A^c)} = \\ &= \frac{\mathbb{P}(B_1|A) \cdot \mathbb{P}(B_2|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B_1|A) \cdot \mathbb{P}(B_2|A) \cdot \mathbb{P}(A) + \mathbb{P}(B_1|A^c) \cdot \mathbb{P}(B_2|A^c) \cdot \mathbb{P}(A^c)} = \\ &= \frac{(0.95)^2 \cdot 0.005}{(0.95)^2 \cdot 0.005 + (0.01)^2 \cdot 0.995} \approx 0.98,\end{aligned}$$

where we have used conditional independence.

6 Discrete random variables

Definition 6.1. *A **random variable** is a (measurable) function $X : \Omega \to \mathbb{R}$.*

In words, a random variable is a quantity the value of which depends on randomness (on the outcome of the random experiment). Random variables are usually denoted by capital letters $(X, Y, Z, \dots)$ or greek letters $(\xi, \eta, \zeta ...)$ Measurability means that for every interval $I \subset \mathbb{R}$ the set

$$\{\omega \in \Omega | X(\omega) \in I\}$$

is an event, that is, it is included in the sigma-algebra $\mathcal{F}$. This ensures that it makes sense to consider the probability $\mathbb{P}(X \in I)$.

Definition 6.2. *A random variable is **discrete** if its range is a countable set. That is, X is discrete if the values it can take can be listed:*

$$x_1, x_2, \dots, x_k, \cdots \in \mathbb{R}.$$

*The **probability mass function** of a random variable is $p(= p_X) : \mathbb{R} \to \mathbb{R}$ defined by*

$$p(x) = \begin{cases} \mathbb{P}(X = x_k) & \text{if } x = x_k \text{ for some } k = 1, \dots \\ 0 & \text{otherwise.} \end{cases}$$

The characteristic properties of probability mass functions are

$$\begin{aligned} &p_k \geq 0, \qquad \forall k \geq 1; \\ &\sum_{k=1}^{\infty} p_k = 1. \end{aligned}$$

Consider the following examples:

- Flip three fair coins. Let Y denote the number of coins that turn up Head. Then Y can take the values $0, 1, 2$ and 3, and $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 3) = \frac{1}{8}$ while $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = \frac{3}{8}$.
- Keep rolling a fair die until the value 6 shows up for the first time. Let Z denote the number of rolls. Then Z can take any positive integer value, and $\mathbb{P}(Z = k) = (\frac{5}{6})^{k-1} \cdot \frac{1}{6}$.

We will se later that Y and Z are examples of binomially and geometrically distributed random variables, respectively.

Expected value

Definition 6.3. *The **expected value** of the discrete random variable X is defined as*

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} x_k P(X = x_k)$$

if the series is absolutely convergent.

Some comments:

- Note that $\mathbb{E}(X)$ is fixed number (in contrast with X, which is a random quantity). It can be regarded as the "center of mass" of the probability mass function. Another interpretation is related to the law of large numbers, to be discussed later.
- If the series converges converges only conditionally, but not absolutely, than we say that the expected value does not exist. Recall that in this case the sum of the series depends on the order in which the terms are added which is certainly to be avoided.

Expectation of a function of a random variable

Introductory example. Let X take values $0, 1$ and -1 with probabilities $\mathbb{P}(X = 0) = 0.5$, $\mathbb{P}(X = 1) = 0.2$ and $\mathbb{P}(X = -1) = 0.3$, and let $Y = X^2$. Then $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 0.5$ and thus $\mathbb{E}Y = 0.5$. Here we give an alternative way to compute $\mathbb{E}Y$.

Lemma 6.4. *Let X be a random variable with $\mathbb{E}X < \infty$, and let $g : \mathbb{R} \to \mathbb{R}$ be such that, for $Y = g(X)$, $\mathbb{E}Y < \infty$. Then*

$$\mathbb{E}Y = \mathbb{E}(g(X)) = \sum_{k=1}^{\infty} g(x_k)\mathbb{P}(X = x_k)$$

The lemma shows that if we are interested only in $\mathbb{E}(g(X))$, then we do not have to determine the distribution of $g(X)$.

Proof. Let y_ℓ; $\ell = 1, \dots$ denote the values that Y can take. As g may fail to be one-to-one, several x_k may be mapped to the same y_ℓ. We group the values x_k accordingly.

$$\sum_{k=1}^{\infty} g(x_k)\mathbb{P}(X = x_k) = \sum_{\ell=1}^{\infty} \sum_{k:g(x_k)=y_\ell} g(x_k)\mathbb{P}(X = x_k) =$$
$$= \sum_{\ell=1}^{\infty} y_\ell \left(\sum_{k:g(x_k)=y_\ell} \mathbb{P}(X = x_k) \right) = \sum_{\ell=1}^{\infty} y_\ell \mathbb{P}(Y = y_\ell) = \mathbb{E}Y.$$

□

In the introductory example, $\mathbb{E}X = -0.1$ so $(\mathbb{E}X)^2 = 0.01$, while $\mathbb{E}(X^2) = 1$. This shows that, in general, $\mathbb{E}(g(X)) \neq g(\mathbb{E}X)$. However, for $a, b \in \mathbb{R}$:

$$\mathbb{E}(aX + b) = \sum_{k=1}^{\infty} (ax_k + b)\mathbb{P}(X = x_k) = a\sum_{k=1}^{\infty} x_k \mathbb{P}(X = x_k) + b\sum_{k=1}^{\infty} \mathbb{P}(X = x_k) = a\mathbb{E}X + b.$$

For $r \geq 1$, $\mathbb{E}(X^r)$ (if exists) is called the rth moment of the random variable X.

Variance

Given a random variable X, let us denote (for brevity) the notation $\mu = \mathbb{E}X$, and let us define the **variance** of X by

$$Var\, X = \mathbb{E}((X - \mu)^2) = \sum_{k=1}^{\infty} (x_k - \mu)^2 \mathbb{P}(X = x_k).$$

This measures how strongly X fluctuates about its mean. Alternative Formula:

$$Var\, X = \sum_{k=1}^{\infty} (x_k^2 - 2\mu x_k + \mu^2)\mathbb{P}(X = x_k) =$$
$$= \sum_{k=1}^{\infty} x_k^2 \mathbb{P}(X = x_k) - 2\mu \sum_{k=1}^{\infty} x_k \mathbb{P}(X = x_k) + \mu^2 \sum_{k=1}^{\infty} \mathbb{P}(X = x_k) =$$
$$= \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

$\mathbb{D}X = \sqrt{Var\, X}$ is called the **standard deviation** of X. For linear transformations we have

$$Var(aX + b) = \mathbb{E}(aX + b - \mathbb{E}(aX + b))^2 = \mathbb{E}(aX + b - (a\mu + b))^2 =$$
$$= \mathbb{E}(a(X - \mu))^2 = a^2\mathbb{E}(X - \mu)^2 = a^2\, Var(X),$$

and thus

$$\mathbb{D}(aX + b) = |a|\mathbb{D}X.$$

7 The binomial distribution

Recall the setting of independent trials. A random variable that arises naturally in this context is the number of successes out of n trials.

Definition 7.1. *Let $n \geq 1$ be an integer and $p \in (0,1)$ a real parameter. The random variable X is binomially distributed with parameters n and p (notation $X \sim Binom(n,p)$) if it can take values $k = 0, 1, \ldots, n$ and*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n).$$

Let us compute the expected value and the variance of the binomial distribution. Consider first $n = 1$. This case is called an *indicator random variable* as we have just one trial, and X indicates whether a success occurs. We have

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

Now let us move on the the case $n \geq 2$. Let t be a real parameter. We have, for any integer $k \geq 1$,

$$\left.\frac{d}{dt}\right|_{(t=1)} (t^k) = \left.(kt^{k-1})\right|_{(t=1)} = k.$$

Moreover, if $k \geq 2$, then

$$\left.\frac{d^2}{dt^2}\right|_{(t=1)} (t^k) = \left.(k(k-1)t^{k-2})\right|_{(t=1)} = k(k-1).$$

Using this, linearity of differentiation and the binomial theorem, we have

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \left(\left.\frac{d}{dt}\right|_{(t=1)} (t^k) \right) \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \left.\frac{d}{dt}\right|_{(t=1)} \left(\sum_{k=1}^{n} \binom{n}{k} (tp)^k (1-p)^{n-k} \right) = \left.\frac{d}{dt}\right|_{(t=1)} ((tp+1-p)^n) = \\ &= \left.\left(np(tp+1-p)^{n-1}\right)\right|_{(t=1)} = np. \end{aligned}$$

With a similar computation

$$\begin{aligned} \mathbb{E}(X(X-1)) &= \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \left(\left.\frac{d^2}{dt^2}\right|_{(t=1)} (t^k) \right) \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \left.\frac{d^2}{dt^2}\right|_{(t=1)} \left(\sum_{k=1}^{n} \binom{n}{k} (tp)^k (1-p)^{n-k} \right) = \left.\frac{d^2}{dt^2}\right|_{(t=1)} ((tp+1-p)^n) = \\ &= \left.\left(n(n-1)p^2(tp+1-p)^{n-2}\right)\right|_{(t=1)} = n(n-1)p^2, \end{aligned}$$

which then implies

$$\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}X = n(n-1)p^2 + np;$$
$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = np - np^2 = np(1-p),$$

and thus

$$\mathbb{D}X = \sqrt{np(1-p)}.$$

In what follows, we investigate two important asymptotics of the binomial distribution.

Bernoulli's Law of Large Numbers

Let us fix some $p \in (0,1)$, and consider a sequence of random variables $X(= X^n) \sim Binom(n,p)$, with this *fixed value of p, but* $n \to \infty$. That is, we keep performing the same trial many times. Note that as $\mathbb{E}X = np \to \infty$, so does the mass associated to X, hence, it is more appropriate to consider $\frac{X}{n}$ instead of X, which is the *proportion of successes among the n trials.*

Proposition 7.2 (Bernoulli's Law of Large Numbers)**.** *Let X be as above. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| > \varepsilon\right) \to 0 \qquad \text{as } n \to \infty.$$

Bernoulli's Law of Large Numbers expresses, in a mathematically correct way, our experience that the proportion of successes approaches, as the number of trials grows, to the individual success rate p. It can be proved by analysing the binomial mass function. This is not included here, as Proposition 7.2 will follow as a particular case of the Weak Law of Large Numbers, which is stated and proved in section 23. Note, however, that

$$\left\{\left|\frac{X}{n} - p\right| > \varepsilon\right\} = \{|X - \mathbb{E}X| > n\varepsilon\}$$

and, as $n \to \infty$, no matter how small ε is, $n\varepsilon$ is asymptotically much larger than the standard deviation $\mathbb{D}X = \sqrt{np(1-p)}$.

The Poisson limit of the binomial distribution

Here we study an asymptotic regime of the binomial which is *quite different from the case of the Bernoulli Law of Large Numbers.* In particular, let $n \to \infty$, however, do not keep p fixed, instead, let $p(= p_n) \to 0$, in such a way that $n \cdot p \to \lambda$, where $\lambda > 0$ is a fixed parameter. Note that this way $\mathbb{E}X$ remains uniformly bounded, and accordingly, it is reasonable to consider the asymptotic behavior of the mass function of X.

Proposition 7.3. *Let us fix $\lambda > 0$ and let $n \to \infty$, $p(= p_n) \to 0$ such that $np \to \lambda$. Then, for any fixed integer $k \geq 0$, we have*

$$\binom{n}{k} p^k (1-p)^{n-k} \longrightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Proof. Recall from Calculus that

$$\lim_{x \to +\infty} \left(1 - \frac{1}{x}\right)^x = e^{-1},$$

which implies

$$\lim_{n\to\infty} (1-p)^n = \lim_{n\to\infty} \left(\left(1 - \frac{1}{p^{-1}}\right)^{p^{-1}} \right)^{p \cdot n} = (e^{-1})^\lambda = e^{-\lambda}. \tag{7.1}$$

Now

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n(n-1)\dots(n-k+1)}{k!} p^k (1-p)^{-k} (1-p)^n = \\ &= \frac{1}{k!} \cdot (np)((n-1)p)\dots((n-k+1)p) \cdot (1-p)^{-k} \cdot (1-p)^n \end{aligned}$$

and the last factor tends to $e^{-\lambda}$ by (7.1). As $p \to 0$ and k is fixed, $(1-p)^{-k} \to 1$. $\frac{1}{k!}$ is constant, and as $np \to \lambda$, we have also $(n-1)p \to \lambda$ and similarly the further factors up to $(n-k+1)p \to \lambda$, so

$$(np)((n-1)p)\dots((n-k+1)p) \to \lambda^k$$

which completes the proof of the Proposition. □

8 The Poisson distribution

Definition 8.1. *Given a parameter $\lambda > 0$, the random variable X is Poisson distributed with parameter λ (notation $X \sim Poi(\lambda)$) if*

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}; \qquad k = 0, 1, 2, ...$$

We have seen that the Poisson distribution arises in a particular asymptotic regime of the binomial distribution, when $n \to \infty$, $p = p(n) \to 0$ such that $np \to \lambda$. In words, the Poisson distribution is a good model for the number of successes when having *many independent* trials such that *the individual success rate is small*. This is a very frequent scenario – here are some Poisson distributed quantities:

- the number of calls received by the call center of a large bank within an hour. (There are many customers, but for each of them the chance of calling within that hour is small.)

- the number of accidents within a month at some busy junction. (There are many cars passing, but for each particular car the chance of a crash is very small.)
- the number of typos in a book chapter. (There are many characters, and for each character the chance of being misspelt is small.)

Here we compute the expected value of a variable $X \sim Poi(\lambda)$:

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

where we changed the index of summation to $m = k - 1$. You may say this is obvious, as we obtained the Poisson as a limit of the binomial – the expected value of which is np, and we had $np \to \lambda$. Yet, what we see here is that the order of taking the limit and the infinite summation (or in a related context, the integration) can be swapped, which is a highly nontrivial issue. We may proceed to compute the variance by noting $\mathbb{E}(X(X-1)) = \mathbb{E}(X^2) - \mathbb{E}X$, and

$$\mathbb{E}(X(X-1)) = \sum_{k=0}^{\infty} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda}\lambda^2 \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2$$

thus

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

Example 8.2. *How many chocolate chips should you plan per muffin to ensure that no more than one percent of customers get upset?*

What I mean is that a customer gets upset if she/he finds no chocolate chips at all in her/his muffin. Let X denote the number of chocolate chips in one particular muffin. *Claim:* X is Poisson distributed.

Assuming that the claim holds, given many customers, by Bernoulli's Law of Large Numbers we have that

$$\frac{\#\{\text{customers who get no muffin}\}}{\text{total number of customers}} \longrightarrow \mathbb{P}(X = 0)$$

hence we want that $e^{-\lambda} = \mathbb{P}(X = 0) < 0.01$, which implies the following lower bound on the parameter of the Poisson distribution: $\lambda > \ln 100 \approx 4.6$.

Now let us argue why the claim holds. To bake the muffins, a large amount of dough is prepared. This will be later chopped into $M \gg 1$ portions of equal size, where each portion corresponds to one muffin. But before splitting it up, $N \gg 1$ chocolate chips are put evenly into the dough. For any particular chip, the chance of landing in the portion that corresponds to my muffin is $1/M$. Hence we have many (N) trials with a small individual success rate ($1/M$), and the number of successes (chocolate chips in my muffin) is Poisson distributed. Moreover, $\lambda \approx N/M$, the number of chocolate chips planned per muffin.

You may skip the somewhat lengthy Example 7d on the length of the longest run in the Ross book. However, the material of pages 155–157 (ninth edition) on the Poisson process is definitely relevant for us.

8.1 The Poisson process

In many of the examples of quantities that are Poisson distributed, there is a *time scale* involved. This opens the perspectives to an exciting branch of probability: the theory of stochastic processes, which studies random phenomena evolving in time.

For the particular case of the Poisson process, we have a *point process*, a countable random subset of the halfline $[0, +\infty)$. Here the halfline is typically interpreted as time. The points in the random set will be referred as *impact*s, which may correspond to the calls at the call center, the accidents at the junction etc.

The Poisson process is defined by three characteristic properties. To formulate these, we need some terminology, which will be useful for future reference as well.

- Consider a continuous function $f : [0, +\infty) \to \mathbb{R}$, which we study for small values $h \to 0+$. $f(h) = o(h)$ ("little o of h") if

$$f(h) = o(h) \quad \Longleftrightarrow \quad \lim_{h \to 0+} \frac{f(h)}{h} = 0.$$

  In particular $h^2 = o(h)$, but $0.01 \cdot h \neq o(h)$. Also, $\sqrt{h} \neq o(h)$, however, $h = o(\sqrt{h})$.

- Consider two discrete random variables X and Y, taking values $x_1, x_2, ...$ and $y_1, y_2, ...$, respectively. X and Y are independent if for any pair of values x_k and y_ℓ the events $\{X = x_k\}$ and $\{Y = y_\ell\}$ are independent. Given random variables $X_1, X_2, ...X_m$, their independence (as a collection) is defined analogously.

Definition 8.3. *Fix some positive parameter λ. A point process on the positive halfline $[0, +\infty)$ is a **Poisson process of intensity** λ if*

P1 ***(Independence)*** *Consider $I_1, I_2, ...I_m$, an arbitrary finite collection of non-overlapping intervals in $[0, +\infty)$, and let $X_1, ..., X_m$ denote the number of impacts in the intervals $I_1, ...I_m$, respectively. The random variables $X_1, ..., X_m$ are independent.*

P2 ***(Homogeneity)*** *Let I be an (infinitesimally small) interval of length h. Then*

$$\mathbb{P}(\text{There is at least one impact in I}) = \lambda \cdot \text{h} + \text{o(h)}$$

P3 ***(No Accumulation)*** *Let I be an (infinitesimally small) interval of length h. Then*

$$\mathbb{P}(\text{There are at least two impacts in I}) = \text{o(h)}$$

Remark. Do not confuse λ, the intensity of the process with the various λ-s that appeared previously as parameters of a Poisson distribution. In particular, the intensity of the process has dimensions $\frac{1}{\text{time}}$. If time is measured in different units, λ has to be rescaled.

Proposition 8.4. *Consider a Poisson process of intensity λ, and let $I_t \subset [0, +\infty)$ be an interval of length t. Let $N(t)$ denote the number of impacts inside I_t. Then $N(t) \sim Poi(\lambda t)$.*

For the proof of this proposition we refer to the Ross book.

Example 8.5. *This is a variation on Example 7e from Ross. Let us assume that on a highway the average number of accidents is* 3 *per month. We start inspecting that highway on a particular day.*

(a) What is the chance that there are at least two accidents in the next twenty days?

(b) What is the chance that at least t *days elapse until the first accident?*

The number of accidents is a Poisson process of intensity $\lambda = 3\frac{1}{\text{month}} = 0.1\frac{1}{\text{day}}$. To solve part (a), let us measure time in days, then

$$N(20) \sim Poi\left(0.1\frac{1}{\text{day}} \cdot 20\text{days}\right) = Poi(2).$$

Hence

$$\begin{aligned}\mathbb{P}(N(20) > 2) &= 1 - \mathbb{P}(N(20) = 0) - \mathbb{P}(N(20) = 1) = 1 - e^{-2} - \frac{2}{2!}e^{-2} = \\ &= 1 - 2e^{-2} \approx 1 - 2 \cdot 0.135 = 0.73.\end{aligned}$$

To solve part (b), let T denote the random variable that measures the time (in days) that elapses until the first accident. Then

$$\mathbb{P}(T > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t} = e^{-0.1 \cdot t}.$$

9 Further discrete distributions

Here we discuss the material of section 4.8 from the Ross book.

9.1 The geometric distribution

Let $p \in (0, 1)$, the individual success rate in a sequence of independent trials (for example, $p = \frac{1}{6}$ for subsequent rolls of a fair die). The random variable X is geometrically distributed with parameter p (i. e. $X \sim Geom(p)$) if it can take the values $k = 1, 2, ...$ and

$$\{X = k\} \quad \Longleftrightarrow \quad \text{The first success is at the kth trial.}$$

For brevity, let us introduce $q = 1 - p$. Then, as $X = k$ means there are $k - 1$ failures in a row, followed by a success, the mass function is

$$\mathbb{P}(X = k) = q^{k-1}p.$$

Note

$$\sum_{k=1}^{\infty} q^{k-1}p = p\sum_{m=0}^{\infty} q^m = p \cdot \frac{1}{1-q} = 1$$

where we have summed up a geometric series, which is the reason for the name of this distribution. To compute expected value and variance, introduce the function $g(q) = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$. As $q \in (0,1)$, this power series converges. Hence, denoting differentiation w.r. to q by prime,

$$\begin{aligned} \sum_{k=1}^{\infty} kq^{k-1} &= g'(q) = \frac{1}{(1-q)^2}, \\ \sum_{k=2}^{\infty} k(k-1)q^{k-2} &= g''(q) = \frac{2}{(1-q)^3}. \end{aligned}$$

Now

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} kq^{k-1}p = pg'(q) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

In particular, in subsequent rolls of a fair die, the *expected time* when the first 6 occurs is at the 6th roll. Also,

$$\mathbb{E}(X(X-1)) = \sum_{k=2}^{\infty} k(k-1)q^{k-1}p = pqg''(q) = \frac{2pq}{(1-q)^3} = \frac{2q}{p^2},$$

thus

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}(X(X-1)) + \mathbb{E}X - \mathbb{E}X^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2},$$

where we have used $q + p = 1$. Then $\mathbb{D}(X) = \frac{\sqrt{q}}{p}$.

The geometric distribution has another remarkable feature, the **memoryless property**. Let k and n be arbitrary positive integers.

$$\begin{aligned} \mathbb{P}(X \geq k) &= q^{k-1}, \qquad \mathbb{P}(X > n) = q^n \\ \mathbb{P}(X \geq k+n | X > n) &= \frac{q^{k+n-1}}{q^n} = q^{k-1} = \mathbb{P}(X \geq k). \end{aligned}$$

In words: if there was no success in the first n trials, the probability that we have to make another k trials for the first success is independent of n. (What has happened previously does not change the distribution of the additional time needed for the first success.)

9.2 The negative binomial distribution

Fix two parameters, an integer $r \geq 1$ and a success rate $p \in (0,1)$. Consider again a sequence of independent trials. Then X has a negative binomial distribution with parameters r and p (i. e. $X \sim NegBinom(r,p)$) if

$$\{X = k\} \quad \Longleftrightarrow \quad \text{The rth success is at the kth trial.}$$

Apparently, X can take values $r, r+1, ...$ Try to figure out the mass function! ($PP(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$.)

Claim: $\mathbb{E}X = \frac{r}{p}$ and $Var\, X = \frac{rq}{p^2}$. The proof is postponed to the next section.

9.3 The hypergeometric distribution

This distribution will be familiar from problems on *sampling without replacement* (eg. drawing cards, capturing tagged elk etc.). Setting: there are N balls in a box, M blue balls and $N-M$ red balls. n balls are drawn. Let X denote the number of blue balls among the n balls drawn. Then

$$\mathbb{P}(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

Claim: $\mathbb{E}X=\frac{nM}{N}$. The proof is postponed to the next section.

10 Expected values of sums of random variables

Here we discuss the content of section 4.9 from Ross and some related material.

Lemma 10.1. *Let X and Y be arbitrary (discrete) random variables. We have $\mathbb{E}(X+Y)=\mathbb{E}X+\mathbb{E}Y$.*

Proof. Let X and Y take values $x_1,x_2,...$ and $y_1,y_2,...$, respectively. Let $p(x_k,y_l)=\mathbb{P}(\{X=x_k\}\cap\{Y=y_\ell\})$ (this is called the joint mass function, to be discussed later). Note that the events $\{X=x_k\}$, $k=1,2,...$ make a partition of the phase space. Similarly, the events $\{Y=y_\ell\},\ell=1,2,...$ make a partition of the phase space. Hence

$$\sum_\ell p(x_k,y_l)=\mathbb{P}(X=x_k),\qquad \sum_k p(x_k,y_l)=\mathbb{P}(Y=y_\ell)$$

Now

$$\begin{aligned}\mathbb{E}(X+Y) &= \sum_{k,\ell}(x_k+y_\ell)p(x_k,y_\ell)=\sum_{k,\ell}x_kp(x_k,y_\ell)+\sum_{k,\ell}y_\ell p(x_k,y_\ell)= \\ &= \sum_k x_k\sum_\ell p(x_k,y_\ell)+\sum_\ell y_\ell\sum_k p(x_k,y_\ell)= \\ &= \sum_k x_k\mathbb{P}(X=x_k)+\sum_\ell y_\ell\mathbb{P}(Y=y_\ell)=\mathbb{E}X+\mathbb{E}Y.\end{aligned}$$

□

By induction, for arbitrary random variables $X_1,X_2,...X_N$ we have

$$\mathbb{E}(X_1+X_2+...+X_N)=\mathbb{E}X_1+\mathbb{E}X_2+...+\mathbb{E}X_N. \tag{10.1}$$

Lemma 10.2. *Let X and Y be independent (discrete) random variables. We have $Var(X+Y)=Var\,X+Var\,Y$.*

Proof. We keep using the notation from the proof of Lemma 10.1. Note that as X and Y are independent,

$$p(x_k, y_\ell) = \mathbb{P}(X = x_k)\mathbb{P}(Y = y_\ell).$$

This implies:

$$\mathbb{E}(X \cdot Y) = \sum_{k,\ell} x_k y_\ell p(x_k, y_\ell) = \left(\sum_k x_k \mathbb{P}(X = x_k)\right) \cdot \left(\sum_\ell y_\ell \mathbb{P}(Y = y_\ell)\right) = (\mathbb{E}X) \cdot (\mathbb{E}Y). \tag{10.2}$$

Now, using Lemma 10.1 and (10.2)

$$\mathbb{E}((X+Y)^2) = \mathbb{E}X^2 + \mathbb{E}Y^2 + 2\mathbb{E}(X \cdot Y) = \mathbb{E}X^2 + \mathbb{E}Y^2 + 2(\mathbb{E}X)(\mathbb{E}Y). \tag{10.3}$$

On the other hand, using again Lemma 10.1,

$$(\mathbb{E}(X+Y))^2 = (\mathbb{E}X)^2 + (\mathbb{E}Y)^2 + 2(\mathbb{E}X)(\mathbb{E}Y). \tag{10.4}$$

Subtracting (10.4) from (10.3) completes the proof of the Lemma. □

By induction, for *independent* random variables $X_1, X_2, ...X_N$ we have

$$Var(X_1 + X_2 + ... + X_N) = Var\, X_1 + Var\, X_2 + ... + Var\, X_N. \tag{10.5}$$

Simple as they may seem, Formulas (10.1) and (10.5) have *loads of useful applications.*

10.1 Indicator variables

Given an event A, *the indicator variable associated to* A is

$$\eta_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

Note that for an indicator random variable, if $p = \mathbb{P}(A)$, we have

$$\begin{aligned} \mathbb{E}\eta_A &= 1 \cdot \mathbb{P}(A) + 0 \cdot (1 - \mathbb{P}(A)) = p, \\ \mathbb{E}(\eta_A^2) &= 1^2 \cdot \mathbb{P}(A) + 0^2 \cdot (1 - \mathbb{P}(A)) = p, \\ Var(\eta_A) &= \mathbb{E}(\eta_A^2) - (\mathbb{E}\eta_A)^2 = p - p^2 = p(1-p). \end{aligned} \tag{10.6}$$

In many cases, it turns out to be a very useful idea to split up a random variable as a sum of indicators.

The binomial distribution

Let $X \sim Binom(n,p)$. Introduce, for $i = 1, 2, ..., n$

$$\eta_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success,} \\ 0 & \text{if the } i\text{th trial is a failure.} \end{cases}$$

Then

- $X = \eta_1 + \eta_2 + ... + \eta_n$. To see this, think of someone who observes the sequence of independent trials, and puts a tick on a piece of paper each time a trial succeeds. Then, after the sequence has terminated, the number of ticks will be exactly the number of successes, the actual value that X takes.
- The random variables $\eta_1, \eta_2, ..., \eta_n$ are independent, as they correspond to independent trials.
- By (10.6), $\mathbb{E}\eta_i = p$ and $Var\,\eta_i = p(1-p)$, for $i = 1, ..., n$.

This way

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}(\eta_1 + ...\eta_n) = \mathbb{E}\eta_1 + ... + \mathbb{E}\eta_n = np, \\ Var\,X &= Var(\eta_1 + ...\eta_n) = Var\,\eta_1 + ... + Var\,\eta_n = np(1-p). \end{aligned} \tag{10.7}$$

We have seen this before, but, actually, this argument describes what is going on.

The hypergeometric distribution

Let X have hypergeometric distribution, as in subsection 9.3. Let us think of the balls as if they were labeled with $1, ..., N$, in such a way that the blue balls have the labels $1, ..., M$ (and thus the red balls carry the labels $M+1, ..., N$). For $i = 1, ..., M$ we introduce the following indicator variables:

$$\eta_i = \begin{cases} 1 & \text{if ball } \#i \text{ is among the } n \text{ balls drawn,} \\ 0 & \text{otherwise.} \end{cases}$$

Now:

$$X = \eta_1 + \cdots + \eta_M.$$

To see this, note that there are exactly as many blue balls among the n balls drawn, as many indicators $\eta_1, \ldots, \eta_M$ "fire" (are equal to 1). Now, for any $i = 1, \ldots, M$, we have:

$$\mathbb{E}\eta_i = \mathbb{P}(\eta_i = 1) = \mathbb{P}(\text{ball } \#i \text{ is drawn}) = \frac{n}{N}.$$

To see this, think of the process as if it did not stop after selecting n balls, we keep on drawing until all the N balls are drawn. This way, the outcome of our random experiment

is a permutation of the balls. We consider the first n items of this permutation as being selected. Any specific ball, hence in particular ball $\#i$, has an equal chance of taking any of the N possible positions, so the chance it takes one of the first n positions – it is selected – is n/N.

Now using (10.1) we arrive at

$$\mathbb{E}X = \mathbb{E}(\eta_1 + \dots \eta_M) = M\frac{n}{N},$$

which is what we stated back in subsection 9.3.

Note that the indicators η_i are not independent this time, so (10.5) does not apply.

The elevator problem

5 people board the elevator on the first floor of an 11-storied building. Each of them picks one of the floors $2, 3, \dots, 11$ as a destination, evenly, and independently of each other. Let X denote the number of times the elevator stops. Let us introduce the following indicator variables:

$$\eta_i = \begin{cases} 1 & \text{if the elevator stops on floor } \#i, \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 2, 3, \dots, 11$. We have to compute $\mathbb{E}\eta_i = \mathbb{P}(\eta_i = 1)$. Now $\{\eta_i = 0\}$ means that none of the 5 passengers pick floor $\#i$, that is, all of them select one of the other 9 possible destination floors. Since they act independently, the chance of that is $(\frac{9}{10})^5$. We arrive at

$$\mathbb{E}\eta_i = 1 - \mathbb{P}(\eta_i = 0) = 1 - (0.9)^5,$$

and thus by (10.1) at

$$\mathbb{E}X = \mathbb{E}(\eta_2 + \dots + \eta_{11}) = 10(1 - (0.9)^5) = 4.0951.$$

The indicators η_i are not independent this time either, so (10.5) does not apply.

10.2 Stopping times

Here the idea is to split up a random variable, which measures the time until a particular event, as a sum of *geometric* random variables (cf. subsection 9.1). These geometrically distributed terms correspond to the random intervals that elapse between consecutive intermediate stops, hence the name.

The negative binomial distribution

Let $X \sim NegBinom(r, p)$, as in subsection 9.2. Recall that X is the number of trials needed for the rth success. Let us introduce the random variables $T_1, T_2, \dots, T_r$, where:

- T_1 is the number of trials needed for the first success;

- T_2 is obtained by counting the number of trials after the first success until the second success;
- and so on: T_i is the number of trials made after the $(i-1)$st success, until the ith success.

Then

- $X = T_1 + T_2 + \cdots + T_r$;
- for any $i = 1, \ldots, r$; the random variable T_i is geometrically distributed with parameter p;
- as we consider a sequence of independent trials, the variables $T_1, ..., T_r$ are independent.

In short, X arises as a sum of i.i.d. (independent, identically distributed) – in fact, geometrically distributed – random variables. Also, by subsection 9.1, $\mathbb{E}T_i = \frac{1}{p}$ and $Var(T_i) = \frac{q}{p^2}$. Thus, by Formulas (10.1) and (10.5), we have:

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}(T_1 + T_2 + \cdots + T_r) = r\frac{1}{p} = \frac{r}{p}, \\ Var\, X &= Var(T_1 + T_2 + \cdots + T_r) = r\frac{q}{p^2} = \frac{rq}{p^2}, \end{aligned}$$

which is what was stated back in subsection 9.2.

The coupon collector problem

We have already introduced this problem back in February: a certain product (chocolate bar, soda...) can be any of M different types. Any item purchased is of a type i ($i = 1, \ldots, M$), with chance $1/M$, independently of other items. Let the random variable S denote *the number of items one has to purchase to collect all M types*. (Analogy: collecting state quarters.)

Think about splitting up S as a sum of stopping times, to find its expected value and variance. This is similar to the case of the negative binomial discussed above, but not entirely the same.

As discussed in class, $S = T_1 + T_2 + \cdots + T_M$, where T_i ($i = 1, \ldots M$) is the number of items one has to purchase, after the $(i-1)$th type has been collected, to collect the ith type. Then $T_i \sim Geom(\frac{M_i+1}{M})$, since, after $i-1$ types have been collected, I am happy if a new item is any of the remaining $M - (i-1)$ types. So

$$\mathbb{E}(S) = \mathbb{E}(T_1 + \cdots + T_M) = 1 + \frac{M}{M-1} + \frac{M}{M-2} + \cdots + M.$$

11 Cumulative distribution functions

From now on, the random variables are not necessarily discrete.

Definition 11.1 (Cumulative distribution function)**.** *Let X be an arbitrary random variable. Then the (cumulative) distribution function of X is*

$$F_X : \mathbb{R} \to \mathbb{R}, \qquad F_X(x) = \mathbb{P}(X \leq x).$$

Comments:

- Do not confuse X, which is a random variable with x, which is a real number.
- If there is no risk of confusion, the subscript X is often dropped and the distribution function is written as $F(x)$.

Our fist example is a discrete random variable.

Example 11.2. *Let us flip three fair coins (say a quarter, a nickel and a dime), and let X denote how many of them turns up Head. Determine the distribution function $F_X(x)$.*

X is a familiar, binomially distributed random variable, it takes the values 0 and 3 with probability 1/8 each; and the values 1 and 2 with probability 3/8 each. To determine $F_X(x)$:

- let $x < 0$. Then $F(x) = \mathbb{P}(X \leq x) = 0$, as X cannot take negative values.
- $F(0) = \mathbb{P}(X \leq 0)$, and $X \leq 0$ can only occur for this particular random variable if $X = 0$. Hence $F(0) = 1/8$. Similarly, for $x \in [0, 1)$, $X \leq x$ can only occur if $X = 0$, hence $F(x) = 0$ on this interval.
- we have that, for this particular random variable, $X \leq 1$ if either $X = 0$ or $X = 1$. Hence
$$F(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}.$$
Arguing as before, for $x \in [1, 2)$, $X \leq x$ can only occur if $X = 0$ or $X = 1$, hence $F(x) = \frac{1}{2}$ on this interval.
- proceeding analogously with this reasoning we arrive at the function depicted on Figure 1.

This example demonstrates the general fact that *for a discrete random variable ξ, $F_\xi(x)$ is a step function*, the jumps of which are at the values that ξ can take, and the heights of these jumps are the probabilities with which the particular values are taken. Now let us consider a very different example.

Example 11.3. *Let us consider a Poisson process of intensity λ, and let T denote the time that elapses until the first impact, as in Example 8.5, part (b).*

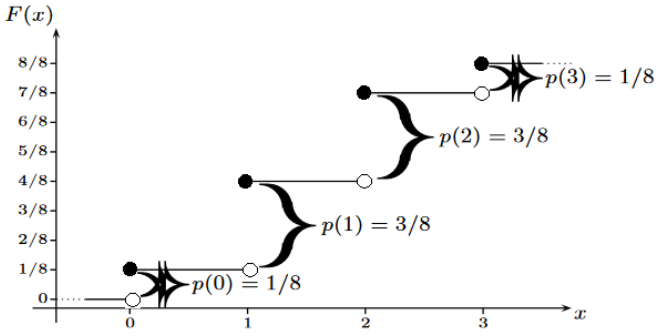

Figure 1: Cumulative distribution function for Example 11.2

Apparently $T \geq 0$, hence for any $t < 0$ we have $F(t) = 0$. Now let us consider $t \geq 0$. Then, as already computed in Example 8.5:

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - \mathbb{P}(T > t) = 1 - \mathbb{P}(N(t) = 0) = 1 - e^{-\lambda t}.$$

Hence we arrive at

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

See also Figure 2. This function, as opposed to the one depicted on Figure 1, is *not* a step function, it is continuous, actually, differentiable except for the single point 0.

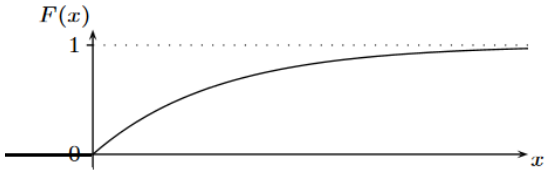

Figure 2: Cumulative distribution function for Example 11.3

Properties of cumulative distribution functions

Let $F(x)$ be the distribution function of some random variable X.

1. $F(x)$ **is nondecreasing**: for $x_1 < x_2$, we have $F(x_1) \leq F(x_2)$. Indeed, the events $E_1 = \{X \leq x_1\}$ and $E_2 = \{X \leq x_2\}$ satisfy $E_1 \subset E_2$, hence $F(x_1) = \mathbb{P}(E_1) \leq \mathbb{P}(E_2) = F(x_2)$.

2. $\lim\limits_{\mathbf{x} \to +\infty} \mathbf{F}(\mathbf{x}) = \mathbf{1}$. By monotonicity (1. above), it is enough to show $\lim\limits_{n \to \infty} F(n) = 1$ and refer to the sandwich principle. To this end, let $A_n = \{X \leq n\}$. We have

$$\lim_{n \to +\infty} F(n) = \lim_{n \to +\infty} \mathbb{P}(A_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbb{P}(\Omega) = 1,$$

since the A_n make an increasing sequence of events.

3. $\lim_{\mathbf{x}\to-\infty} \mathbf{F}(\mathbf{x}) = \mathbf{0}$. It is enough to show $\lim_{n\to\infty} F(-n) = 0$ and refer to the sandwich principle. To this end, let $B_n = \{X \leq -n\}$. We have

$$\lim_{n\to\infty} F(-n) = \lim_{n\to+\infty} \mathbb{P}(B_n) = \mathbb{P}(\bigcap_{n=1}^{\infty} B_n) = \mathbb{P}(\emptyset) = 0,$$

since the B_n form a decreasing sequence of events.

4. $F(x)$ **is right continuous.** Again referring to monotonicity (1. above), it is enough to show that, for any fixed $x_0 \in \mathbb{R}$, $\lim_{n\to\infty} F(x_0 + \frac{1}{n}) = F(x_0)$. Let $C_n = \{X \leq x_0 + \frac{1}{n}\}$, which is a decreasing sequence and

$$\bigcap_{n=1}^{\infty} C_n = \{X \leq x_0\} \;\Rightarrow\; \lim_{n\to+\infty} \mathbb{P}(C_n) = \mathbb{P}(X \leq x_0) = F(x_0).$$

Comments:

- The above four properties characterize distribution functions. This means that, on the one hand, for any random variable X the distribution function $F_X(x)$ has these properties. On the other hand, it holds true that for any $F(x)$ that has the above four properties, there exist a random variable X such that $F(x) = F_X(x)$.

- $F(x)$ is, *in general, not continuous* from the left. For any $x_0 \in \mathbb{R}$ fixed, let us introduce the notations

$$F(x_0 - 0) = \lim_{x\to x_0-0} F(x), \qquad F(x_0 + 0) = \lim_{x\to x_0+0} F(x)$$

for the left and right limits of $F(x)$ at x_0, respectively. Then 4. above states that $F(x_0 + 0) = F(x_0)$, for any $x_0 \in \mathbb{R}$. On the other hand,

$$F(x_0 - 0) = \lim_{n\to\infty} \mathbb{P}(X \leq x_0 - \frac{1}{n}) = \mathbb{P}(\bigcap_{n=1}^{\infty} \{X \leq x_0 - \frac{1}{n}\}) = \mathbb{P}(X < x_0).$$

In particular:

$$F(x_0) - F(x_0 - 0) = \mathbb{P}(X \leq x_0) - \mathbb{P}(X < x_0) = \mathbb{P}(X = x_0).$$

So, $F_X(x)$ has a jump at some point $x_0 \in \mathbb{R}$ if and only if $\mathbb{P}(X = x_0) > 0$. Accordingly, if $F_X(x)$ is *continuous* on its entire domain $\mathbb{R}$, this means that *no fixed value is taken by X with positive probability.* In other words, X does not have any discrete component.

The significance of the distribution function is that **any information on the distribution of X can be expressed in terms of $F_X(x)$.** Let, in particular $a < b$ be arbitrary. Then:

$$\mathbb{P}(X \in (a, b]) = \mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a). \qquad (11.1)$$

12 Absolutely continuous random variables

Definition 12.1. *The random variable X is absolutely continuous if there exists an integrable function $f_X : \mathbb{R} \to \mathbb{R}$ such that*

$$F_X(a) = \int_{-\infty}^{a} f_X(x)dx, \qquad \forall a \in \mathbb{R}. \tag{12.1}$$

In this case $f_X(x)$ is the density function of X.

Recall from calculus that in this case $F(x)$ is (one of) the antiderivative(s) of $f(x)$. Accordingly, building upon (11.1):

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$

in accordance with the fundamental theorem of calculus. This then generalizes to

$$\mathbb{P}(X \in B) = \int_{B} f(x)dx$$

for any Borel measurable set $B \subset \mathbb{R}$. In particular, for any $x_0 \in \mathbb{R}$,

$$\mathbb{P}(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0,$$

that is, no fixed value is taken by X with positive probability. Recall that this is equivalent to the continuity of $F(x)$. In fact, $F(x)$ is not only continuous, it is (almost everywhere) differentiable: $\frac{dF}{dx}|_{x=x_0} = f(x_0)$ for almost every $x_0 \in \mathbb{R}$.

Comment: The range of an absolutely continuous random variable X is not restricted to a countable set, the probability is "smeared all over the place". As no fixed number is taken with positive probability, one may wonder how to interpret $f_X(x_0)$ for some $x_0 \in \mathbb{R}$. Now let consider some $\varepsilon > 0$ infinitesimally small. Then, assuming that $f(x)$ is continuous at x_0:

$$\mathbb{P}(x_0 \leq X \leq x_0 + \varepsilon) = \int_{x_0}^{x_0+\varepsilon} f_X(x)dx \approx f_X(x_0)\varepsilon,$$

accordingly, $f_X(x_0)$ can be interpreted as *the rate with which the probability of falling into a small interval about x_0 scales down* as the size of the interval tends to 0.

Density functions have the following characteristic properties:

- $f(x) \geq 0$ for (almost) every $x \in \mathbb{R}$,
- $\int_{-\infty}^{+\infty} f(x)dx = 1.$

Compare these with what we have learned about the probability mass function of a discrete random variables! In a nutshell, in the absolutely continuous case we will have the same formulas as in the discrete case, just summation will be replaced by integration. You will have to recall what you learned in calculus!

Example 12.2. *The density function of a random variable X is*

$$f_X(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{A}{x^3} & \text{if } x \geq 1. \end{cases}$$

(a) $A = ?$

(b) $\mathbb{P}(2 \leq X \leq 3) = ?$

To find A, we have to check that X satisfies the two characteristic properties of density functions. $f(x) \geq 0$ holds for every $x \in \mathbb{R}$ if $A \geq 0$. For the other property:

$$\begin{aligned} 1 = \int_{-\infty}^{+\infty} f(x)dx &= \int_1^{\infty} \frac{A}{x^3}dx = \lim_{L\to\infty} \left[-\frac{A}{2x^2}\right]_1^L = \\ &= \lim_{L\to\infty} \left(-\frac{A}{2L^2} + \frac{A}{2}\right) = \frac{A}{2}, \implies A = 2. \end{aligned}$$

To solve part (b):

$$\mathbb{P}(2 \leq X \leq 3) = \int_2^3 f(x)dx = \int \frac{2}{x^3}dx = \left[\frac{1}{x^2}\right]_2^3 = -\frac{1}{9} + \frac{1}{4} = \frac{5}{36}.$$

There are several further examples of similar character in section 5.1 of the Ross book. Also, Homework set #7 (to be posted soon) will contain several such exercises.

Expected values

Definition 12.3. *The expected value of an absolutely continuous random variable is defined as*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x)dx.$$

Convergence issues analogous to the ones discussed in the discrete case may arise. In particular, it is required that the integral is absolutely convergent, otherwise, there are three possibilities, $\mathbb{E}X = +\infty$, $\mathbb{E}X = -\infty$, or the expected value does not exist.

Proposition 12.4. *If X is an absolutely continuous random variable, and $g : \mathbb{R} \to \mathbb{R}$ is measurable, then $Y = g(X)$ is another random variable. We have*

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

if the integral on the RHS is absolutely convergent.

For the proof, see section 5.2 of the Ross book. Let me just mention that the argument we had for the analogous statement in the discrete case does not work here.

From here the general formalism runs pretty much parallel to the discrete case. Let us keep our variable X fixed. It is a common notation to write $\mu = \mu_X := \mathbb{E}X$ (the subscript is dropped if there is no ambiguity).

Definition 12.5. *The variance of the variable X is defined by*

$$Var\, X = \mathbb{E}((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx.$$

The standard deviation of X is $\mathbb{D}X = \sqrt{Var\, X}$.

The following properties are proved exactly the same way as in the discrete case, just replace summation by integration:

- alternative Formula for the variance:

$$Var\, X = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2,$$

- generally, $\mathbb{E}(g(X))$ and $g(\mathbb{E}X)$ are different,
- linear rescaling: for $a, b \in \mathbb{R}$ we have

$$\mathbb{E}(a\cdot X+b) = a\cdot\mathbb{E}(X)+b, \qquad Var(a\cdot X+b) = a^2\cdot Var(X), \qquad \mathbb{D}(a\cdot X+b) = |a|\cdot\mathbb{D}(X). \tag{12.2}$$

13 Uniform distribution

Definition 13.1. *Let $\alpha < \beta$ two real numbers. The random variable X is uniformly distributed on the interval $[\alpha, \beta]$ if its density $f_X(x)$ is equal to some positive constant c on the interval $[\alpha, \beta]$, and vanishes outside the interval. Notation: $X \sim UNI[\alpha, \beta]$.*

Note that the requirement $\int_{-\infty}^{\infty} f_X(x)dx = 1$ fixes the constant c and we have:

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

and, by (12.1)

$$F_X(x) = \begin{cases} 0 & \text{if } x < \alpha, \\ \frac{x-\alpha}{\alpha-\beta} & \text{if } \alpha \le x \le \beta, \\ 1 & \text{if } x > \beta. \end{cases}$$

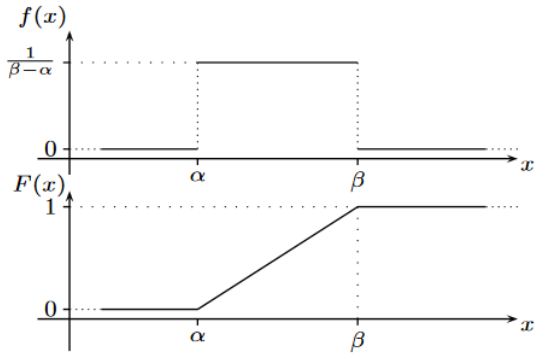

Figure 3: Density and distribution function for the uniform distribution

See also Figure 3.

Here we compute the expected value and the variance for the uniform distribution. Let us consider first the special case $Y \sim UNI[0,1]$, then $f_Y(x) = 1$ for $x \in [0,1]$ and $f_Y(x) = 0$ otherwise. Hence

$$\begin{aligned} \mathbb{E}Y &= \int_{-\infty}^{\infty} x f_Y(x)dx = \int_0^1 x dx = \frac{1}{2}; \\ \mathbb{E}(Y^2) &= \int_{-\infty}^{\infty} x^2 f_Y(x)dx = \int_0^1 x^2 dx = \frac{1}{3}, \\ Var(Y) &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Now let $X \sim UNI[\alpha, \beta]$, then $X = (\beta - \alpha)Y + \alpha$, and using (12.2):

$$\begin{aligned} \mathbb{E}X &= (\beta - \alpha)\mathbb{E}Y + \alpha = \frac{\alpha + \beta}{2}, \\ Var(X) &= (\beta - \alpha)^2 \, Var\, Y = \frac{(\beta - \alpha)^2}{12}. \end{aligned}$$

For problems on the uniform distribution, a useful formula is as follows. Let $\alpha \le \alpha_1 < \beta_1 \le \beta$, then

$$\mathbb{P}(\alpha_1 \le X \le \beta_1) = \int_{\alpha_1}^{\beta_1} f_X(x)dx = \frac{\beta_1 - \alpha_1}{\beta - \alpha} = \frac{\text{length of "useful" interval}}{\text{total available length}}. \tag{13.1}$$

This is analogous to the formula we had in combinatorial problems. Some applications:

- Let $X \sim UNI[0,5]$. What is the chance that the integer part of X is an odd number?

$$\mathbb{P}(\text{integer part odd}) = \mathbb{P}(X \in [1,2)) + \mathbb{P}(X \in [3,4)) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5} = 0.4.$$

- Let $Y \sim UNI[0,1]$. What is the chance that the numbers Y, $1-Y$ and $\frac{1}{2}$ arise as the sidelengths of a triangle?

$$\begin{aligned} Y + 1 - Y > \frac{1}{2} \quad &\Longrightarrow \quad \text{always}, \\ Y + \frac{1}{2} > 1 - Y \quad &\Longrightarrow \quad Y > \frac{1}{4}, \\ 1 - Y + \frac{1}{2} > Y \quad &\Longrightarrow \quad Y < \frac{3}{4}, \end{aligned} \tag{13.2}$$

  hence

$$\mathbb{P}\left(\text{can from a triangle}\right) = \mathbb{P}\left(\frac{1}{4} < Y < \frac{3}{4}\right) = \frac{1}{2}.$$

See Example 3c in section 5.3 of the Ross book for another application. Also, such problems are included in Homework set #7.

Although we do not discuss multidimensional distributions in this section, it is worth mentioning that Formula (13.1) has a natural generalization to higher dimensions. For example, if the random point P is uniformly distributed on some bounded domain $D \subset \mathbb{R}^2$, then for $D_1 \subset D$,

$$\mathbb{P}(P \in D_1) = \frac{Area(D_1)}{Area(D)}.$$

As an application, recall the Bertrand's paradox, Example 3d in section 4.3 of Ross, which we discussed on the very first class of this course, back in Budapest.

Distributions that are neither discrete, nor absolutely continuous

Discrete and absolutely continuous are two major classes of random variables, however, there are distributions that do not belong to any of these two categories. This can happen in different ways.

Distributions that have a discrete and an absolutely continuous component. This is Example 10a in section 4.10 of the Ross book. See also Figure 9 there. Such a random variable can be realized as follows. Flip a fair coin. If the coin turns up Head, let X be uniformly distributed on the interval $[0, 0.5]$. If the coin turns up Tail, let X follow a discrete distribution. In particular the *conditional* probabilities are $\mathbb{P}(X = 1|\text{Tail}) = 1/3$, $\mathbb{P}(X = 2|\text{Tail}) = 1/2$ and $\mathbb{P}(X = 3|\text{Tail}) = 1/6$.

Distributions that have neither discrete, nor absolutely continuous components. This is a different scenario, more interesting mathematically, and related to fractals. Watch the "devil's staircase" video on this, linked from piazza.

14 Normal (Gaussian) random variables

This is probably the most important class of continuous distributions. We start with the case of the *standard normal* distribution.

The standard normal distribution

Definition 14.1. *A random variable Z has standard normal distribution (notation: $Z \sim \mathcal{N}(0,1)$) if its density function is*

$$f_Z(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Accordingly, the cumulative distribution function of Z is

$$F_Z(z) = \Phi(z) = \int_{-\infty}^{z} \varphi(x)dx.$$

The graph of the function $\varphi(x)$ is the well known bell curve, see Figure 4. The function $\Phi(x)$ cannot be expressed in terms of elementary functions. It is an analytic function, see Figure 5 for its graph. Its values are given in the *standard normal table*, available for instance in the resources section of piazza. When solving homeworks or taking exams, you may use this table.

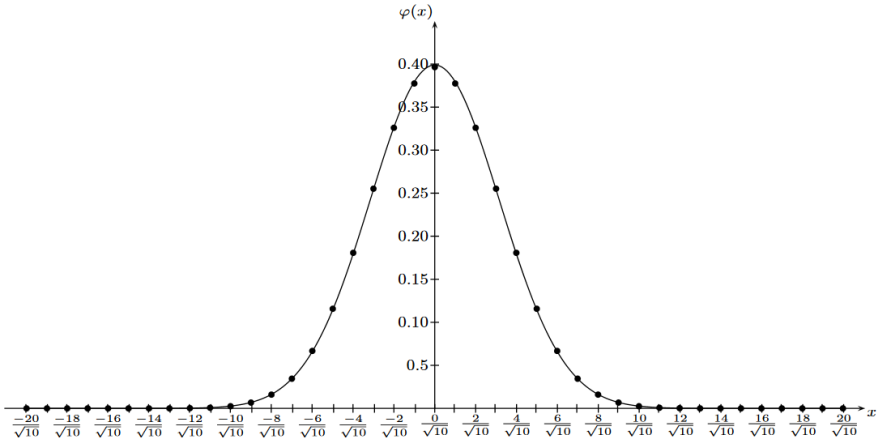

Figure 4: The standard normal density $\varphi(x)$

Note that $\varphi(x)$ is an even function – this precisely means that the distribution of $Z \sim \mathcal{N}(0,1)$ is symmetric about the origin. This implies in particular

$$\Phi(0) = \frac{1}{2}, \qquad \Phi(-x) = 1 - \Phi(x).$$

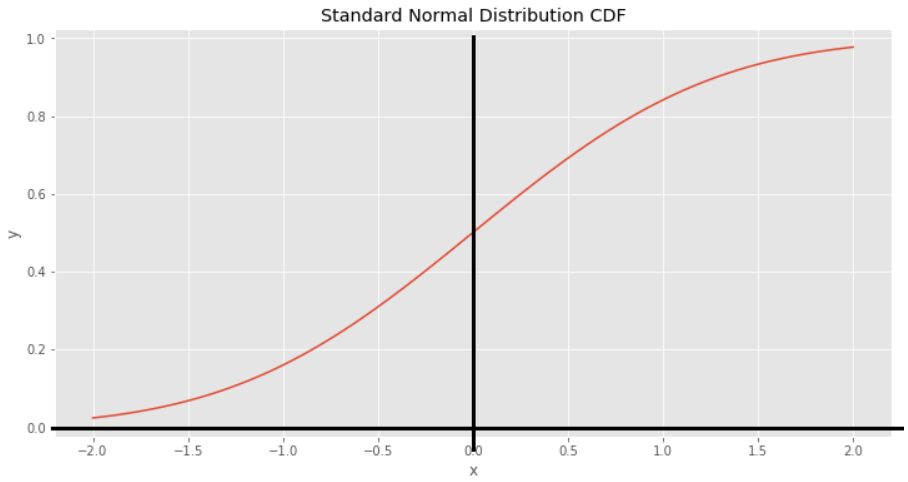

Figure 5: The standard normal distribution function $\Phi(x)$

We still need to check that $\varphi(x)$ is a proper density function. $\varphi(x) \geq 0$ is obvious. Now let

$$I = \int_{-\infty}^{\infty} e^{-x^2/2} dx.$$

We still need to verify

$$\int_{-\infty}^{\infty} \varphi(x) dx \;\Leftrightarrow\; I = \sqrt{\pi} \;\Leftrightarrow\; I^2 = 2\pi.$$

Now:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2/2} dx \cdot \int_{-\infty}^{\infty} e^{-y^2/2} dy = \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dxdy = \\ &= \int_{\theta=0}^{2\pi} \int_{r=0}^{+\infty} e^{-\frac{r^2}{2}} r dr d\theta = (2\pi) \left[-e^{-\frac{r^2}{2}} \right]_{r=0}^{\infty} = 2\pi. \end{aligned}$$

where we have used polar coordinates.

Now we compute the expected value and the variance of $Z \sim \mathcal{N}(0,1)$. First note that there are no issues with convergence: since there exists $C > 0$ such that

$$e^{-x^2/2} \leq C e^{-|x|}$$

and, for arbitrary $k \geq 1$, $\int_0^{+\infty} x^k e^{-x} dx$ is absolutely convergent, so is $\int_{-\infty}^{\infty} x^k e^{-x^2/2} dx$. This

means in particular that limits to $\pm\infty$ can be taken in an arbitrary manner. Now

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} x\varphi(x)dx = (2\pi)^{-1/2} \int_{-\infty}^{\infty} xe^{-x^2/2}dx = 0,$$

since the function is odd, and the interval is symmetric about the origin. To proceed,

$$Var(Z) = \mathbb{E}(Z^2) - (\mathbb{E}Z)^2 = \mathbb{E}(Z^2) = \int_{-\infty}^{\infty} x^2\varphi(x)dx = (2\pi)^{-1/2} \int_{-\infty}^{\infty} x^2e^{-x^2/2}dx,$$

and to compute $\int_{-\infty}^{\infty} x^2e^{-x^2/2}dx$ we use integration by parts with $u = x$ and $dv = xe^{-x^2/2}dx$. Then $du = dx$ and $v = -e^{-x^2/2}$, so

$$\int_{-\infty}^{\infty} x^2e^{-x^2/2}dx = \left[-xe^{-x^2/2}\right]_{-\infty}^{+\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}dx = I = \sqrt{2\pi}$$

as the first term vanishes by L'Hospital's rule. We arrive at

$$Var(Z) = 1.$$

To define normal distributions with other parameters, it is convenient to introduce some further terminology.

Standardization

Definition 14.2. *A random variable Y is standard if $\mathbb{E}Y = 0$ and $Var(Y) = 1$. Let X be an arbitrary random variable, and let $\mu = \mathbb{E}X$ and $\sigma^2 = Var(X)$. The standardization of X is*

$$Y = \frac{X - \mu}{\sigma} \iff X = \sigma Y + \mu.$$

It is easy to check that Y is indeed a standard random variable:

$$\begin{aligned} \mathbb{E}Y &= \mathbb{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\left(\mathbb{E}X - \mu\right) = 0 \\ Var\,Y &= Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\,Var(X) = 1. \end{aligned}$$

Let $F_X(x)$ and $F_Y(x)$ denote the cumulative distribution functions of X and Y, respectively. Then

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(\sigma Y + \mu \le x) = \mathbb{P}\left(Y \le \frac{x-\mu}{\sigma}\right) = F_Y\left(\frac{x-\mu}{\sigma}\right).$$

In the absolutely continuous case let $f_X(x)$ and $f_Y(x)$ denote the density functions of X and Y, respectively. Differentiation gives:

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

General normal distributions

Definition 14.3. *Fix $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. X is normally distributed with expected value μ and variance σ^2 (notation $X \sim \mathcal{N}(\mu, \sigma^2)$) if there exists $Z \sim \mathcal{N}(0,1)$ such that $X = \sigma Z + \mu$.*

If $X \sim \mathcal{N}(\mu, \sigma^2)$, by the above Formulas

$$\begin{aligned} F_X(x) &= \Phi\left(\frac{x-\mu}{\sigma}\right), \\ f_X(x) &= \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \end{aligned}$$

where $\exp(t) = e^t$. See also Figure 6.

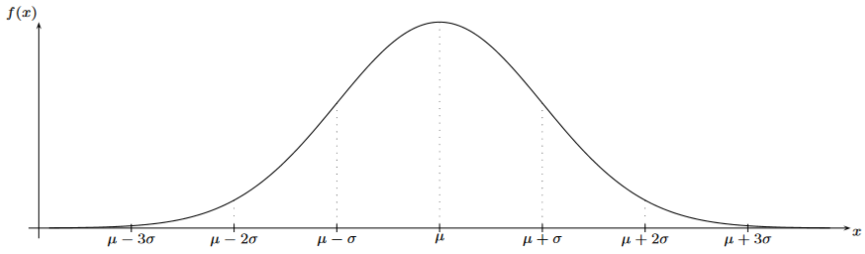

Figure 6: The density function of $\mathcal{N}(\mu, \sigma^2)$

By construction, the class of normal distributions is invariant under rescaling. That is, if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$ then for $X' = aX + b$ we have $X' \sim \mathcal{N}(\mu', (\sigma')^2)$ with $\mu' = a\mu + b$ and $(\sigma')^2 = a^2\sigma^2$.

Normally distributed random variables arise in many situations. The main reason for this is the central limit theorem. Computations of probabilities of normally distributed random variables are based on the standard normal table, as in the example below.

Example 14.4. *The amount of beer served in a pub is normally distributed with expected value 1 pint and standard deviation 2 oz.*

(a) What is the chance that I get less than 13 oz?

(b) I have a tolerance level a oz, which means that I complain iff I get less than a oz. If this happens on the average once out of 20 occasions, what is my tolerance level?

Let X denote the amount of beer measured in oz. Then $X \sim \mathcal{N}(16, 4)$. (Watch out when it is σ and when σ^2!) This means $Z = \frac{X-\mu}{\sigma} = \frac{X-16}{2}$ is standard normal. For part (a):

$$\begin{aligned}\mathbb{P}(X \leq 13) &= \mathbb{P}\left(\frac{X-16}{2} \leq \frac{-3}{2}\right) \approx \mathbb{P}(Z \leq -1.66) = \Phi(-1.66) = 1 - \Phi(1.66) \approx \\ &\approx 1 - 0.9515 = 0.0485,\end{aligned}$$

using the standard normal table.

For (b), we have to search backwards in the table.

$$0.05 = \mathbb{P}(X \leq a) = \mathbb{P}\left(\frac{X-16}{2} \leq \frac{a-16}{2}\right) = \mathbb{P}\left(Z \leq \frac{a-16}{2}\right) = \Phi\left(\frac{a-16}{2}\right)$$

Hence, using the relation $\Phi(-x) = 1 - \Phi(x)$ we get

$$\Phi\left(\frac{16-a}{2}\right) = 0.95 = \Phi(1.65)$$

which, by monotonicity (hence invertibility) of Φ implies

$$\frac{16-a}{2} = 1.65 \iff a = 16 - 2 \cdot 1.65 = 12.7.$$

Further problems of similar character are included in Homework #7.

15 The de Moivre-Laplace Central limit theorem

Let $X(= X^{(n)}) \sim Binom(n, p)$ with p *fixed* and $n \to \infty$. (For example, we keep rolling the *same* die ($p = \frac{1}{6}$) *many* times ($n \to \infty$).)

In such an asymptotic regime (which is, it is important to emphasize, quite different from the Poisson regime!), the distribution is so much "stretched out" that we can hardly observe anything about it if we want to study it directly. In particular $\mathbb{E}X = np$ also tends to ∞. As a first step, let us *center* X and consider:

$$X - \mathbb{E}X = X - np$$

Now the expected value is moved to the origin, yet, the mass function remains stretched out, both along the positive and along the negative semi-axis.

We have discussed earlier *Bernoulli's law of large numbers*, which concerns a rescaled version of our (centered) variable, namely

$$Y = (Y^{(n)}) = \frac{X}{n} - p = \frac{X - np}{n}; \qquad \Longrightarrow \qquad \mathbb{P}(|Y| \geq \varepsilon) \to 0$$

for any $\varepsilon > 0$. This means that if we divide by n, all the randomness is scaled out, the distribution of Y gets concentrated on 0 in the limit. In other words, dividing by n is an overkill.[1] This is not a surprise, as

$$\mathbb{D}(X) = \sqrt{npq} \qquad (q = 1 - p)$$

and thus $\mathbb{D}(X) = o(n)$ (recall the little o notation).

The de Moivre-Laplace theorem concerns the *standardization of* X. The idea is to consider, instead of Y, the variable

$$Z = (Z^{(n)}) = \frac{X - \mathbb{E}X}{\mathbb{D}(X)} = \frac{X - np}{\sqrt{npq}}. \tag{15.1}$$

The theorem states that Z is not only standard, actually, it can be regarded as *standard normal* in the asymptotic as $n \to \infty$. So here is the statement of the theorem.

Theorem 15.1 (de Moivre-Laplace CLT)**.** *Let $X(= X^{(n)}) \sim Binom(n, p)$ with p fixed and $n \to \infty$. Let us consider the standardization* (15.1)*. Then, for any $a < b$ we have*

$$\mathbb{P}\left(a < \frac{X - np}{\sqrt{npq}} \leq b\right) \longrightarrow \Phi(b) - \Phi(a) \qquad \text{as } n \to \infty. \tag{15.2}$$

Although the full proof of this theorem is not provided here, nonetheless, we sketch the main steps. Formula (15.2), which is our ultimate goal, will be refereed to as the *global form of the de Moivre-Laplace theorem.*

Step #1: Stirling's formula. This is to approximate $n!$ for large values of n.

$$n! \approx \frac{n^n}{e^n} \cdot \sqrt{2\pi n}, \quad \text{more precisely} \quad \frac{n! \cdot e^n}{n^n \cdot \sqrt{n}} = C + O\left(\frac{1}{n}\right),$$

where $C = \sqrt{2\pi}$. Here, given two functions $u, v : \mathbb{N} \to \mathbb{R}$, $u(n) = O(v(n))$ (u is big O of v) means that there exists some $K > 0$ such that $u(n) \leq Kv(n)$ (cf. with little o we discussed at the Poisson process). The argument below proves that the (logarithm of the) sequence converges to some constant C, with the error term $O(1/n)$. The fact that $C = \sqrt{2\pi}$ follows form the de Moivre-Laplace theorem itself (namely, the limit has to be a probability distribution, which fixes the constant).

Stirling's formula relies on the observation that for any $k \geq 1$

$$0 \leq a_k = \int_k^{k+1} \ln x dx - \frac{\ln\, k \;+\; \ln\,(k+1)}{2} \leq \frac{1}{k^2}. \tag{15.3}$$

This arises by comparing the integral with the area of the trapezoid. The bound form below follows as the function $\ln(x)$ is concave down. The bound from above follows when

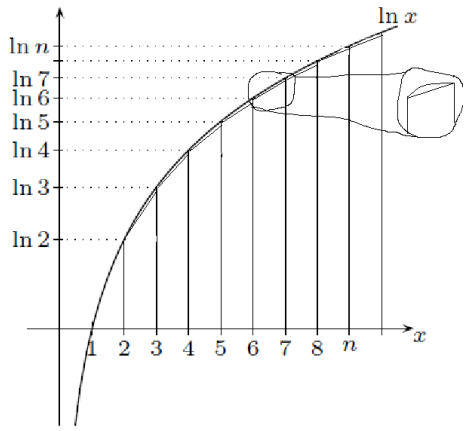

Figure 7: The idea behind Stirling's formula

computing the error term of the trapezoidal rule – the second derivative of $\ln(x)$ is $-\frac{1}{x^2}$. See also Figure 7.

Now let us sum up the expression in (15.3) on $k = 1, ..., (n-1)$. we have:

$$\begin{aligned} S_n &= \sum_{k=1}^{n-1} a_k = \int_1^n \ln x dx - \left(\frac{\ln 1}{2} + \ln 2 + \cdots + \ln(n-1) + \frac{\ln(n)}{2} \right) = \\ &= [x \ln x - x]_1^n - \left(\ln(2 \cdot 3 \cdot \ldots (n-1) \cdot n) - \frac{\ln(n)}{2} \right) = \\ &= n \ln(n) - \ln(n) - \ln(n!) + \ln(\sqrt{n}) = \ln \left(\frac{n^n \cdot \sqrt{n}}{e^n \cdot n!} \right). \end{aligned}$$

Also, (15.3) implies that S_n is a series with positive terms, bounded from above by:

$$S_n \leq \sum_{k=1}^{n-1} \frac{1}{k^2}$$

which is a convergent series. Hence there exists some $S > 0$ such that

$$S = \lim_{n \to \infty} S_n = \lim_{n \to \infty} \ln \left(\frac{n^n \cdot \sqrt{n}}{e^n \cdot n!} \right)$$

Also, by the tail estimate on $\sum \frac{1}{k^2}$:

$$0 \leq S - S_n \leq \sum_{k=n}^{\infty} \frac{1}{k^2} = O\left(\frac{1}{n} \right).$$

[1]Let us note, nonetheless, that in this setting $\frac{X}{n}$, the *empirical success frequency*, is definitely a very reasonable quantity to consider. See the applications further down in this section.

Step #2: The local form of the de Moivre-Laplace theorem. To appreciate this, recall that

$$X \sim Binom(n,p); \qquad Z = \frac{X-np}{\sqrt{npq}}.$$

Accordingly, X can take the values $k = 0, 1, \ldots, n$, with probabilities

$$\mathbb{P}(X=k) = \binom{n}{k} p^k q^{n-k}; \quad k = 0, 1, \ldots, n.$$

Since the standardization Z is an affine linear rescaling of X, for any k there exists exactly one $z_k \in \mathbb{R}$ such that $X = k$ iff $Z = z_k$:

$$\begin{array}{ll} \text{values of } X: & 0\,,\, 1\,, \ldots,\, k\,, \ldots, n \\ & \Updownarrow, \Updownarrow, \ldots, \Updownarrow, \ldots, \Updownarrow \\ \text{values of } Z: & z_0, z_1, \ldots, z_k, \ldots, z_n. \end{array}$$

In fact, $z_k = \frac{k-np}{\sqrt{npq}}$. The local form of the de Moivre Laplace theorem states that

$$\mathbb{P}(Z=z_k) = \mathbb{P}(X=k) = \binom{n}{k} p^k q^{n-k} \;=\; \frac{\varphi(z_k)}{\sqrt{npq}} + O\left(\frac{1}{n}\right) \tag{15.4}$$

for any $k \geq 1$ such that $|k-np| = O(\sqrt{n})$. (15.4) follows from Stirling's formula by some computation, which we do not detail here.

Step #3: Local form implies global form. We use the notation from Step #2. Note that

$$\Delta z = z_k - z_{k-1} = \frac{1}{\sqrt{npq}}, \quad k = 1, 2, \ldots, n.$$

Now, using (15.4):

$$\mathbb{P}(a < Z \leq b) = \sum_{k: a < z_k \leq b} \mathbb{P}(Z = z_k) = \sum_{k: a < z_k \leq b} \frac{\varphi(z_k)}{\sqrt{npq}} + O\left(\frac{1}{\sqrt{n}}\right), \tag{15.5}$$

The bound of the second term here follows as there are $O(\sqrt{n})$ terms of magnitude $O(1/n)$. Consequently, the second term is $O\left(\frac{1}{\sqrt{n}}\right)$, which tends to 0 as $n \to \infty$.

On the other hand, the first term in the RHS of (15.5) is a Riemann sum, in particular

$$\sum_{k: a < z_k \leq b} \frac{\varphi(z_k)}{\sqrt{npq}} = \sum_{k: a < z_k \leq b} \varphi(z_k) \cdot \Delta z \;\longrightarrow \int_a^b \varphi(z) dz = \Phi(b) - \Phi(a),$$

as $n \to \infty$, which is precisely (15.2).

Applications

Example 15.2. *A fair coin is flipped* 40 *times. What is the chance that there are exactly* 20 *Heads?*

It may be argued that $n = 40$ is not that large, yet, we will see that the approximation given by the de Moivre-Laplace CLT is not that far from the actual probability. Let X denote that number of Heads. Then:

$$X \sim Binom(n, p) = Binom(40, 0.5)$$

hence

$$\mathbb{E}X = 40 \cdot 0.5 = 20; \qquad \mathbb{D}(X) = \sqrt{40 \cdot 0.5 \cdot 0.5} = \sqrt{10},$$

hence the standardization is $Z = \frac{X-20}{\sqrt{10}}$ It is important to point out that we approximate a discrete distribution with a continuous distribution. Hence, literally speaking, for any k the chance of $\mathbb{P}(X = k)$ is 0 by this approximation. To handle this issue, we assign to each of the values $X = k$, the interval $X \in (k-0.5, k+0.5]$. This is called the *continuity correction.* Now

$$\begin{aligned}
\mathbb{P}(X = 20) &= \mathbb{P}(19.5 < X \leq 20.5) = \mathbb{P}\left(-\frac{0.5}{\sqrt{10}} < \frac{X-20}{\sqrt{10}} \leq \frac{0.5}{\sqrt{10}}\right) \approx \\
&\approx \Phi\left(\frac{0.5}{\sqrt{10}}\right) - \Phi\left(-\frac{0.5}{\sqrt{10}}\right) = 2\Phi\left(\frac{0.5}{\sqrt{10}}\right) - 1 \approx \\
&\approx 2\Phi(0.16) - 1 \approx 2 \cdot 0.5636 - 1 = 0.1272.
\end{aligned}$$

In fact

$$\mathbb{P}(X = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} = 0.1254...$$

so the approximation is pretty good. It is worth noting for future reference:

$$Z \sim \mathcal{N}(0,1) \implies \mathbb{P}(|Z| \leq a) = \mathbb{P}(-a \leq Z \leq a) = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1, \qquad (15.6)$$

for any $a > 0$, by $\Phi(-a) = 1 - \Phi(a)$.

Example 15.3. *Consider a college with* 400 *students. Each of the students are expected to show up at a lecture with* 0.6 *chance, independently of each other. How many seats are needed to ensure, with* 99% *probability, that every student who shows up at the lecture can take a seat?*

It is important to observe how this question is formulated. If we wanted to ensure that all attendees can take a seat with 100% *probability*, then apparently 400 seats would be needed. There is a possibility that all the 400 students show up, however, the probability of this is tiny. If we want to go only for 99%, that changes the picture significantly. Let X denote the number of students who attend the lecture. Then

$$X \sim Binom(n, p) = Binom(400, 0.6)$$

hence

$$\mathbb{E}X = 400 \cdot 0.6 = 240; \qquad \mathbb{D}(X) = \sqrt{400 \cdot 0.6 \cdot 0.4} = \sqrt{96} \approx 9.8,$$

and we are looking for the smallest $k \geq 1$ such that

$$\mathbb{P}(X \leq k) \geq 0.99.$$

Now

$$\mathbb{P}(X \leq k) = \mathbb{P}\left(\frac{X - 240}{9.8} \leq \frac{k - 240}{9.8}\right) = \Phi\left(\frac{k - 240}{9.8}\right)$$

So we want

$$\Phi\left(\frac{k - 240}{9.8}\right) \geq 0.99 = \Phi(2.32)$$

which, by monotonicity of Φ is equivalent to

$$\frac{k - 240}{9.8} \geq 2.32 \iff k \geq 262.736;$$

so 263 seats will do!

Example 15.4. *How many times a fair coin has to be flipped to ensure that the proportion of Heads is between* 0.49 *and* 0.51*, with* 0.95 *probability?*

Again, observe how the question is formulated. By the Law of Large Numbers, the proportion of Heads converges to 0.5 in the following sense. Given any *precision* level ε, as the *number of trials* n grows, the proportion of Heads will lie in the ε neighborhood of 0.5, with higher and higher *certainty* (or probability) $1 - \delta$. The de Moivre-Laplace CLT this way provides a quantitative refinement of the Bernoulli Law of Large Numbers. We can never be 100% sure, no matter how many trials are made, there is always a tiny-tiny chance that all flips are Tails (for instance).

There are three quantities in relation here, the number of trials n, the precision ε and the certainty $1 - \delta$. If two of these are given, the third can be determined. The question above is formulated in such a way that we are looking for a lower bound on the number of trials n that ensures a particular precision with a particular certainty.

Let X denote the number of Heads. Then

$$X \sim Binom(n, p) = Binom(n, 0.5)$$

hence

$$\mathbb{E}X = n \cdot 0.5; \qquad \mathbb{D}(X) = \sqrt{n \cdot 0.5 \cdot 0.5} = \sqrt{n \cdot 0.25} = 0.5\sqrt{n}.$$

Now n is unknown, and we want that

$$\begin{aligned} 0.95 \quad \leq \quad & \mathbb{P}\left(0.49 \leq \frac{X}{n} \leq 0.51\right) = \mathbb{P}(|X - 0.5n| \leq 0.01n) \\ & \mathbb{P}\left(\left|\frac{X - 0.5n}{0.5\sqrt{n}}\right| \leq 0.02\sqrt{n}\right) = 2\Phi(0.02\sqrt{n}) - 1, \end{aligned}$$

where we have used (15.6). This is equivalent to

$$\Phi(0.02\sqrt{n}) \geq 0.975 = \Phi(1.96) \iff 0.02\sqrt{n} \geq 1.96 \iff n \geq 9604.$$

I will comment in class how this relates to survey making (cf. Homework #7B).

16 The exponential distribution

Let $\lambda > 0$. The random variable X is exponentially distributed with parameter λ (notation: $X \sim Exp(\lambda)$) if

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

and, accordingly

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

See also Figure 8. Note that we have already seen this distribution in Example 8.5, part (b): this is the distribution of the time that elapses until the first impact in the Poisson process. The exponential distribution arises frequently as a model for various *waiting times*.

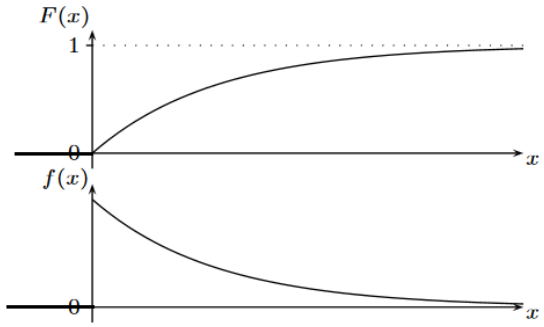

Figure 8: Distribution function and density for $X \sim Exp(\lambda)$.

Computing moments is an exercise on integration by parts:

$$\mathbb{E}X = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-xe^{-\lambda x}\right]_0^\infty + \frac{1}{\lambda}\int_0^\infty \lambda e^{-\lambda x} dx = \frac{1}{\lambda};$$

and for any $k \geq 2$

$$\mathbb{E}X^k = \int_0^\infty x^k \lambda e^{-\lambda x} dx = \left[-x^k e^{-\lambda x}\right]_0^\infty + \frac{k}{\lambda}\int_0^\infty \lambda x^{k-1} e^{-\lambda x} dx = \frac{k}{\lambda}\mathbb{E}X^{k-1}.$$

The recursion gives, in particular

$$\mathbb{E}X^2 = \frac{2}{\lambda^2}, \qquad \text{thus } Var(X) = \frac{1}{\lambda^2} \quad \text{and} \quad \mathbb{D}(X) = \frac{1}{\lambda}.$$

A remarkable feature of the exponential distribution is the *memoryless property* (cf. the analogous property we have seen for the geometric random variable). Let $t > 0$ and $s > 0$ be arbitrary. Then

$$\begin{aligned} \mathbb{P}(X > s) &= e^{-\lambda s}; \\ \mathbb{P}(X > t + s \,|\, X > t) &= \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(X > s). \end{aligned}$$

In words: no matter how long we have already been waiting, the chance that we have to wait for at least another s units of time remains the same.

17 The distribution of a function of a random variable

Example 17.1. *Let $X \sim Exp(\lambda)$, and let $Y = X^2$. Determine the density $f_Y(y)$ of Y.*

We determine first the cumulative distribution function of Y. Note first that as $\mathbb{P}(X \geq 0) = 1$, we have $\mathbb{P}(Y \geq 0) = 1$, hence $F_Y(y) = 0$ for $y < 0$. Let us now consider $y \geq 0$:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(X \leq \sqrt{y}) = 1 - e^{-\lambda\sqrt{y}}.$$

Now differentiation with respect to y gives:

$$f_Y(y) = \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{\lambda}{2\sqrt{y}} \cdot e^{-\lambda\sqrt{y}} & \text{if } y > 0. \end{cases}$$

Proposition 17.2. *Let X have density function $f_X(x)$ and let us consider $Y = k(Y)$ such that $k : \mathbb{R} \to \mathbb{R}$ is*

- *continuously differentiable*
- *strictly monotonic (either strictly decreasing, or strictly increasing).*

Note that in this case $k : \mathbb{R} \to \mathbb{R}$ is invertible – let us denote its inverse by $k^{-1}(y)$. The density function of Y can be computed by the following formula:

$$f_Y(y) = f_X(k^{-1}(y)) \cdot \left| \frac{d(k^{-1}(y))}{dy} \right| \tag{17.1}$$

Comments:

- The *support of* X is the set

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

We have $\mathbb{P}(X \in supp\, X) = 1$. For Proposition 17.2, it is enough that $k : \mathbb{R} \to \mathbb{R}$ is strictly monotonic restricted to $supp\, X$. This is the case in Example 17.1.

- Instead of the formal proof, we give a heuristic argument (which, actually, can be made rigorous) to explain Formula (17.1). Recall that the value $f_X(x)$ is the scaling factor by which the probability of falling in an interval of length dx about x decays as $dx \to 0$. Now the preimage of some small interval $[y, y + dy]$ by $k : \mathbb{R} \to \mathbb{R}$ is another small interval $[x, x + dx]$ where

$$y = k(x) \;\Leftrightarrow\; x = k^{-1}(y); \qquad \frac{dy}{dx} = \left|\frac{dk(x)}{dx}\right| \;\Leftrightarrow\; \frac{dx}{dy} = \left|\frac{d(k^{-1}(y))}{dy}\right|.$$

Accordingly

$$f_Y(y)dy \approx \mathbb{P}(y \le Y \le y + dy) = \mathbb{P}(x \le X \le x + dx) \approx f_X(x)dx$$

and division by dy yields (17.1). See Figure 9 for the case when $Y = X^2$, and X is supported on the positive halfline.

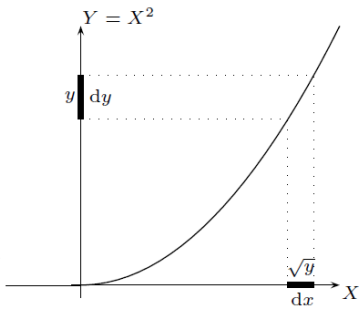

Figure 9: Density transform formula.

Example 17.3 (Standard Cauchy distribution)**.** *Let* $\Theta \sim UNI\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ *and* $X = \tan\Theta$.

We have

$$f_\Theta(\vartheta) = \begin{cases} \frac{1}{\pi} & \text{if } \vartheta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$x = k(\vartheta) = \tan(\vartheta) \Leftrightarrow \vartheta = k^{-1}(x) = \tan^{-1}(x) \qquad \text{so} \qquad \frac{d(\tan^{-1}(x))}{dx} = \frac{1}{1+x^2}$$

and thus

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

See section 6.3, in particular Figure 7 in the Ross book for a geometric problem in which the standard Cauchy distribution arises. An interesting feature of this random variable is that $\mathbb{E}X$ does not exist.

18 Independent random variables

Recall that we have already introduced the notion of independence for discrete random variables when we discussed the Poisson process. Here this is generalized to arbitrary random variables.

Definition 18.1. *Let X and Y be arbitrary random variables. X and Y are independent if for any pair of intervals I_1 and I_2 we have*

$$\mathbb{P}(X \in I_1 \text{ and } Y \in I_2) = \mathbb{P}(X \in I_1) \cdot \mathbb{P}(X \in I_2).$$

Independence means that

$$\mathbb{P}(Y \in I_2 \,|\, X \in I_1) = \mathbb{P}(Y \in I_2).$$

That is, no matter how we fix the value of X, the conditional distribution of Y remains unaffected. This behavior is in strong contrast with the case of $Y = k(X)$ studied in the previous section, when fixing the value of X determined Y entirely. There is a large scale of intermediate cases. In other words, the joint distribution of X and Y contains much more information than just the distributions os X and Y separately (the marginal distributions). Unfortunately, we do not have the time to elaborate on joint distributions in detail.

From now on, we mostly (but not entirely) focus on the independent case. If the variables X and Y are independent, then their marginal distributions determine their joint distribution, and questions on their joint behavior can be addressed in terms of them. For example, let X and Y have cumulative distribution functions $F_X(x)$ and $F_Y(x)$, respectively, and let $W = \max(X, Y)$. Then, for any $x \in \mathbb{R}$:

$$F_W(x) = \mathbb{P}(W \leq x) = \mathbb{P}(X \leq x \text{ and } Y \leq x) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq x) = F_X(x) \cdot F_Y(x).$$

Sums of independent random variables

In this section the following question is studied: if the distributions of the independent random variables X and Y are known, how is their sum $Z = X + Y$ distributed? Let us consider first a discrete example.

Proposition 18.2. *Let X and Y be independent, $X \sim Poi(\lambda)$ and $Y \sim Poi(\mu)$. Then $X + Y \sim Poi(\lambda + \mu)$.*

For any $k \geq 0$ we have:

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{i=-\infty}^{\infty} \mathbb{P}(X+Y=k|X=i)\mathbb{P}(X=i) = \\
&= \sum_{i=-\infty}^{\infty} \mathbb{P}(Y=k-i|X=i)\mathbb{P}(X=i) = \sum_{i=-\infty}^{\infty} \mathbb{P}(Y=k-i)\mathbb{P}(X=i) = \\
&= \sum_{i=0}^{k} e^{-\lambda}\frac{\lambda^i}{i!}e^{-\mu}\frac{\mu^{k-i}}{(k-i)!} = \\
&= e^{-(\lambda+\mu)}\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda^i\mu^{k-i} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}.
\end{aligned}$$

Hence we see that the one parameter family of Poisson distributions is closed under adding independent copies. The terminology is that the family of Poisson distributions is stable, which is rather exceptional (see the case of uniform distributions below).

Now we move on to absolutely continuous examples. In this case X and Y have some density functions $f_X(x)$ and $f_Y(y)$, respectively. The density function of $X + Y$ is given by the Formula:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

In real analysis, this is called the *convolution* of the two functions f_X and f_Y. See Ross, section 6.3 for a proof. As an application, we have (see again Ross, section 6.3 for the computation):

Example 18.3. *Let X and Y be independent, both $X \sim UNI[0,1]$ and $Y \sim UNI[0,1]$. Then the density of $X + Y$ is:*

$$f_{X+Y}(t) = \begin{cases} t & \text{if } t \in [0,1], \\ 2-t & \text{if } t \in [1,2], \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 10.

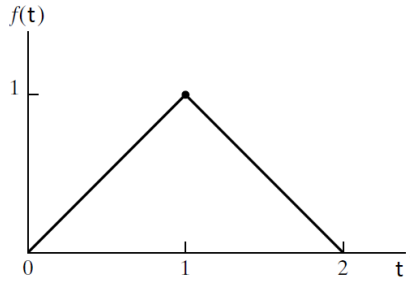

Figure 10: Convolution of two i.i.d. $UNI[0,1]$.

Finally, as most important application, let us consider the sums of independent normally distributed random variables.

Proposition 18.4. *Let X_1 and X_2 be independent, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then*

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Recall Lemmas 10.1 and 10.2, which extend from the discrete case to general random variables. Hence it is not surprising that $\mathbb{E}(X_1 + X_2) = \mu_1 + \mu_2$ and $Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2$. What is, however, specific to normal distributions is that this family is stable under convolutions. See Ross, section 6.3.3 for the computation.

In the context of Proposition 18.4 we have $-X_2 \sim \mathcal{N}(-\mu_2, \sigma_2^2)$, and thus

$$X_1 - X_2 = X_1 + (-X_2) \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2); \tag{18.1}$$

that is, the expected values are subtracted, but the variances add.

Proposition 18.4 has many applications, which may easily come up in Homework and Exam problems. Here is an example.

Example 18.5. *In a bottle of mineral water the amount of sodium has expected value* 35 *mg and standard deviation* 3 *mg, while the amount of calcium has expected value* 55 *mg and standard deviation* 4 *mg. The amounts of these minerals can be regarded independent and normally distributed. What is the probability that in my bottle (a) the total amount of the two minerals exceeds* 100 *mg; (b) there is more sodium than calcium?*

Let X denote the amount of sodium, and Y denote the amount of calcium in my bottle, measured in miligrams. Then $X \sim \mathcal{N}(35, 9)$ while $Y \sim \mathcal{N}(55, 16)$, and they are independent. Hence, by Proposition 18.4 and (18.1)

$$X + Y \sim \mathcal{N}(90, 25) \qquad \text{and } X - Y \sim \mathcal{N}(-20, 25).$$

To solve (a):

$$\begin{aligned}\mathbb{P}(X+Y\geq 100) &= 1-\mathbb{P}(X+Y\leq 100)=1-\mathbb{P}\left(\frac{X+Y-90}{5}\leq 2\right)= \\ &= 1-\Phi(2)=1-0.9772=0.0228.\end{aligned}$$

To solve (b):

$$\begin{aligned}\mathbb{P}(X>Y) &= \mathbb{P}(X-Y>0)=1-\mathbb{P}(X-Y\leq 0)= \\ &= 1-\mathbb{P}\left(\frac{X-Y+20}{5}\leq 4\right)=1-\Phi(4)\approx 0.\end{aligned}$$

19 Further indicator problems

Recall the problems discussed in section 10.1. Here we include some further examples. Indicator variables are useful for some problems in Homework #8 and may occur on the Final Exam as well. Recall that the idea is to represent some discrete random variable X as $X=\eta_1+\cdots+\eta_K$, where the η_i are indicators. Then $\mathbb{E}X=\sum\mathbb{E}\eta_i$. If, furthermore, the indicators are independent, then $Var(X)=\sum Var(\eta_i)$.

Example 19.1. *Let us recall the matching problem: N phones are distributed among N people. Let X denote the number of people who are matched with their phones. Let us determine $\mathbb{E}X$ and $Var(X)$.*

Recall that at the beginning of the semester we determined $\mathbb{P}(X=0)$ by the inclusion-exclusion formula. Now let

$$\eta_i=\begin{cases}1 & \text{if person } \#i \text{ is matched with her/his phone,}\\ 0 & \text{otherwise,}\end{cases}\qquad i=1,\dots N.$$

Then $X=\eta_1+\cdots+\eta_N$. Also, for any $i=1,\dots,N$ we have:

$$\mathbb{E}\eta_i=\mathbb{E}\eta_1=\mathbb{P}(\eta_1=1)=\frac{(N-1)!}{N!}=\frac{1}{N}$$

and thus

$$\mathbb{E}X=\sum_{i=1}^{N}\mathbb{E}\eta_i=N\frac{1}{N}=1.$$

To proceed to $Var(X)$, note that these indicators are *not independent*. So we rely on

the following pattern:

$$\begin{aligned} Var(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \ (= \mathbb{E}X^2 - 1); \\ X^2 &= \sum_{i=1}^{N} \eta_i^2 + \sum_{i<j} 2\eta_i\eta_j; \\ \mathbb{E}\eta_i^2 &= \mathbb{E}\eta_i \ (= 1/N); \quad (i = 1, \dots, N) \\ \mathbb{E}(\eta_i\eta_j) &= \mathbb{P}(\eta_i = 1 \text{ and } \eta_j = 1); \quad i < j \\ \mathbb{E}X^2 &= \sum_{i=1}^{N} \mathbb{E}\eta_i^2 + \sum_{i<j} 2\mathbb{E}\eta_i\eta_j \ \left(= 1 + 2\binom{N}{2}\mathbb{P}(\eta_1 = 1 \text{ and } \eta_2 = 1)\right). \end{aligned} \tag{19.1}$$

Here $\sum_{i<j}$ is the sum on all pairs such that $i < j$, so there are $\binom{N}{2}$ terms. Now

$$\{\eta_i = 1 \text{ and } \eta_j = 1\} = \{\text{both } \#i \text{ and } \#j \text{ are matched with their phones}\},$$

so

$$\mathbb{E}\eta_i\eta_j = \mathbb{P}(\eta_i = 1 \text{ and } \eta_j = 1) = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)}$$

and thus, using (19.1)

$$\mathbb{E}X^2 = 1 + 2\binom{N}{2}\frac{1}{N(N-1)} = 2; \implies Var\, X = \mathbb{E}X^2 - 1 = 2 - 1 = 1.$$

Example 19.2. *A fair coin is flipped* 100 *times. The outcome of this experiment is a sequence of the type* $THHTTTHTTHHT\dots$ *A run of Heads is a maximal subsequence of consecutive symbols* H*. Let* X *denote the number of runs in the sequence. Determine* $\mathbb{E}X$*.*

Let

$$\eta_i = \begin{cases} 1 & \text{if a run of Heads starts at position } \#i, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \dots 100.$$

Then

$$\mathbb{E}\eta_1 = \mathbb{P}(\eta_1 = 1) = \mathbb{P}(\text{the first symbol is } H) = \frac{1}{2},$$

however, for $i = 2, \dots, 100$:

$$\begin{aligned} \mathbb{E}\eta_i &= \mathbb{P}(\eta_i = 1) = \\ &= \mathbb{P}\left(\text{the symbols at positions } \#(i-1) \text{ and } \#i \text{ are } TH, \text{ respectively}\right) = \frac{1}{4}, \end{aligned}$$

So:

$$\mathbb{E}X = \sum_{i=1}^{100} \mathbb{E}\eta_i = \mathbb{E}\eta_1 + \sum_{i=2}^{100} \mathbb{E}\eta_i = \frac{1}{2} + \frac{99}{4} = 25.25.$$

See sections 7.2, 7.3 and 7.4 of the Ross book for further applications of indicator variables.

20 Covariance

Definition 20.1 (Covariance)**.** *Consider two random variables X and Y, and let us denote by $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$ their expected values. The covariance of X and Y is defined as*

$$Cov(X,Y) = \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right).$$

Alternative formula:

$$Cov(X,Y) = \mathbb{E}(X\cdot Y - \mu_Y\cdot X - \mu_X\cdot Y + \mu_X\cdot\mu_Y) = \mathbb{E}(X\cdot Y) - \mu_X\cdot\mu_Y = \mathbb{E}(X\cdot Y) - \mathbb{E}X\cdot\mathbb{E}Y.$$

Lemma 20.2. *Let X and Y be independent. Then $Cov(X,Y) = 0$.*

The proof of Lemma 20.2 is just a reference to (10.2), which extends form discrete to general random variables. However, it is important to note that the converse of this statement is, in general, false, as the following counterexample shows.

Example 20.3. *Let X take values $-1, 0$ and $+1$, each with probability $\frac{1}{3}$, and let $Y = X^2$. Then $Cov(X,Y) = 0$. Yet, Y is a function of X, so they are NOT independent.*

To see this, note that apparently $\mathbb{E}X = 0$, while $X\cdot Y = X$ (since $Y = 1$ if and only if $X \neq 0$), hence $\mathbb{E}(X\cdot Y) = \mathbb{E}X = 0$.

Covariance has the following properties:

- *symmetric*: apparently $Cov(X,Y) = Cov(Y,X)$;
- *positive definite*: $Cov(X,X) = Var(X) \geq 0$, moreover $Var(X) = 0$ only in the trivial case when X is deterministically equal to the constant $\mathbb{E}X$.
- *bilinear*: by symmetry, it is enough to check that, given random variables X_1, X_2, Y, and constants $a_1, a_2, b \in \mathbb{R}$:

$$Cov(a_1X_1 + a_2X_2 + b, Y) = a_1\,Cov(X_1,Y) + a_2\,Cov(X_2,Y). \tag{20.1}$$

  This follows as

$$\mathbb{E}(a_1X_1 + a_2X_2 + b) = a_1\mu_1 + a_2\mu_2 + b$$

  (where $\mu_i = \mathbb{E}X_i$, $i = 1, 2$) and thus

$$a_1X_1 + a_2X_2 + b - \mathbb{E}(a_1X_1 + a_2X_2 + b) = a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2)$$

  so (20.1) follows form the definition of covariance and the linearity of expectation.

You may recall from linear algebra that these properties characterize inner products. Indeed, covariance can be regarded as an inner product on some linear space.

A further consequence of bilinearity is that, given arbitrary random variables $X_1, \ldots, X_n$, we have

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j). \end{aligned}$$

Definition 20.4 (Correlation coefficient)**.** *The correlation coefficient of two random variables X and Y is defined as:*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\mathbb{D}X \cdot \mathbb{D}Y}$$

Proposition 20.5. *For arbitrary random variables X and Y:*

$$-1 \le \rho(X,Y) \le 1.$$

Moreover $|\rho(X,Y)| = 1$ if and only if X and Y are in exact linear relation. More precisely:

$$\rho(X,Y) = 1 \iff Y = m \cdot X + d \quad \text{for some } m > 0 \text{ and } b \in \mathbb{R}, \tag{20.2}$$

$$\rho(X,Y) = -1 \iff Y = -m \cdot X + d \quad \text{for some } m > 0 \text{ and } b \in \mathbb{R}. \tag{20.3}$$

Proof. Let us introduce $\sigma_X = \mathbb{D}X$ and $\sigma_Y = \mathbb{D}Y$ for brevity. Let, furthermore,

$$Z = \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}; \qquad W = \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}.$$

Then, by bilinearity of covariance:

$$Var(Z) = \frac{Var\, X}{\sigma_X^2} + \frac{Var\, Y}{\sigma_Y^2} + 2\frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 1 + 1 + 2\rho(X,Y)$$

hence

$$\rho(X,Y) = -1 \iff Var\, Z = 0 \iff Z = \text{const},$$

which is equivalent to (20.3). Similarly,

$$Var(W) = \frac{Var\, X}{\sigma_X^2} + \frac{Var\, Y}{\sigma_Y^2} - 2\frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 1 + 1 - 2\rho(X,Y),$$

hence

$$\rho(X,Y) = 1 \iff Var\, W = 0 \iff W = \text{const},$$

which is equivalent to (20.2). □

Comments:

- So what correlation actually measures is the extent of *linear* dependence among the two random variables.
- Note

  $$|\rho(X,Y)| \le 1 \iff |Cov(X,Y)| \le \sqrt{Var(X)} \cdot \sqrt{Var(Y)}, \tag{20.4}$$

  which is the Cauchy-Schwartz inequality.

21 Tower rule

Given two random variables X and Y, we may consider, for any $x_0 \in \mathbb{R}$, the *conditional distribution of* Y given that $X = x_0$, and accordingly, the *conditional expectation* $\mathbb{E}(Y|X = x_0)$.

$\mathbb{E}(Y|X = x_0)$ has a double character: on the one hand, it is an average (an expectation), on the other hand, it depends on x_0. In a sense "half of the randomness" is integrated out, the "other half" associated to the value of X, is still undetermined. Accordingly, we may just write $\mathbb{E}(Y|X)$, which is a random variable, a function of X.

Now, we may take the expected value of this later random variable, which we denote by $\mathbb{E}(\mathbb{E}(Y|X))$. The *tower rule* states that the expected value $\mathbb{E}Y$ can be computed in two steps as

$$\mathbb{E}Y = \mathbb{E}(\mathbb{E}(Y|X)).$$

An analogy is the law of total probability. There is a lot more to say about the tower rule and conditional expectations, this topic is to be revisited in the enrichment period. Let us just include here two applications.

Example 21.1. *Let T be uniformly distributed on $[0, 10]$. A detector is functional for a random time interval which has length T. During that interval, it detects the impacts of particles, while no particles are detected before and after this time interval. The particles arrive according to a Poisson process of intensity 2. Let X denote the number of particles detected. $\mathbb{E}X = ?$*

Solution: Note that if $T = t$, the conditional distribution is $X|T = t \sim Poi(2t)$. Hence $\mathbb{E}(X|T = t) = 2t$, or briefly $\mathbb{E}(X|T) = 2T$. Now, by the tower rule

$$\mathbb{E}X = \mathbb{E}(\mathbb{E}(X|T)) = \mathbb{E}(2T) = 2\mathbb{E}T = 2 \cdot 5 = 10$$

where we have used that $T \sim UNI[0, 10]$.

Example 21.2. *In a casino, the following game is played: a fair die is rolled until a 6 occurs for the first time, when the player earns 20 dollars. However, each time the value i is rolled prior to that 6, the player has to pay i dollars. What is the expected gain of the player in such a game?*

Let X denote the gain of the player. Try to find $\mathbb{E}X$ using the tower rule with one of the following two auxiliary random variables:

Solution #1 Introduce N, the number of rolls.

Solution #2 Introduce Y, the value obtained at the first roll.

22 Moment generating functions

Definition 22.1. *Let X be a random variable. The moment generating function of X is defined as*

$$M_X : \mathbb{R} \to \mathbb{R}; \qquad t \mapsto M_X(t) = \mathbb{E}(e^{tX}).$$

If there is no ambiguity we may drop the subscript X and simply write $M(t)$.

Comment: The integral may not be finite for all $t \in \mathbb{R}$. What we assume is that there exists some $t_0 > 0$ such that $M_X(t) < \infty$ for $t \leq t_0$. This is indeed the case of all major distributions discussed here (eg. Poisson, normal, exponential etc.)
Basic properties:

- $M(0) = 1$,
- Let prime denote differentiation w.r.t. t. Then

$$\begin{aligned} M'(t) = \mathbb{E}(Xe^{tX}) &\implies M'(0) = \mathbb{E}X \\ M''(t) = \mathbb{E}(X^2e^{tX}) &\implies M''(0) = \mathbb{E}(X^2) \end{aligned}$$

  and proceeding with this, for the nth derivative at 0:

$$M^{(n)}(0) = \mathbb{E}(X^n).$$

- Let X and Y be independent. Then

$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX}) \cdot \mathbb{E}(e^{tY}) = M_X(t) \cdot M_Y(t). \tag{22.1}$$

- Let Z be the standardization of X, that is, $X = \sigma Z + \mu$. Then

$$M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}(e^{t(\sigma Z+\mu)}) = e^{t\mu} \cdot \mathbb{E}(e^{\sigma tZ}) = e^{t\mu} M_Z(t\sigma). \tag{22.2}$$

Let $X \sim Poi(\lambda)$. Then

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \exp(\lambda e^t) = \exp(\lambda(e^t - 1)).$$

Let $Z \sim \mathcal{N}(0, 1)$. Then

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{tx} \varphi(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} + tx - \frac{t^2}{2} + \frac{t^2}{2}\right) dx = \\ &= e^{t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-t)^2\right) dx = e^{t^2/2}. \end{aligned} \tag{22.3}$$

This implies, by (22.2):

$$X \sim \mathcal{N}(\mu, \sigma^2) \iff M_X(t) = e^{\mu t} \cdot e^{\frac{t^2\sigma^2}{2}}.$$

The *logarithmic moment generating function* of X is defined as

$$\Psi_X(t) = \ln(M_X(t))$$

and has the following properties:

$$\begin{aligned} & & \Psi_X(0) &= 0, \\ \Psi'_X(t) = \frac{M'_X(t)}{M_X(t)} & \implies & \Psi'_X(0) &= \frac{M'_X(0)}{M_X(0)} = \mathbb{E}X, \\ \Psi''_X(t) = \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{(M_X(t))^2} & \implies & \Psi''_X(0) &= \frac{M''_X(0) - (M'_X(0))^2}{1} = Var\, X. \end{aligned} \tag{22.4}$$

23 The Weak Law of Large Numbers

23.1 Useful inequalities

Proposition 23.1 (**Markov's inequality**). *Let X be a nonnegative random variable – i.e. $\mathbb{P}(X \geq 0) = 1$ – and let $\mu = \mathbb{E}X$. Then, for any $a > 0$:*

$$\mathbb{P}(X \geq a) \leq \frac{\mu}{a}.$$

Proof. Let us fix $a > 0$, and introduce the indicator variable

$$\eta_a = \begin{cases} 1 & \text{if } X \geq a, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$a \cdot \eta_a \leq X \qquad \implies \qquad a\mathbb{P}(X \geq a) = \mathbb{E}(a \cdot \eta_a) \leq \mathbb{E}X = \mu,$$

and division by a results in Markov's inequality.

□

Comment. This inequality is void if $a \leq \mu$, and it is useful for $a \gg \mu$. For example, it states that the probability that X takes values ten times greater than $\mathbb{E}X$ cannot exceed 0.1. It is remarkable that this bound holds irrespective of the actual distribution of X.

Proposition 23.2 (**Chebyshev's inequality**). *Let X be an arbitrary random variable and let $\mu = \mathbb{E}X$, $\sigma^2 = Var(X)$. Then, for any $b > 0$:*

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

Proof. Let us introduce $Y = (X - \mu)^2$ which is a nonnegative random variable. Hence we may apply Markov's inequality with b^2 playing the role of a. Note also $\mathbb{E}Y = Var(X) = \sigma^2$.

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

□

Again, Chebyshev's inequality is useful when $b \gg \sigma$. It is widely applicable as it holds irrespective of the distribution of X. Accordingly, the bounds are often not very sharp. For example, by Chebysev's inequality

$$\mathbb{P}(|X - \mu| \geq 3\sigma) \leq \frac{1}{9} \approx 0.111.$$

However, if we know that X is normally distributed ($X \sim \mathcal{N}(\mu, \sigma^2)$) then, using (15.6)

$$\mathbb{P}(|X - \mu| \leq 3\sigma) = \mathbb{P}(|Z| \leq 3) = 2\Phi(3) - 1 = 0.9974 \implies \mathbb{P}(|X - \mu| \geq 3\sigma) = 0.0036,$$

which is far less.

23.2 The Weak Law of Large Numbers

Setting. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of *independent, identically distributed* (in short: i.i.d.) random variables. This is highly relevant, for example, for (idealized) statistics applications. As the variables have the same distribution, they also have the same expected values and variances, which we denote as

$$\mu = \mathbb{E}X_1 (= \mathbb{E}X_n); \qquad \sigma^2 = Var(X_1) (= Var(X_n)); \quad (\forall n \geq 1).$$

Another notation:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Theorem 23.3 (**Weak Law of Large Numbers**). *Let us consider an i.i.d. sequence with the notations introduced above. Let $\varepsilon > 0$ be arbitrary. Then*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \longrightarrow 0 \qquad \text{as } n \to \infty.$$

Interpretation. In words, the sequence of random variables $\frac{S_n}{n}$ – the "empirical mean with growing sample size" – converges to μ in probability. There are other relevant forms of convergence which are, actually, stronger than convergence in probability. This is the reason for the terminology "weak law of large numbers".

Proof. We have

$$\mathbb{E}S_n = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = n\mu$$

and by independence also

$$Var(S_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2.$$

We will apply Chebyshev's inequality to S_n.

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = \mathbb{P}(|S_n - n\mu| \geq n\varepsilon) \leq \frac{Var(S_n)}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. □

Comments.

- Consider a sequence of independent trials. Let, for $i \geq 1$,

$$X_i = \eta_{A_i} = \begin{cases} 1 & \text{if the } i\text{th trial is a success,} \\ 0 & \text{if the } i\text{th trial is a failure,} \end{cases}$$

that is, the indicator variable associated to the ith success. These random variables are i.i.d. Also $S_n \sim Binom(n, p)$ and $\mathbb{E}X_1 (= \mathbb{E}X_i) = p$. Hence, in this special case, the WLLN states that for any $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \longrightarrow 0, \qquad \text{as } n \to \infty,$$

which is just Bernoulli's Law of Large Numbers.

- Taking a look at the computation it turns out that we could have proved something stronger. Namely, instead of

$$\frac{S_n}{n} - \mu = \frac{S_n - n\mu}{n} = \frac{S_n - \mathbb{E}S_n}{n}$$

we could have considered

$$\frac{S_n - n\mu}{n^\alpha}, \qquad \text{for some } \alpha > 0.$$

Then

$$\mathbb{P}\left(\left|\frac{S_n - n\mu}{n^\alpha}\right| \geq \varepsilon\right) = \mathbb{P}(|S_n - n\mu| \geq n^\alpha\varepsilon) \leq \frac{Var(S_n)}{n^{2\alpha}\varepsilon^2} = \frac{\sigma^2}{n^{2\alpha-1}\varepsilon^2} \to 0$$

as long as $2\alpha - 1 > 0 \iff \alpha > \frac{1}{2}$. This indicates that the *true order of fluctuations for* S_n *is* $\sqrt{n}$, a fact that we will revisit in the Central Limit Theorem.

24 The Central Limit Theorem

We will use the notations of the previous section.

Theorem 24.1 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n, \ldots$ be an i.i.d. sequence, $\mu = \mathbb{E}X_1$, $\sigma^2 = Var\, X_1$, $S_n = X_1 + \cdots + X_n$. Then, for any $a \in \mathbb{R}$*

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq a\right) \longrightarrow \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx. \tag{24.1}$$

In words: the sequence of variables $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges in distribution to the standard normal $Z \sim \mathcal{N}(0,1)$.

We will prove this theorem for the special case when the random variable X_1 (and hence all the X_k, $k \geq 1$) have a finite moment generating function $M(t) = M_{X_1}(t) (= M_{X_k}(t))$ for any $t \in \mathbb{R}$. (It would be enough that the variance is finite, but then the proof would require different tools.) Also, we will rely on the following Lemma, which we include without proof.

Lemma 24.2. *Let*

$$Y_1, \ldots, Y_n, \ldots$$

be a sequence of random variables with cumulative distribution functions

$$F_1(x), \ldots, F_n(x), \ldots,$$

and moment generating functions

$$M_1(t), \ldots, M_n(t), \ldots,$$

respectively. Let, furthermore, W be a random variable (to which the sequence converges) with cumulative distribution function $F_W(x)$ and moment generating function $M_W(t)$, respectively. Assume that

$$M_n(t) \longrightarrow M_W(t), \qquad \forall t \in \mathbb{R}.$$

Then

$$F_n(a) \longrightarrow F_W(a) \quad \text{whenever} \quad F_W(\cdot) \text{ is continuous at } a \in \mathbb{R}. \tag{24.2}$$

Proof. of Theorem 24.1.

First step: let us assume that $\mu = 0$ and $\sigma = 1$, in other words, that the i.i.d. random variables $X_1, \ldots, X_n, \ldots$ are standard. In this case $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ reduces to $\frac{S_n}{\sqrt{n}}$.

We will apply Lemma 24.2 with $Y_n = \frac{S_n}{\sqrt{n}}$ and $W = Z \sim \mathcal{N}(0,1)$. Since Z is a continuous random variable, (24.2) applies to any $a \in \mathbb{R}$ once we establish that

$$\left(M_{\frac{S_n}{\sqrt{n}}}(t) \to M_Z(t) = e^{t^2/2};\ \forall t \in \mathbb{R}\right) \Longleftrightarrow \left(\Psi_{\frac{S_n}{\sqrt{n}}}(t) \to \Psi_Z(t) = \frac{t^2}{2}\ \forall t \in \mathbb{R}\right), \tag{24.3}$$

as the logarithm is a continuous function (we have used (22.3)).

Now, as the variables $X_1, \dots, X_n$ are i.i.d., by (22.1) we have

$$M_{S_n}(t) = M_{X_1+\dots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \dots M_{X_n}(t) = (M(t))^n.$$

Then

$$M_{\frac{S_n}{\sqrt{n}}}(t) = \mathbb{E}\left(\exp\left(t\frac{S_n}{\sqrt{n}}\right)\right) = M_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

and thus

$$\Psi_{\frac{S_n}{\sqrt{n}}}(t) = n \cdot \Psi\left(\frac{t}{\sqrt{n}}\right).$$

Now as $\mathbb{E}(X_1) = 0$ and $Var(X_1) = 1$, we may use Taylor expansion and (22.4), to obtain, for small s:

$$\Psi(s) = 0 + 0 + \frac{s^2}{2} + O(s^3) = \frac{s^2}{2} + O(s^3)$$

and thus

$$\Psi_{\frac{S_n}{\sqrt{n}}}(t) = n \cdot \Psi\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2} + O\left(\frac{1}{\sqrt{n}}\right)$$

which readily implies (24.3).

Second step: extend to arbitrary μ and σ. Let us introduce

$$\widetilde{X}_k = \frac{X_k - \mu}{\sigma}, \quad k = 1, 2, \dots, n, \dots$$

which is an i.i.d. sequence of standard variables and

$$\widetilde{S}_n = \widetilde{X}_1 + \dots + \widetilde{X}_n = \frac{X_1 - \mu}{\sigma} + \dots + \frac{X_1 - \mu}{\sigma} = \frac{S_n - n\mu}{\sigma}$$

and thus

$$\frac{\widetilde{S}_n}{\sqrt{n}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

However, by the first step the convergence (24.1) applies to $\frac{\widetilde{S}_n}{\sqrt{n}}$, which this way extends to the general case. □

Applications of the CLT

Example 24.3 (de Moivre-Laplace CLT)**.** *Here we discuss the special case of independent trials as at the WLLN.*

$$X_i = \eta_{A_i} = \begin{cases} 1 & \text{if the } i\text{th trial is a success,} \\ 0 & \text{if the } i\text{th trial is a failure,} \end{cases}$$

Also $S_n \sim Binom(n, p)$, $\mu = \mathbb{E}X_1 (= \mathbb{E}X_i) = p$ and $\sigma = \mathbb{D}(X_1) (= \mathbb{D}X_i) = \sqrt{p(1-p)}$. Hence Theorem 24.1 reduces to the de Moivre-Laplace CLT (Theorem 15.1) in this special case.

Example 24.4. *We would like to measure some quantity (say, the distance of a star, in light years...). n measurements are made, under identical and independent conditions; the result is an i.i.d. sequence $X_1, ..., X_n$ with unknown $\mu = \mathbb{E}X_1$. The estimate for μ is $\frac{S_n}{n}$, and the question is how close this is to the true μ, and in what sense. Let us assume $\sigma = 5$.*

(a) Assume $n = 100$ measurements are made. Determine the smallest $\delta > 0$ such that μ differs from the estimation $\frac{S_n}{n}$ by not more than δ with probability 0.95? (Statistics terminology: determine the confidence interval of μ at significance level 95%.)

(b) How many measurements are needed to ensure that $\frac{S_n}{n}$ and μ differ by not more than 0.4 with probability 0.98?

For part (a):

$$0.95 \geq \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \leq \delta\right) = \mathbb{P}\left(\left|\frac{S_n - n\mu}{\sqrt{n}\sigma}\right| \leq \frac{\sqrt{n}\delta}{\sigma}\right) \approx 2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) - 1$$

that is

$$0.975 = \Phi(1.96) \geq \Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) \iff \delta \geq \frac{1.96\sigma}{\sqrt{n}} = 0.98.$$

For part (b), with a similar computation:

$$0.99 = \Phi(2.32) \geq \Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) \iff n \geq \left(\frac{2.32\sigma}{\delta}\right)^2 = \left(\frac{2.32 \cdot 5}{0.4}\right)^2 = 841.$$

Example 24.5. *Consider 48 real numbers the values of which are rounded off to the closest integer. The round off errors for these numbers are i.i.d. with uniform distribution on $[-0.5, 0.5]$. Let R and T denote the sum of the 48 real numbers, and the sum of the integers obtained by rounding off the numbers individually, respectively. What is the chance that rounding off R to the closest integer results in T?*

Let $X_1, \ldots, X_{48}$ denote the consecutive i.i.d. round off errors, and $S_{48} = X_1 + \cdots + X_{48}$. As $X_i \sim UNI[-0.5, 0, 5]$, we have $\mu = 0$ and $\sigma^2 = \frac{1}{12}$. Also, $R = T + S_{48}$. Hence

$$\mathbb{P}(|R - T| \leq 0.5) = \mathbb{P}(|S_{48}| \leq 0.5) = \mathbb{P}\left(\frac{S_{48}}{\sqrt{n}\sigma} \leq \frac{0.5}{\sqrt{48}(\sqrt{12})^{-1}}\right) \approx 2\Phi(0.25) - 1 = 0.1974$$

25 Joint distributions

25.1 Joint cumulative distribution functions

For two random variables X and Y, the joint cumulative distribution function $F : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$F(x, y) = \mathbb{P}(X \leq x \text{ and } Y \leq y), \qquad \text{for } (x, y) \in \mathbb{R}^2.$$

That is, $F(a_1, a_2)$ is the probability of the event that the random point (X, Y) lies in the quadrant with top right corner (a_1, a_2).

For $x_0 \in \mathbb{R}$ consider

$$F_X(x_0) = \mathbb{P}(X \leq x_0) = \lim_{y \to +\infty} \mathbb{P}(X \leq x_0, Y \leq y) = \lim_{y \to +\infty} F(x_0, y),$$

the *marginal distribution function* of X. Similarly for $y_0 \in \mathbb{R}$ the marginal distribution function of Y is defined as

$$F_Y(y_0) = \lim_{x \to +\infty} F(x, y_0).$$

Some further limit properties:

$$\begin{aligned} \lim_{y \to -\infty} F(x_0, y) &= 0, \quad \forall x_0 \in \mathbb{R}, \\ \lim_{x \to -\infty} F(x, y_0) &= 0, \quad \forall y_0 \in \mathbb{R}, \\ \lim_{x \to +\infty, y \to +\infty} F(x, y) &= 1. \end{aligned}$$

For a single random variable, another important property was that the cumulative distribution function is nondecreasing. This generalizes to the bivariate case as follows. Let $a_1 < b_1$ and $a_2 < b_2$ be arbitrary. Then

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0. \tag{25.1}$$

Question 25.1. *Find a geometric interpretation of the property* (25.1).

25.2 Discrete case: joint and marginal mass functions

The pair (X, Y) is jointly discrete if both X and Y are discrete random variables. This means that it is possible to enumerate the values that X and Y can take as $x_1, \ldots, x_k, \ldots$ and as $y_1, \ldots, y_\ell, \ldots$. Accordingly, the planar points the random (X, Y) may take can be enumerated as $(x_1, y_1), (x_1, y_2), \ldots, (x_k, y_\ell), \ldots$. Define the *joint mass function* of (X, Y) by

$$p(x_k, y_\ell) = \mathbb{P}(X = x_k \text{ and } Y = y_\ell).$$

Obvious properties:

$$\begin{aligned} p(x_k, y_\ell) &\geq 0, \qquad \forall (k, \ell); \\ \sum_{k,\ell} p(x_k, y_\ell) &= 1. \end{aligned}$$

It is useful to think of the joint mass function as a table the entries of which add up to 1.

The marginal mass functions of X and Y are defined as

$$\mathbb{P}(X = x_k) = \sum_{\ell} p(x_k, y_\ell); \qquad \text{and} \qquad \mathbb{P}(Y = y_\ell) = \sum_{k} p(x_k, y_\ell), \tag{25.2}$$

respectively, and can be thought of as column sums and row sums of the table, respectively.

Example 25.2. *There are* 5 *balls in a urn,* 2 *blue balls,* 2 *red balls and* 1 *white ball. Two balls are drawn (without replacement). Let* X *denote the number of white balls among the* 2 *balls drawn, and let* Y *denote the number of red balls among the* 2 *balls drawn. Determine the joint mass function and then deduce the marginal mass functions for this example.*

25.3 Jointly absolutely continuous case: joint and marginal densities

Definition 25.3. *The pair of random variables* (X, Y) *has a jointly absolutely continuous distribution if there exists some* $f : \mathbb{R}^2 \to \mathbb{R}$ *(the joint density) such that, for any (Borel measurable) set* $A \subset \mathbb{R}^2$ *we have*

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

Obvious properties:

$$\begin{aligned} f(x, y) &\geq 0, \ \forall (x, y) \in \mathbb{R}^2; \\ \iint_{\mathbb{R}^2} f(x, y)dxdy &= 1. \end{aligned}$$

The joint cumulative distribution function and the joint density function are related as

$$F(a_1, a_2) = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} f(x, y)dydx$$

and accordingly

$$f(x_0, y_0) = \frac{\partial^2 F}{\partial x \partial y}(x_0, y_0).$$

(Recall Formula (25.1) at this point.)

Marginal densities

Let (X, Y) be jointly absolutely continuous with joint density $f(x, y)$. Then, for the marginal distribution function of X:

$$F_X(x_0) = \mathbb{P}(X \leq x_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{\infty} f(t, y) dy dt = \int_{-\infty}^{x_0} f_X(t) dt,$$

where

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

In other words, the (marginal) distribution of X is absolutely continuous with *marginal density* $f_X(x)$. Similarly, the (marginal) distribution of Y is absolutely continuous with marginal density

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

Instead of summing up the masses along columns and rows as in the discrete case, $f_X(x)$ and $f_Y(y)$ are obtained by integrating the joint density along vertical lines and horizontal lines, respectively.

Remark 25.4. *So in particular we have seen that if (X, Y) are jointly absolutely continuous, then (the marginals of) both X and Y are absolutely continuous. However, the converse of this statement is, in general, false. Can you think of an example where both X and Y are absolutely continuous, but they are NOT jointly absolutely continuous?*

Example 25.5. *Let the joint density of (X, Y) be given by the formula*

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & \text{if } x > 0, y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- *Determine the marginal densities,*
- *compute the probability $\mathbb{P}(X > Y)$*

for this example.

Two dimensional uniform distribution

Definition 25.6. *Let $D \subset \mathbb{R}^2$ be a bounded domain. The distribution of (X, Y) is uniform on D if*

$$f(x, y) = \begin{cases} \frac{1}{Area(D)} & \text{if } (x, y) \in D \\ 0 & \text{otherwise.} \end{cases}$$

Example 25.7. *Let (X, Y) be uniformly distributed on the triangle with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$. Determine the joint density and the marginal densities.*

25.4 Independence

Recall the notion of independence form Definition 18.1. Note that independence is equivalent to

$$F(x,y) = F_X(x) \cdot F_Y(y), \qquad \forall (x,y) \in \mathbb{R}^2.$$

Furthermore, if (X,Y) are *discrete*, then they are independent if and only if

$$p(x_k, y_\ell) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_\ell), \qquad \forall k, \ell.$$

If (X,Y) are *jointly absolutely continuous*, then they are independent if and only if

$$f(x,y) = f_X(x) \cdot f_Y(y), \qquad \forall (x,y) \in \mathbb{R}^2.$$

Question 25.8. *Check the examples of this section for independence. Some further related problems are 6.20, 6.21 an 6.22 from the Ross book, which you may find in the file supplement_HW8A.pdf available in the resources section of piazza.*

An important special case is when $X \sim UNI[\alpha_1, \beta_1]$, $Y \sim UNI[\alpha_2, \beta_2]$ and they are independent. Equivalently, the random point (X,Y) is uniformly distributed on the rectangle $[\alpha_1, \beta_1] \times [\alpha_2, \beta_2]$.

Question 25.9. *A woman and a man decide to meet at some location. If both of them arrive, independently, at a time uniformly distributed between noon and 1pm, what is the probability that the first to arrive has to wait more than 10 minutes?*

Example 25.10 (Buffon's needle problem)**.** *A table is ruled with equidistant parallel lines that are* 2 *inches apart. A needle of length* 1 *inch is flipped onto the table. What is the probability that the needle intersects one of the lines?*

Let X denote the distance (in inches) of the midpoint of the needle to the closest line, and let Θ denote the angle that the needle makes with (one of) the lines. Then $X \sim UNI[0,1]$, while $\Theta \sim UNI[0, \frac{\pi}{2}]$ and they are independent. Equivalently, the pair (Θ, X) is uniformly distributed on the rectangle $[0, \frac{\pi}{2}] \times [0,1]$. Hence

$$\begin{aligned} \mathbb{P}(\text{intersection}) &= \mathbb{P}\left(X \le \frac{1}{2}\sin\Theta\right) = \\ &= \frac{1}{\pi/2 \cdot 1} \int_0^{\pi/2} \frac{1}{2}\sin\vartheta \, d\vartheta = \frac{1}{\pi}. \end{aligned} \tag{25.3}$$

26 Conditional distributions

Discrete case

It is pretty straightforward to define the conditional mass function of X on $\mathbb{P}(Y = y_0) > 0$. For example in this case the conditional mass function of X given $Y = 0$ is:

$Y\backslash X$	0	1
0	0.4	0.2
1	0.1	0.3

$$\begin{aligned}\mathbb{P}(X=0|Y=0) &= \frac{0.4}{0.6}=\frac{2}{3},\\ \mathbb{P}(X=1|Y=0) &= \frac{0.2}{0.6}=\frac{1}{3}.\end{aligned}$$

It is also clear that X and Y are independent if and only if the conditional mass function is the same for every $Y=y_0$ (and coincides with the marginal mass function).

Jointly absolutely continuous case

Let us fix y_0 with $f_Y(y_0)>0$ and $x_0\in\mathbb{R}$ arbitrary. Then, for dy infinitesimally small, consider

$$\begin{aligned}\mathbb{P}(X\le x_0|y_0\le Y\le y_0+dy) &= \frac{F(x_0,y_0+dy)-F(x_0,y_0)}{F_Y(y_0+dy)-F_Y(y_0)}=\\ &= \frac{\frac{F(x_0,y_0+dy)-F(x_0,y_0)}{dy}}{\frac{F_Y(y_0+dy)-F_Y(y_0)}{dy}} \xrightarrow{dy\to 0} \frac{\frac{\partial F}{\partial y}(x_0,y_0)}{F_Y'(y_0)}=\\ &\quad \frac{\int_{-\infty}^{x_0}\frac{\partial^2 F}{\partial x\partial y}(x,y_0)dx}{f_Y(y_0)}=\int_{-\infty}^{x_0} f_{X|Y}(x|Y=y_0)dx.\end{aligned}$$

This motivates the following definition.

Definition 26.1. *Let $y_0\in\mathbb{R}$ be such that $f_Y(y_0)>0$. The conditional density of X given $Y=y_0$ is defined as*

$$f_{X|Y}(x|Y=y_0)=\frac{f(x,y_0)}{f_Y(y_0)}.$$

What is remarkable is that we condition on $Y=y_0$, although this has zero probability. Here are two examples.

Example 26.2. *Let the joint distribution of (X,Y) be uniform on the triangle with vertices $(0,0)$, $(1,0)$ and $(1,1)$. Determine the conditional densities for $X|Y=y_0$ and $Y|X=x_0$ for all values y_0 and x_0 that are relevant.*

Conclusion: if the joint distribution of (X,Y) is uniform on some domain $D\subset\mathbb{R}^2$, then the conditional distributions are always uniform, on some interval(s) (which depend on the conditioning). The marginal distributions are, however, typically not uniform.

Example 26.3. *Let $X\sim UNI[0,1]$ and for any $x_0\in(0,1)$ let $Y|X=x_0\sim UNI[0,x_0]$ (shortly $Y|X\sim UNI[0,X]$). Determine the joint density, and then the marginal density of Y.*

Conditional expectation

Recall section 21 on the conditional expectation and the tower rule. More generally, we may consider the conditional expectation of any $Z = g(X, Y)$ as

$$\mathbb{E}(g(X,Y)|X = x_0) = \int_{-\infty}^{\infty} g(x_0, y) f_{Y|X}(y|X = x_0) dy$$

which can be regarded as a random variable, as it depends on $X = x_0$. The tower rule is then

$$\mathbb{E}(\mathbb{E}(g(X,Y)|X)) = \int_{-\infty}^{\infty} \mathbb{E}(g(X,Y)|X = x) f_X(x) dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy = \mathbb{E}(g(X,Y)).$$

Another useful relation is that, when $g(x, y) = h_1(x) \cdot h_2(y)$, we have

$$\mathbb{E}(h_1(X) \cdot h_2(Y)|X) = h_1(X) \cdot \mathbb{E}(h_2(Y)|X)$$

As an application, consider Example 26.3.

$$\begin{aligned}
\mathbb{E}X &= \frac{1}{2}, \\
\mathbb{E}Y &= \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}\left(\frac{X}{2}\right) = \frac{1}{2} \cdot \mathbb{E}X = \frac{1}{4}, \\
\mathbb{E}(X \cdot Y) &= \mathbb{E}(\mathbb{E}(X \cdot Y|X)) = \mathbb{E}(X \cdot \mathbb{E}(Y|X)) = \mathbb{E}\left(\frac{X^2}{2}\right) = \frac{1}{2} \cdot \mathbb{E}X^2 = \frac{1}{6}, \\
Cov(X,Y) &= \frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{24}.
\end{aligned}$$

Conditional variance

For $X = x_0$ fixed, on top of $\mathbb{E}(Y|X)$, the variance of the conditional distribution of Y may be considered. This way the conditional variance $Var(Y|X)$ is obtained, which, as it depends on $X = x_0$, is a random variable. If the expected value of this random variable is taken, then

$$\mathbb{E}(Var(Y|X)) = \mathbb{E}(\mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2) = \mathbb{E}(Y^2) - \mathbb{E}((\mathbb{E}(Y|X))^2),$$

by the tower rule. Swapping roles, we may consider the variance of the conditional expectation:

$$Var(\mathbb{E}(Y|X)) = \mathbb{E}((\mathbb{E}(Y|X))^2) - (\mathbb{E}(\mathbb{E}(Y|X)))^2 = \mathbb{E}((\mathbb{E}(Y|X))^2) - (\mathbb{E}Y)^2.$$

Adding the two relations the *conditional variance formula* is obtained:

$$Var\, Y = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X)).$$

As an application, let us revisit the following example

Example 26.4. *Let T be uniformly distributed on $[0, 10\,sec]$. A detector is functional for a random time interval which has length T. During that interval, it detects the impacts of particles, while no particles are detected before and after this time interval. The particles arrive according to a Poisson process of intensity $\lambda = 2\,sec^{-1}$. Let X denote the number of particles detected. $\mathbb{E}X = ?$ $Var\,X = ?$*

Solution: Note that if $T = t$, the conditional distribution is $X|T = t \sim Poi(\lambda t)$. Hence $\mathbb{E}(X|T = t) = \lambda t$, or briefly $\mathbb{E}(X|T) = \lambda T$. Now, by the tower rule

$$\mathbb{E}X = \mathbb{E}(\mathbb{E}(X|T)) = \mathbb{E}(\lambda T) = \lambda \mathbb{E}T = 2\,\mathrm{sec}^{-1} \cdot 5\,\mathrm{sec} = 10$$

where we have used that $T \sim UNI[0, 10\,\mathrm{sec}]$. Also $Var(X|T) = \lambda T$, so

$$\begin{aligned} \mathbb{E}(Var(X|T)) &= \mathbb{E}(\lambda T) = 10, \\ Var(\mathbb{E}(X|T)) &= Var(\lambda T) = \lambda^2\, Var(T) = 4\,\mathrm{sec}^{-2} \cdot \frac{100}{12}\,\mathrm{sec}^2 = \frac{100}{3}, \\ Var(X) &= \mathbb{E}(Var(X|T)) + Var(\mathbb{E}(X|T)) = 10 + \frac{100}{3}. \end{aligned}$$

Conditional expectation and prediction

Let us consider first just a single random variable X (with $\mu = \mathbb{E}X$ and $\sigma = \mathbb{D}X$), which we would like to estimate by a deterministic value $c \in \mathbb{R}$. Then in the sense of mean square displacement, the error of the estimation is

$$\mathbb{E}((X-c)^2) = \mathbb{E}((X-\mu+\mu-c)^2) = \mathbb{E}((X-\mu)^2)+2(\mu-c)\mathbb{E}((X-\mu))+(c-\mu)^2 = \sigma^2+(\mu-c)^2,$$

which is Steiner's theorem. In particular, the best estimation for X is $\mu = \mathbb{E}X$, for which the mean square error is σ^2.

Now given two random variables X and Y, we may be looking for some functional relation such that

$$Y = k(X) + \text{random fluctuations}$$

In other words, given that X takes a specific value, we would like to predict Y in such a way that minimizes the error arising from random fluctuations. Then, by the above argument, the best prediction is $k(X) = \mathbb{E}(Y|X)$, the mean square error of which is $Var(Y|X)$.

Example 26.5. *A random quantity $S \sim \mathcal{N}(\mu, \sigma^2)$ is sent as a signal, and received at the other end of a channel as another random variable R for which the conditional distribution is $R|S \sim \mathcal{N}(S, 1)$. If the value $R = r$ is received, what is the best estimate for what has been sent?*

By the above considerations, we want to find $\mathbb{E}(S|R = r)$. Now

$$\begin{aligned} f_{S|R}(s|R = r) &= \frac{f(s,r)}{f_R(r)} = \frac{f_{R|S}(r|S = s) \cdot f_S(s)}{f_R(r)} = \\ &= C(r) \exp\left(-\frac{1}{2}\left(\frac{(s-\mu)^2}{\sigma^2} + (r-s)^2\right)\right). \end{aligned}$$

where $C(r)$ is some normalizing factor. Then

$$\begin{aligned}\frac{(s-\mu)^2}{\sigma^2}+(r-s)^2 &= s^2\left(1+\frac{1}{\sigma^2}\right)-2s\left(r+\frac{\mu}{\sigma^2}\right)+c_1(r)= \\ &= \frac{\sigma^2+1}{\sigma^2}\left(s^2-2s\left(\frac{\sigma^2}{1+\sigma^2}r+\frac{1}{1+\sigma^2}\mu\right)+\mu_r^2\right)+c_2(r)\end{aligned} \tag{26.1}$$

where $\mu_r=\frac{\sigma^2}{1+\sigma^2}r+\frac{1}{1+\sigma^2}\mu$. Hence

$$f_{S|R}(s|R=r)=\widetilde{C}(r)\exp\left(-\frac{1}{2}\cdot\frac{(s-\mu_r)^2}{\sigma_r^2}\right),$$

that is, $S|R\sim\mathcal{N}(\mu_R,\sigma_R^2)$, with

$$\sigma_R^2=Var(S|R)=\frac{\sigma^2}{\sigma^2+1}$$

which is independent of R, and

$$\mu_R=\mathbb{E}(S|R)=\frac{\sigma^2}{1+\sigma^2}R+\frac{1}{1+\sigma^2}\mu$$

which is a convex combination of R and μ.

Correlation coefficient and indicator variables

As covariance and correlation coefficient came up, let me include one more problem on this.

Problem 26.6. *Consider an ordinary deck of 52 cards where each card has one of the 4 possible suits* $(\spadesuit,\heartsuit,\clubsuit,\diamondsuit)$ *and one of the* 13 *possible values* $(2,3,\dots,10,J,Q,K,A)$. *Add two jokers two get a deck of 54 cards. Shuffle the deck and draw a card. Put the card back and repeat this procedure* 30 *times. (So this is sampling with replacement.) Let* X *denote the number of times the card drawn is a spade, and let* Y *denote the number of times the card drawn is an ace. Determine* $Cov(X,Y)$.

27 Multidimensional transformations

Let (X,Y) have joint density $f(x,y)$, and let $(U,V)=k(X,Y)$, where $k:\mathbb{R}^2\to\mathbb{R}^2$ is

- smooth (continuously differentiable)
- one-to-one

Comment: it is enough if k has these properties when restricted to the domain $U \subset \mathbb{R}^2$ on which f is supported.

Our aim is to calculate the joint density $g(u,v)$ of (U,V). For a domain $D_1 \subset \mathbb{R}^2$, let $D_2 = k(D_1)$ (and thus $D_1 = k^{-1}(D_2)$). Then

$$\begin{aligned}\mathbb{P}((U,V) \in D_2) &= \mathbb{P}((X,Y) \in D_1) = \iint\limits_{D_1} f(x,y)dxdy = \iint\limits_{D_2} f(k^{-1}(u,v))\frac{\partial(x,y)}{\partial(u,v)}dudv = \\ &= \iint\limits_{D_2} f(k^{-1}(u,v)) \cdot \frac{1}{J}\, dudv;\end{aligned}$$

where

$$J = \frac{\partial(u,v)}{\partial(x,y)} = \left|\det\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix}\right|,$$

the *Jacobian* of $k : \mathbb{R}^2 \to \mathbb{R}^2$.

Hence

$$g(u,v) = f(k^{-1}(u,v)) \cdot \frac{1}{J}$$

for the joint density of (U,V).

It is important that on top of determining the Jacobian, one has to keep track of the two dimensional domains, as demonstrated in the following example.

Example 27.1. *Let (X,Y) be independent, $X \sim UNI[0,1]$; $Y \sim UNI[0,1]$. Let $U = X \cdot Y$ and $V = Y/X$. Determine $g(u,v)$, the joint density of (U,V).*

We have

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

We have to determine the image of the unit square by k. $y/x = v$ corresponds to a line of slope v that goes through the origin, hence the range of this variable is $0 < v < +\infty$. For any value of v fixed, we have to determine the range of u.

- if $v = 1$, we are after the intersection of the line (x,x) with the unit square, which is a segment between $(0,0)$ and $(1,1)$, hence we have $0 < u < 1$.
- if $0 < v < 1$,we are after the intersection of the line (x,vx) with the unit square, which is a segment between $(0,0)$ and $(1,v)$, hence we have $0 < u < v$.
- if $1 < v$, we are after the intersection of the line (x,vx) with the unit square, which is a segment between $(0,0)$ and $(1/v,1)$, hence we have $0 < u < 1/v$.

In a summary the image of the unit square is the domain

$$\mathbb{R}^2 \supset U = \{(u,v) \,|\, 0 < v \leq 1; 0 < u < v\} \bigcup \{(u,v) \,|1 < v < +\infty; 0 < u < 1/v\}.$$

For the Jacobian

$$J = \frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} y & -y/x^2 \\ x & 1/x \end{vmatrix} = \frac{2y}{x} = 2v.$$

Hence the joint density is

$$g(u,v) = \begin{cases} \frac{1}{2v} & \text{if } (u,v) \in U, \\ 0 & \text{otherwise.} \end{cases}$$

28 The multivariate normal distribution

Some terminology

Throughout this section, vectors in $\mathbb{R}^n$ will be underlined, while $n \times n$ matrices will be double underlined.

For the joint distribution of (arbitrary) random variables $(X_1, X_2 \ldots, X_n)$, the vector of expected values is

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{pmatrix},$$

and the covariance matrix $\underline{\underline{C}}$ is defined as

$$C_{i,j} = Cov(X_i, X_j) \qquad i = 1, \ldots, n;\ j = 1, \ldots, n.$$

In particular, for $n = 2$

$$\underline{\underline{C}} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

Where $\sigma_i^2 = Var(X_i)$ $(i = 1, 2)$, and ρ is the correlation coefficient. $\underline{\underline{C}}$ is always a symmetric positive definite matrix by the Cauchy-Schwartz inequality (20.4).

The multivariate normal distribution

Let $\underline{\mu}$ be a vector of expected values and $\underline{\underline{C}}$ a covariance matrix. Let, furthermore, $\underline{\underline{A}} = \underline{\underline{C}}^{-1}$, which is another symmetric positive definite matrix. The variables $(X_1, \ldots, X_n)$ are *jointly normally distributed* with vector of expected values $\underline{\mu}$ and covariance matrix $\underline{\underline{C}}$ (in short $\underline{X} \sim \mathcal{N}(\underline{\mu}, \underline{\underline{C}})$) if their joint density is

$$f(x_1, \ldots, x_n) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \cdot \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \underline{\underline{A}} (\underline{x} - \underline{\mu}) \right),$$

where $\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, and the superscript T stands for transpose.

Special case. If $\underline{\mu} = \underline{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ (the origin) and $\underline{\underline{C}} = \underline{\underline{A}} = \underline{\underline{I}} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$, the unit matrix, then

$$f(x_1, \dots, x_n) = \varphi(x_1) \cdots \varphi(x_n),$$

that is, the variables are independent, and $X_i \sim \mathcal{N}(0,1)$ $(i = 1, \dots, n)$.

Proposition 28.1. *If $\underline{X} \sim \mathcal{N}(\underline{\mu}, \underline{\underline{C}})$, then there exist independent variables $Y_i \sim \mathcal{N}(0,1)$ $(i = 1, \dots, n)$ and an invertible $n \times n$ matrix $\underline{\underline{B}}$ such that*

$$\underline{X} = k^{-1}(\underline{Y}) = \underline{\underline{B}}\,\underline{Y} + \underline{\mu} \quad \Longleftrightarrow \quad \underline{Y} = k(\underline{X}) = \underline{\underline{B}}^{-1}(\underline{X} - \underline{\mu}).$$

Proof. As $\underline{\underline{A}}$ is symmetric and positive definite, it is diagonalizable. That is, there exist n positive eigenvalues $\lambda_1, \dots, \lambda_n$ (counted with multiplicity) and associated eigenvectors $\underline{u}_1, \dots, \underline{u}_n$ which form an orthonormal base. Furthermore, denoting

$$\underline{\underline{P}} = \begin{pmatrix} & & \\ \underline{u}_1 & \cdots & \underline{u}_n \\ & & \end{pmatrix} \qquad \text{and} \qquad \underline{\underline{D}} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

we have

$$\underline{\underline{D}} = \underline{\underline{P}}^{-1}\,\underline{\underline{A}}\,\underline{\underline{P}} \quad \Longleftrightarrow \quad \underline{\underline{A}} = \underline{\underline{P}}\,\underline{\underline{D}}\,\underline{\underline{P}}^{-1},$$

where $\underline{\underline{D}}$ is a diagonal and $\underline{\underline{P}}$ is an orthogonal matrix.

Denote, furthermore

$$\underline{\underline{D}}^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix}$$

and define

$$\underline{\underline{B}}^{-1} := \underline{\underline{D}}^{1/2}\,\underline{\underline{P}}^{-1} \quad \Longleftrightarrow \quad \underline{\underline{B}} = \underline{\underline{P}}\,\underline{\underline{D}}^{-1/2}.$$

Then we have

$$\underline{\underline{A}} = \underline{\underline{P}}\,\underline{\underline{D}}\,\underline{\underline{P}}^{-1} = \underline{\underline{P}}\,\underline{\underline{D}}^{1/2}\underline{\underline{D}}^{1/2}\,\underline{\underline{P}}^{-1} = (\underline{\underline{B}}^{-1})^T\underline{\underline{B}}^{-1} \tag{28.1}$$

(as $\underline{\underline{P}}^{-1} = \underline{\underline{P}}^T$ for an orthogonal matrix, while $\underline{\underline{D}}^{1/2}$ is symmetric). So if

$$\underline{y} = k(\underline{x}) = \underline{\underline{B}}^{-1}(\underline{x} - \underline{\mu})$$

then as this transformation is linear, it has a constant Jacobian

$$J = \frac{\partial \underline{y}}{\partial \underline{x}} = \det(B^{-1}) = \det(\underline{\underline{D}}^{1/2}) = \sqrt{\det \underline{\underline{D}}} = \sqrt{\det \underline{\underline{A}}}, \tag{28.2}$$

where we have used that $\det P = 1$ for an orthogonal matrix. Also

$$(\underline{x} - \underline{\mu})^T\underline{\underline{A}}(\underline{x} - \underline{\mu}) = (\underline{x} - \underline{\mu})^T(\underline{\underline{B}}^{-1})^T\underline{\underline{B}}^{-1}(\underline{x} - \underline{\mu}) = \underline{y}^T\underline{y}. \tag{28.3}$$

So, the joint density of $\underline{Y} = (Y_1, \ldots, Y_n)$ is

$$\begin{aligned} g(\underline{y}) &= \frac{1}{J} \cdot f(\underline{x}) = \frac{1}{\sqrt{\det A}} \cdot \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \cdot \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{A}}(\underline{x}-\underline{\mu})\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\,\underline{y}^T\,\underline{y}\right) \end{aligned}$$

which completes the proof. □

Note also that $\underline{X} = \underline{\underline{B}}\underline{Y} + \underline{\mu}$, so

$$\begin{aligned} Cov(X_i, X_j) &= Cov\left(\sum_{k=1}^n B_{ik}Y_k, \sum_{\ell=1}^n B_{j\ell}Y_\ell\right) = \sum_{k=1}^n \sum_{\ell=1}^n B_{ik}B_{j\ell}\delta_{k\ell} = \\ &= \sum_{k=1}^n B_{ik}B_{jk} = (\underline{\underline{B}}\,\underline{\underline{B}}^T)_{ij} = (\underline{\underline{A}}^{-1})_{ij}, \end{aligned}$$

which shows that $\underline{\underline{C}} = \underline{\underline{A}}^{-1}$ is indeed the covariance matrix for $\underline{X}$.

If $\underline{X} = (X_1, \ldots, X_n)$ are jointly normally distributed, then we have the following properties.

- If the variables are uncorrelated, that is, for all pair $i \neq j$ we have $Cov(X_i, X_j) = 0$, then $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ $(i = 1, \ldots, n)$ and the variables are, actually, independent. Recall that in general zero covariance does not imply independence, so this is a special feature of the jointly normally distributed case.
- Any linear combination $\sum_{i=1}^n a_i X_i$ of the variables is normally distributed.
- The conditional distributions are normal, too. For example, if $n = 2$, then $X_1|X_2 \sim \mathcal{N}(\mu_{X_2}, \hat{\sigma}^2)$, where $\hat{\sigma}^2 = \sigma_1^2(1-\rho^2)$ and $\mu_{X_2} = \mu_1 + \rho\dfrac{\sigma_1}{\sigma_2}(X_2 - \mu_2)$. this can be verified by computations analogous to what we have done in Example 26.5.

Example 28.2. *The height-weight example, see the excel file.*

Example 28.3. *Theoretical Exercise 7.41 in the Ross book.*

29 Almost sure convergence

Definition 29.1. *A sequence of random variables $Y_n : \Omega \to \mathbb{R}$ converges in probability to 0 if for any $\varepsilon > 0$*

$$\mathbb{P}(|Y_n| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Definition 29.2. *A sequence of random variables $Y_n : \Omega \to \mathbb{R}$ converges to 0 almost surely if for any $\varepsilon > 0$*

$$\mathbb{P}(\{\omega \in \Omega \,|\, Y_n(\omega) \to 0\}) = 1.$$

Example 29.3. *Here is an example of a sequence of random variables for which $Y_n \to 0$ in probability, but not almost surely. Let $\Omega = [0,1]$ with the Lebesgue measure. For $n \geq 1$, there is a unique $k \geq 0$ and $r \in \{0, \dots, 2^k - 1\}$ such that $n = 2^k + r$. Let*

$$Y_n(\omega) = \begin{cases} 1 & \text{if } \frac{r}{2^k} \leq \omega \leq \frac{r+1}{2^k}, \\ 0 & \text{otherwise.} \end{cases}$$

Borel-Cantelli lemmas

Let $A_n \subset \Omega$, $n \geq 1$ be a sequence of events, and

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k; \qquad \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

In words: $\limsup_{n\to\infty} A_n$ is the event that infinitely many of the A_n occur, while $\liminf_{n\to\infty} A_n$ is the event that all but finitely many of the A_n occur. It follows for the de Morgan laws that

$$\left(\limsup_{n\to\infty} A_n\right)^C = \liminf_{n\to\infty} A_n^C; \qquad \left(\liminf_{n\to\infty} A_n\right)^C = \limsup_{n\to\infty} A_n^C.$$

Lemma 29.4 (Borel-Cantelli lemmas)**.** *Let A_n be a sequence of events.*

(1) If $\sum_{k=0}^{\infty} \mathbb{P}(A_k) < \infty$, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$.

(2) If the events A_n are independent and $\sum_{k=0}^{\infty} \mathbb{P}(A_k) = \infty$, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$.

Proof. To prove (1), use that for any $N \geq 1$:

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) \leq \mathbb{P}\left(\bigcup_{k=N}^{\infty} A_k\right) \leq \sum_{k=N}^{\infty} \mathbb{P}(A_k)$$

which is the tail of a convergent sequence, hence tends to 0 as $N \to \infty$.

To prove (2), we may use independence of the events A_n and the identity $1 - x \leq e^{-x}$ to obtain

$$\mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^C\right) = \prod_{k=n}^{\infty} (1 - \mathbb{P}(A_k)) \leq \prod_{k=n}^{\infty} e^{-\mathbb{P}(A_k)} \leq \exp\left(-\sum_{k=n}^{\infty} \mathbb{P}(A_k)\right).$$

Then

$$\mathbb{P}\left(\left(\limsup_{n\to\infty} A_n\right)^C\right) = \mathbb{P}\left(\liminf_{n\to\infty} A_n^C\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^C\right) \leq \lim_{n\to\infty} \exp\left(-\sum_{k=n}^{\infty} \mathbb{P}(A_k)\right) = 0.$$

as any tail of a divergent series (with non-negative terms) is infinite. □

Strong law of large numbers

Consider am i.i.d. sequence of random variables, $X_1, X_2, \ldots, X_n, \ldots$; with $\mu = \mathbb{E}X (= \mathbb{E}X_1)$ and $S_n = X_1 + \cdots + X_n$. The weak law of large numbers states that $\frac{S_n}{n} \to \mu$ in probability.

Theorem 29.5 (Strong Law of Large Numbers)**.** *Consider an i.i.d. sequence with the notations specified above. Then*

$$\frac{S_n}{n} \to \mu \qquad \textit{almost surely, as } n \to \infty.$$

Proof. We prove this statement under the additional assumption $\mathbb{E}X^4 < \infty$. Note that this implies that there exists some $\tilde{K} > 0$ such that $\mathbb{E}(|X|^p) \leq \tilde{K}$ for $p = 1, 2, 3, 4$.

Centered case. Let us assume for now that $\mu = 0$, hence our aim is to bound $\frac{S_n}{n}$ from above. We have

$$S_n^4 = (X_1 + \cdots + X_n)^4 = \sum_{i=1}^{n} X_i^4 + \sum_{i<j} \binom{4}{2} X_i^2 X_j^2 + \ldots$$

where the additional terms are of the form $cX_iX_j^2X_k$ with $j \neq i$ and $k \neq i$. When the expected value of such a term is taken, we have, by independence and $\mu = 0$ that $\mathbb{E}(cX_iX_j^2X_k) = c \cdot \mathbb{E}X_i \cdot \mathbb{E}(X_j^2X_k) = 0$. For the other two types of terms, we may use the upper bound on $\mathbb{E}(|X|^p)$ to get:

$$\mathbb{E}S_n^4 \leq Kn^2 \qquad \text{for some } K > 0.$$

Now, for any $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) = \mathbb{P}(S_n^4 \geq n^4\varepsilon^4) \leq \frac{Kn^2}{n^4\varepsilon^4} = \frac{K}{\varepsilon^4} \cdot \frac{1}{n^2}$$

by the Markov inequality. Note that the numbers on the RHS form a summable series. Introduce, for $m \geq 1$ fixed,

$$A_n^m = \left\{\left|\frac{S_n}{n}\right| \geq \frac{1}{m}\right\}; \qquad B_m = \limsup_{n\to\infty} A_n^m$$

By the first Borel-Cantelli lemma, $\mathbb{P}(B_m) = 0$ for any $m \geq 1$. That is, $\mathbb{P}(B_m^C) = 1$ for any $m \geq 1$. Hence

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} B_m^C\right) = 1$$

which precisely means that

$$\mathbb{P}\left(\left\{\forall \varepsilon > 0\ \exists n_0 \geq 1 :\ \forall n \geq n_0 \left|\frac{S_n}{n}\right| < \varepsilon\right\}\right) = 1,$$

which is exactly the statement of the SLLN.

For extension to the *general case* of $\mu \neq 0$, let $\widehat{X}_n = X_n - \mu$, which are centered, so the previous case applies. Then note that

$$\left\{ \frac{S_n}{n} \to \mu \right\} = \left\{ \frac{\widehat{S}_n}{n} \to 0 \right\}.$$

□

30 Markov chains

Given a countable set $\mathcal{S}$, a sequence of $\mathcal{S}$-valued random variables $X_0, X_1, \ldots, X_n, \ldots$ is a Markov chain if

$$\mathbb{P}(X_n = i_n \,|\, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_1 = i_1, X_0 = i_0) = \mathbb{P}(X_n = i_n \,|\, X_{n-1} = i_{n-1})$$

for any $n \geq 1$ and $i_n, i_{n-1}, \ldots, i_1, i_0 \in \mathcal{S}$. Throughout, time homogeneous Markov chains are considered which means

$$\mathbb{P}(X_n = j \,|\, X_{n-1} = i) = \mathbb{P}(X_1 = j \,|\, X_0 = i) = p_{ij}$$

where p_{ij} is the *transition matrix*. We have

- $p_{ij} \geq 0$, $\forall i, j$;
- $\sum_j p_{ij} = 1$ $\forall i$.

Consider the Gambler's ruin, Ehrenfest chain, Weather chain examples from Durrett. Here is another example:

Example 30.1. *A drunk person is walking around randomly in a small town the map of which is displayed on Figure 11. Each time he is at one of the corners, he picks evenly one of the available streets, except the street he has just arrived from. Do the corners visited by the drunk person form a Markov chain? If yes, determine the transition matrix. If no, come up with an alternative suggestion for a Markov chain that describes the process.*

Multistep transition probabilities:

$$\mathbb{P}(X_{n+m} = j | X_n = i) = (p^m)_{ij}$$

where p^m is the mth power of the matrix p.

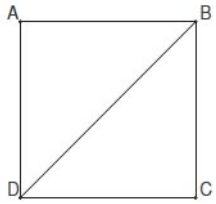

Figure 11: Map of the city for Example 30.1

Classification of states

For $x, y \in \mathcal{S}$, introduce the following notations:

$T_y = \min\{n \geq 1 : X_n = y\}$ the first hitting/return time,
$\rho_{xy} = \mathbb{P}_y(T_y < \infty) = \mathbb{P}(T_y < \infty \,|\, X_0 = y)$ the probability that x can be reached from y,
$x \to y$ if $\rho_{xy} > 0$, i.e. $\exists n \geq 1 : p_{xy}^n > 0$ x communicates with y.

Further terminology:

- If $\rho_{xx} = 1$, then x is called *recurrent*, if $\rho_{xx} < 1$ then x is called *transient*.
- A set $A \subset \mathcal{S}$ is
  - *closed* if for any $x \in A$ and $y \notin A$ we have $x \not\to y$,
  - *irreducible* if for any pair $x, y \in A$ we have $x \to y$.

Proposition 30.2. *If $A \subset \mathcal{S}$ is finite, closed and irreducible then all states $x \in A$ are recurrent.*

Proof. Introduce

$T_y^k = \min\{n > T^{k-1} : X_n = y\}$ the kth hitting/return time,
$N_y = \#\{n \geq 1 : X_n = y\}$ the total number of visits at y,
$\mathbb{E}_x(N_y)$ the expected number of visits at y, when starting from x.

Notice that

$$\{N_y \geq k\} = \{T_y^k < \infty\},$$

hence

$$\mathbb{E}_x(N_y) = \sum_{k=1}^{\infty} \mathbb{P}_x(N_y \geq k) = \sum_{k=1}^{\infty} \mathbb{P}_x(T_y^k < \infty).$$

Also

$$\mathbb{P}_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}.$$

This implies that there are two options:

$$\begin{aligned} y \text{ is transient} &\iff \rho_{yy} < 1 \iff \mathbb{E}_y(N_y) = \frac{\rho_{yy}}{1-\rho_{yy}} < \infty; \\ y \text{ is recurrent} &\iff \rho_{yy} = 1 \iff \mathbb{E}_y(N_y) = \infty. \end{aligned}$$

Using indicator variables, we get another formula

$$\mathbb{E}_x(N_y) = \sum_{k=1}^{\infty} p_{xy}^k.$$

Now, the statement of the Proposition follows from the following two observations:

1. If A is irreducible, and there exists some $y \in A$ which is recurrent, then every $x \in A$ is recurrent. This holds because

$$\mathbb{E}_x(N_x) = \sum_{k=1}^{\infty} p_{xx}^k \geq \rho_{xy} \left(\sum_{j=1}^{\infty} p_{yy}^j \right) \rho_{yx}$$

and we have $\rho_{xy} > 0$ and $\rho_{yx} > 0$ by irreducibility, while the middle factor is infinite as y is recurrent.

2. If A is *finite* and closed, than it contains at least one recurrent state. To see this, assume that all sates in A are transient. Then $\forall x, y \in A$ we have $\mathbb{E}_x(N_y) < \infty$. But

$$\sum_{y \in A} \mathbb{E}_x(N_y) = \sum_{y \in A} \sum_{k=1}^{\infty} p_{xy}^k = \sum_{k=1}^{\infty} \sum_{y \in A} p_{xy}^k = \sum_{k=1}^{\infty} 1 = \infty.$$

□

Corollary 30.3. *If $\mathcal{S}$ is finite, then it arises as a disjoint union $\mathcal{S} = \mathcal{T} \cup \mathcal{R}_1 \cup \cdots \cup \mathcal{R}_K$ for some $K \geq 1$, where every $x \in \mathcal{T}$ is transient while for any $k = 1, \ldots, K$, $\mathcal{R}_k$ is a closed and irreducible (hence recurrent) class.*

Proof. Let $x \in \mathcal{T}$ if there exists some $y \in \mathcal{S}$ such that $x \to y$ but $y \not\to x$. Such states are transient as there is a positive probability that we get to y and then never get back to x.

For $y_1, y_2 \in (\mathcal{S} \setminus \mathcal{T})$, we have $y_1 \to y_2$ if and only if $y_2 \to y_1$, which defines an equivalence relation, and the $\mathcal{R}_k$ are the equivalence classes. □

In most cases, irreducible Markov chains are considered, which means that $\mathcal{S}$ consists of a single closed irreducible class. In case $\mathcal{S}$ is finite, (all states in) this class are recurrent.

Stationary distributions

Even if started from a specific state $X_0 = i \in \mathcal{S}$, at the next time step, X_1 is already random, in particular $\mathbb{P}(X_1 = j | X_0 = i) = P_{ij}$. Hence it is reasonable to consider the evolution of distributions. If we have

$$\alpha_j = \mathbb{P}(X_0 = j), \qquad \left(\alpha_j \geq 0, \forall j \in \mathcal{S}, \sum_{j \in \mathcal{S}} \alpha_j = 1 \right)$$

then

$$\beta_j = \mathbb{P}(X_1 = j) = \sum_{i \in \mathcal{S}} \mathbb{P}(X_1 = j | X_0 = i) \mathbb{P}(X_0 = i) = \sum_{i=1}^{K} \alpha_i P_{ij}$$

where we have assumed that $\#\mathcal{S} = K < \infty$. The moral is that *probability distributions on $\mathcal{S}$ are evolved by multiplication by the transition matrix from the right.*

Definition 30.4. *A vector $(\pi_1, \ldots, \pi_K)$ with $\pi_i \geq 0$ and $\sum_{i=1}^{K} \pi_i = 1$ is a stationary distribution if*

$$\pi_j = \sum_{i=1}^{K} \pi_i P_{ij}, \qquad \forall j = 1, \ldots K$$

that is, if π_i is a left eigenvector of the matrix P_{ij} corresponding to the eigenvalue $\lambda = 1$.

Proposition 30.5. *If $\mathcal{S}$ is finite and the chain is irreducible, then there exists a unique stationary distribution. Moreover, $\pi_i >$ for every $i = 1, \ldots, K$.*

Proof. Recall

$$\sum_{j=1}^{K} P_{ij} = 1; \qquad \forall i = 1, \ldots, K$$

which can be interpreted as follows: $\lambda = 1$ is an eigenvalue of P_{ij} with right eigenvector $(1, \ldots, 1)^T$. Hence $\lambda = 1$ is an element of the spectrum of P, and thus there has to exist at least one left eigenvector $(v_1, \ldots, v_K)$; for which $v_j = \sum_{i=1}^{K} v_i P_{ij}$ for every $j = 1, \ldots K$.

Now we show that $v_j \geq 0$ for every $j = 1, \ldots K$ – more precisely, that all of the components v_j have the same sign. Consider $R = (Id + P)/2$, that is, $R_{ij} = (\delta_{ij} + P_{ij})/2$, the transition matrix for the lazy chain. On the one hand, R apparently has the same eigenvalues and eigenvectors as P. On the other hand, there exists $M \geq 1$ such that $R_{ij}^M > 0$ for every $(i, j) \in \{1, \ldots, K\}^2$. To see this, note that, by irreducibility, for every pair i, j there exist $m (= m(i, j))$ such that $R_{ij}^m > 0$. Let

$$M = \max\{m(i, j) \, : \, (i, j) \in \{1, \ldots, K\}^2\}$$

and we claim that $R_{ij}^M > 0$ for every (i,j), because if you make it to j earlier, you can just "waste your time" there. Now we argue by contradiction, assuming there exist components of negative sign, which would imply:

$$|v_j| = \left| \sum_{i=1}^K v_i R_{ij}^M \right| < \sum_{i=1}^K |v_i| R_{ij}^M \implies$$
$$\implies \sum_{j=1}^K |v_j| < \sum_{j=1}^K \sum_{i=1}^K |v_i| R_{ij}^M = \sum_{i=1}^K |v_i|,$$

a contradiction. To see that all components are, actually, positive, note that

$$v_j = \sum_{i=1}^K v_i R_{ij}^M > 0$$

as on the RHS at least one term is positive, unless $v_i = 0$ for all i, a contradiction.

Finally, to see that the eigenvector v obtained this way is unique (up to a constant factor), assume that the eigenvalue $\lambda = 1$ is of higher multiplicity, then the associated eigenspace would contain a two dimensional plane, and thus a vector perpendicular to v, which would have components of variable signs, a contradiction. □

Convergence

Aperiodicity. For $x \in \mathcal{S}$, let

$$Per(i) = g.c.d\{n \geq 1 \,:\, p_{ii}^n > 0\}$$

where g.c.d. stands for greatest common divisor. If $Per(i) = 1$, then $i \in \mathcal{S}$ is called aperiodic. Equivalently, $i \in \mathcal{S}$ is aperiodic if there exists $N \geq 1$ such that $p_{ii}^n > 0$ for any $n \geq N$. Also, if $i \in \mathcal{S}$ is aperiodic and $i \leftrightarrow j$, then $j \in \mathcal{S}$ is aperiodic, hence aperiodicity is a property of a class. Unless otherwise stated, irreducible, aperiodic Markov chains are considered.

Proposition 30.6. *Let us consider an irreducible, aperiodic Markov chain on a finite state space $\mathcal{S}$. Then, for any pair i, j we have*

$$\lim_{n\to\infty} p_{ij}^n = \pi_j.$$

Also, $\mathbb{E}_x(T_x) = (\pi(x))^{-1}$ for any $x \in \mathcal{S}$.

Proof. (Sketch)
Coupling argument. We may consider the Markov chain on $\mathcal{S} \times \mathcal{S}$ with transition probabilities

$$P_{(x_1,y_1)(x_2,y_2)} = p_{x_1x_2} \cdot p_{y_1y_2}.$$

This chain is irreducible. As a consequence, if X_n and Y_n are two copies of the Markov chain started from two different states, and T denotes the first time that $X_n = Y_n$, then $\mathbb{P}(T < \infty) = 0$. This implies

$$|\mathbb{P}(X_n = j) - \mathbb{P}(Y_n = j)| \leq 2P(T > n) \to 0 \qquad \forall j \in \mathcal{S}, \text{ as } n \to \infty.$$

Now if X_n is started form state i and Y_n is started form the stationary distribution, this boils down to

$$\lim_{n\to\infty} |p_{ij}^n - \pi_j| = 0.$$

To see that $\mathbb{E}_x(T_x) = (\pi(x))^{-1}$, introduce $N_n(x)$, the number of visits at x up to time n. Recall that T_x^k is the time of the kth visit at y. Then

$$\lim_{n\to\infty} \frac{n}{N_n(x)} = \lim_{k\to\infty} \frac{T_x^k}{k} = \mathbb{E}_x(T_x)$$

almost surely. The first equality is because for both fractions the numerator counts the total number of steps, while the denominator counts the number of times the chain is at state x. The second inequality is a law of large numbers when the chain is started form x, and otherwise the initial transient of making to x for the first time does not contribute to the asymptotic. Hence

$$\lim_{n\to\infty} \frac{N_n(x)}{n} = (\mathbb{E}_x(T_x))^{-1}$$

almost surely, also when the chain is started from the stationary state, in which case $\mathbb{E}_\pi N_n(x) = \pi(x)n$. □

Detailed balance

Let us consider irreducible chains with finitely many states ($\#\mathcal{S} = K < \infty$). We know that there is a unique stationary state π_i, namely, the eigenvector $\sum_{i=1}^{K} \pi_i p_{ij} = \pi_j$, however, it may be difficult to solve this explicitly if K is large.

Definition 30.7. *The Markov chain has detailed balance if there exists some vector $\pi_i > 0$ such that*

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \forall i, j \in \{1, \ldots, K\}^2$$

Summation on i implies that in this case π_i is the stationary distribution.
Comments.

- There are chains which do not have detailed balance. See Example 1.29 in Durrett.
- Two important examples for detailed balance are
  - Birth and death chains, in particular, the Ehrenfest chain.
  - Random walks on *undirected* graphs.

Gambler's ruin

For $N \geq 2$ and $p \in (0,1)$, let $\mathcal{S} = \{0, 1, \ldots, N-1, N\}$, and

$$\begin{aligned} &p(i, i+1) = p \quad \text{and} \\ &p(i, i-1) = q = 1-p \quad \text{for } 1 \leq i \leq N-1, \text{ while} \\ &p(0,0) = 1, \\ &p(N,N) = 1. \end{aligned}$$

0 and N are absorbing states. Introduce $\theta = q/p$. Let

$$\begin{aligned} &T_N = \min\{n \geq 1 \,:\, X_n = N\}, \\ &T_0 = \min\{n \geq 1 \,:\, X_n = 0\}, \\ &T = \min(T_0, T_N), \\ &h(k) = \mathbb{P}(T_N < T_0 \,|\, X_0 = k), \qquad k = 1, \ldots, N-1; \\ &g(k) = \mathbb{E}_k T, \qquad k = 1, \ldots, N-1. \end{aligned}$$

Unfair game from the gambler's perspective: Let $p < 1/2$; and thus $\theta > 1$. $X_n = 0$ means that the gambler is bankrupt at time n, while $X_n = N$ means that the gambler has made his goal and stops playing at time N. Let us determine $h(k)$:

$$\begin{aligned} &h(0) = 0, \quad h(N) = 1; \\ &h(k) = qh(k-1) + ph(k+1) \implies \\ &(h(k+1) - h(k)) = \theta(h(k) - h(k-1)) = \cdots = \theta^k (h(1) - h(0)) \end{aligned}$$

and thus

$$1 = h(N) - h(0) = \sum_{k=0}^{N-1} (h(k+1) - h(k)) = (h(1) - h(0)) \sum_{k=0}^{N-1} \theta^k = (h(1) - h(0)) \frac{\theta^N - 1}{\theta - 1}$$

so

$$\begin{aligned} &(h(1) - h(0)) = \frac{\theta - 1}{\theta^N - 1} \implies \\ &h(k) = h(k) - h(0) = \sum_{i=0}^{k-1} (h(i+1) - h(i)) = \sum_{i=0}^{k-1} \theta^i (h(1) - h(0)) = \frac{\theta - 1}{\theta^N - 1} \cdot \frac{\theta^k - 1}{\theta - 1} = \frac{\theta^k - 1}{\theta^N - 1}. \end{aligned}$$

Let us consider roulette when $p = 18/38$, $q = 20/38$ and $\theta = 10/9$. If, for example, $k = 50$ and $N = 100$, then

$$h(50) = \frac{(10/9)^{50} - 1}{(10/9)^{100} - 1} \approx 0.00513$$

Fair game: The case of $p = q = 1/2$ ($\theta = 1$) was considered back in Budapest:

$$h(k) = \frac{k}{N}. \tag{30.1}$$

Unfair game from the casino's perspective: Now $p > 1/2$ and thus $\theta < 1$. We may consider the limit

$$\lim_{N\to\infty} \mathbb{P}_k(T_0 < \infty) = 1 - \lim_{N\to\infty} h(k) = \theta^k \tag{30.2}$$

In particular, for roulette, the chance that the casino will ever go bankrupt decreases exponentially as a function of the initial capital k. For example, if $k = 100$, $\mathbb{P}(T_0 < \infty) = (0.9)^{100} \approx 2.656 \times 10^{-5}$.

Expected exit time for the fair game: For $p = 1/2$ let us determine $g(k)$.

$$\begin{aligned}
&g(0) = 0, \quad g(N) = 0;\\
&g(k) = 1 + \frac{1}{2}g(k-1) + \frac{1}{2}g(k+1) \implies\\
&g(k+1) - g(k) = g(k) - g(k-1) - 2 = \cdots = g(1) - g(0) - 2k = c - 2k
\end{aligned}$$

and thus

$$0 = g(N) = g(N) - g(0) = \sum_{k=0}^{N-1}(g(k+1) - g(k)) = Nc - 2\frac{N(N-1)}{2}.$$

This implies

$$c = N - 1$$

and

$$g(k) = g(k) - g(0) = \sum_{i=0}^{k-1}(g(i+1) - g(i)) = kc - 2\frac{k(k-1)}{2} = k(N-k). \tag{30.3}$$

Infinite state spaces

Reflected random walk. Let

$$\begin{aligned}
&\mathcal{S} = \{0, 1, \dots\},\\
&p(i, i+1) = p \quad \text{for } i \geq 0,\\
&p(i, i-1) = q = 1 - p \quad \text{for } i \geq 1, \text{ and}\\
&p(0, 0) = q = 1 - p.
\end{aligned}$$

Let us use the notations from the gambler's ruin section, in particular $\theta = q/p$.

Drift to the left: $p < 1/2$, and thus $\theta > 1$. This is a birth and death chain, which has detailed balance:

$$\pi(i)p = \pi(i+1)q \implies \pi(i+1) = \theta^{-1}\pi(i) = \theta^{-i}\pi(0)$$

which by $\sum_{i=0}^{\infty} \pi(i) = 1$ gives

$$\pi(k) = (1-\theta^{-1})\theta^{-k+1}$$

As there is a stationary distribution, the expected return time to any state has the finite expectation $(\pi(i))^{-1}$.

Drift to the right: $p > 1/2$, and thus $\theta < 1$. From (30.2) it follows that the chain is *transient*: when started from 0, there is a positive chance that it never returns. With probability 1 it returns only finitely many times, and wanders off to infinity.

Case with no drift: To see what happens in this borderline case, we consider the $N \to \infty$ limits of

- (30.1). When started form 0, the chain returns with probability 1.
- (30.3). However, the expected return time is infinite.

Also, the detailed balance equations would result in $\pi(i) = \pi(0)$ for every $i \geq 1$, but as there are infinitely many states, no such probability distribution exists.

These examples demonstrate that, for irreducible Markov chains on infinite state spaces there are the following three options.

- *Positive recurrence.* There is a stationary distribution, the chain returns to its starting position with probability 1, and the expected return time is finite.
- *Transience.* There is a positive probability that the chain does not return to its starting position. With probability 1 it returns only finitely many times and wanders off to infinity.
- *Nullrecurrence.* The chain returns to its starting position with probability 1, however, the expected return time is infinite. Accordingly, there is no stationary distribution.

Symmetric random walk on $\mathbb{Z}^d$. By symmetry, there cannot exit a stationary distribution, so the chain is either transient or nullrecurrent. It is a famous theorem of György Pólya that the simple symmetric random walk on $\mathbb{Z}^d$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$.

We will investigate first the *case of* $d = 1$. Let us start the chain from 0. Then, by direct inspection

$$\mathbb{P}(X_k = 0) = \begin{cases} \binom{2n}{n} \cdot \frac{1}{2^{2n}} & \text{if } k = 2n, \\ 0 & \text{if } k = 2n+1. \end{cases}$$

Let

$$A_n = \{X_{2n} = 0\},$$

and let us use Stirling's formula to estimate $\mathbb{P}(A_n)$.

$$\begin{aligned}
n! &\approx \frac{n^n}{e^n}\sqrt{2\pi n}, \\
(n!)^2 &\approx \frac{n^{2n}}{e^{2n}} 2\pi n, \\
(2n!) &\approx \frac{(2n)^{2n}}{e^{2n}}\sqrt{4\pi n} = 2^{2n} \cdot \frac{n^{2n}}{e^{2n}} 2\pi n \cdot \frac{c}{\sqrt{n}}, \quad \text{for some } c > 0, \\
\mathbb{P}(A_n) &= \binom{2n}{n} \cdot \frac{1}{2^{2n}} = \frac{c}{\sqrt{n}} + \text{higher order terms.}
\end{aligned}$$

Now as

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$$

(2) of the Borel-Cantelli lemmas suggests that

$$\mathbb{P}(\limsup A_n) = 1 \implies$$

The walk returns to 0 infinitely many times with probability 1.

For $d \geq 2$, consider d independent random walkers $X_n^{(i)}$, $i = 1, \ldots, d$, along the d coordinate axes, all isomorphic to $\mathbb{Z}$. The d dimensional random walk returns to the origin at time $2n$ if and only if all the $X_{2n}^{(i)}$ return to 0 simultaneously. So

$$\begin{aligned}
A_n^d &= \{d \text{ dimensional random walk returns to the origin at time } 2n\}; \\
\mathbb{P}(A_n^d) &= \prod_{i=1}^{d} \mathbb{P}(X_{2n}^{(i)} = 0) = (\mathbb{P}(A_n))^d = \frac{C}{n^{d/2}} + \text{higher order terms}
\end{aligned}$$

So

$$\begin{array}{lcll}
\displaystyle\sum_{n=1}^{\infty} \mathbb{P}(A_n^d) = \infty & \iff & \text{nullrecurrence} & \iff \quad d = 1 \text{ or } 2; \\
\displaystyle\sum_{n=1}^{\infty} \mathbb{P}(A_n^d) < \infty & \iff & \text{transience} & \iff \quad d \geq 3.
\end{array}$$