Cairo University

Faculty of Sciences

Department of Mathematics

An Introduction to Probability Theory

Stat 201

Based on the book of Y. A. Rozanov

Lectures notes edited by Mostafa SABRI

Disclaimer

These lectures notes are heavily based on the book of Rozanov, "Probability Theory : A Concise Course", a book which, as far as I can see, has received universal acclaim from laymen and students alike around the world. The choice of content is exactly what I was looking for, it covers all the basics and all fundamental theorems which a working mathematician is assumed to know no matter his research field. The level of rigor is good, the language makes no use of measure theory (which is great if you are teaching 2nd year undergraduates), and the book doesn't get lost into too many auxiliary directions, which are simply distracting if the course length is on average 11 lectures of 2 hours, due to midterm exams and so on. The exercises too, which are due to the translator Silverman, are very good and adequate.

So why write any notes then ? Well my contribution has simply been to make the material a bit more organized for students. The inline text has been converted into definitions, lemmas and so on. I have added some examples here and there, I removed few results in later chapters, I gave alternative proofs on some occasions, I have sometimes tried to emphasize the importance of certain notions like the axioms of probability. As a rule I did not change what was already good. Hence some chapters are almost identical to the book, others have more pronounced changes, the exercises are all copied from the book. Then I also changed some notations to make them more standard. To summarize, my contribution is small especially to mature mathematicians, but I found it important for students. Finally I added an epilogue.

The notes have been classroom tested for at least two years, I have been able to finish the contents in 10 lectures (sometimes excluding the optional epilogue). This leaves time for occasional holidays, Q&A and so on.

The notes are "to the point" so to speak for educational purposes, it is a good idea to have a parallel, optional, leisurely reading for those who want more. For that I recommend the book of Isaac "The pleasures of probability" for a fun ride, or the textbook of Sheldon Ross for more examples.

Mostafa Sabri

Contents

1. The uniform distribution is discussed in Chapter 4, more distributions like the Cauchy/exponential/geometric distributions are discussed in the exercises.

Chapter 1

Basic concepts

Classical physics teaches us that the universe is completely deterministic. For example, if we toss a fair coin, then in principle by taking into account the force and angle at which the coin was tossed, the air resistance, the face of the coin before the toss, the distance to the ground, then we could in principle solve some differential equations and predict exactly if the outcome will be a head or a tail. But in practice no one does that, this is way too complicated. We prefer to look at this coin toss as a *random experiment*. Intuitively, since the coin is fair, the events of getting a head or a tail should be equally likely, in other words, each should have a 50% chance of appearing, or equivalently, probability 0.5. With this point of view we are abandoning the task of finding the exact outcome of the experiment and just ask for the most likely outcomes that will appear. In this sense, randomness is not an intrinsic property of the experiment, this theory of randomness which we call *probability theory* is simply a way to cope with our limited human capacity to solve equations and hence ask for less precise information which can still be very useful. This point of view is adopted all the time in physics to model complex behaviors; heavy nuclei for example are observed to behave like random matrices.

This was all about classical physics. In quantum physics, randomness is in contrast at the very core of the theory. So in all cases, it is important to have a good theory of probability to understand the world in which we live. Let us start this chapter by formalizing some intuition we have about daily experiments. Below we say that two outcomes are *mutually exclusive* if they cannot both occur at the same time (e.g. head and tail).

Definition 1.1. Consider an experiment with a finite number N of mutually exclusive outcomes, and suppose they are all *equally likely*. Let A be some event

associated to this experiment. Then we define the *probability of A* by

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

where $N(A)$ is the number of outcomes leading to the occurrence of A.

We will usually denote by Ω the set of all outcomes of the experiment.

Example 1.2. Toss a fair coin. What is the probability of getting a head ?

Solution. Here $\Omega = \{H, T\}$ and $A = \{H\}$ so $\mathbb{P}(A) = \frac{1}{2} = 0.5$.

Example 1.3. Throw an unbiased dice. What is the probability of getting an even number ?

Solution. Here $\Omega = \{1, 2, \ldots, 6\}$ and $A = \{2, 4, 6\}$ so $\mathbb{P}(A) = \frac{3}{6} = 0.5$.

Example 1.4. Throw a pair of unbiased dices. What is the probability that both dice show the same number of spots ?

Solution. Here $\Omega = \{(1,1), (1,2), \ldots, (6,6)\} = \{(x,y) : x, y \in \{1, 2, \ldots, 6\}\}$ is the set of all ordered pairs with coordinates in $\{1, \ldots, 6\}$, where the x coordinate represents the first dice and the y coordinate the second dice. There are $6 \times 6 = 36$ such pairs. Also $A = \{(1,1), (2,2), \ldots, (6,6)\}$. Hence, $\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}$.

The rule $\mathbb{P}(A) = \frac{N(A)}{N}$ is of *geometric nature*. It says that the probability of getting A is the *relative size* of A in the set Ω. But why not estimate the probability of A using a *time average* instead ? More precisely :

Definition 1.5. Consider an experiment which is repeated n times under exactly the same conditions, with each trial having no influence on the others. Let $n(A)$ be the total number of times that some event A occurred during these experiments. Then we define the *relative frequency of A* by $\frac{n(A)}{n}$.

Example 1.6. In tossing a fair coin 10,000 times in 10 series of 1000 trials each, it was experimentally observed that the number of heads in each 1000 trials was

$$\{501, 485, 509, 536, 485, 488, 500, 497, 494, 484\}.$$

Hence, the relative frequencies of heads were $0.501, 0.485, \ldots, 0.484$, respectively.

A natural question then is whether we have

$$\mathbb{P}(A) = \lim_{n \to \infty} \frac{n(A)}{n}.$$

This is indeed the case and it is a consequence of the *Law of Large Numbers* which we will learn in Chapter 6.

Example 1.7 (De Méré's paradox)**.** The french Chevalier de Méré noticed after some extensive gaming that upon throwing 3 dices, the total 11 appears more often than 12. But from his point of view these two events (call them A_1 and A_2) should be equally likely because A_1 occurs in six ways : (6:4:1, 6:3:2, 5:5:1, 5:4:2, 5:3:3, 4:4:3) and A_2 also occurs in six ways : (6:5:1, 6:4:2, 6:3:3, 5:5:2, 5:4:3, 4:4:4). Therefore, A_1 and A_2 should have the same probability.

The fallacy in this argument was found by the french polymath Pascal. The problem here is that the outcomes listed by De Méré are not equally likely. For example, the combination 6:4:1 may occur in six different ways : (6,4,1), (6,1,4), (4,6,1), (4,1,6), (1,6,4), (1,4,6). Here we are considering ordered triples with the first coordinate representing the first dice and so on. On the other hand, the combination 4:4:4 can occur in only one way, namely (4,4,4).

Taking this into account, we now check that $N(A_1) = 27$ while $N(A_2) = 25$. So they are not equally probable, which is what De Méré observed experimentally.

To treat more complicated examples we need some rudiments of combinatorial analysis. We start with the following.

Theorem 1.8. (1) *Given n_1 elements $a_1, a_2, \ldots, a_{n_1}$ and n_2 elements $b_1, b_2, \ldots, b_{n_2}$ there are exactly $n_1 n_2$ ordered pairs of the form (a_i, b_j).*

(2) *More generally, if for each $k = 1, \ldots, r$ we have n_k elements of the form $a_1^{(k)}, \ldots, a_{n_k}^{(k)}$, then there are exactly $n_1 n_2 \cdots n_r$ ordered r-tuples of the form $(a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_r}^{(r)})$.*

For example, if we throw 2 dices and one coin, then the number of ordered triples is $6 \times 6 \times 2 = 72$. One of them is e.g. $(1, 5, H)$.

Proof. (1) should be known to the student. One proof is to draw the ordered pairs on the (x, y)-plane and count the number of points. Another proof is to fix the first coordinate and count. In fact, if we fix i, then there are n_2 pairs of the form $(a_i, b_1), \ldots (a_i, b_{n_2})$. Adding this from $i = 1$ to n_1 gives $n_2 + n_2 + \cdots + n_2 = n_1 n_2$ pairs.

For (2) we use mathematical induction. The result is true for $r = 2$ by (1). Assume the result is true for $r - 1$. Then the number of r-tuples $(a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_r}^{(r)})$ is equal to the number of ordered pairs $(a_{i_1}^{(1)}, b_j)$, where b_j are $(r-1)$-tuples of the form $(a_{i_2}^{(2)}, \ldots, a_{i_r}^{(r)})$. By the induction hypothesis the number of b_j is $n_2 \cdots n_r$. And by hypothesis, the number of $a_{i_1}^{(1)}$ is n_1. So applying (1) we get that the number of pairs $(a_{i_1}^{(1)}, b_j)$ is $n_1(n_2 \cdots n_r)$. This proves that the number of r-tuples $(a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_r}^{(r)})$ is $n_1 \cdots n_r$. □

Example 1.9. What is the probability of getting three sixes in a throw of three dices ?

Solution. Here Ω is the set of all ordered triples (x, y, z) with $x, y, z \in \{1, \ldots, 6\}$. By Theorem 1.8 there are $N = 6^3$ such triples. Also $A = \{(6, 6, 6)\}$. Hence $\mathbb{P}(A) = \frac{1}{6^3} = \frac{1}{216}$.

Sampling with replacement : Suppose we choose r objects in succession from a population of n distinct objects $\{a_1, \ldots, a_n\}$, by first choosing the object, recording the result, then returning the object back to the population, then choosing the next object and so on. How many ordered samples do we get ?

Solution. The samples take the form of ordered tuples $(a_{i_1}, \ldots, a_{i_r})$ where a_{i_k} represents the k-th outcome. Each coordinate has exactly n possible outcomes, since the chosen objects are returned. Hence, by Theorem 1.8 the total number is

$$N = n^r \,.$$

Sampling without replacement : Suppose we choose r objects in succession from a population of n distinct objects $\{a_1, \ldots, a_n\}$, by first choosing the object, recording the result, then removing the object from the population before making the next choice. How many ordered samples do we get ?

Solution. The samples are again ordered tuples $(a_{i_1}, \ldots, a_{i_r})$, but now a_{i_1} has n possible outcomes, while a_{i_2} has $(n-1)$ possible outcomes, a_{i_3} has $(n-2)$ possible outcomes and so on. So by Theorem 1.8, the number of samples is

$$N = n(n-1)(n-2)\cdots(n-r+1)\,. \tag{1.1}$$

Number of permutations : Suppose we have n distinct items $a_1, \ldots, a_n$. What is the number of ways in which we can rearrange this sequence ?

Solution. Each rearrangement is an ordered n-tuple $(a_{i_1}, \ldots, a_{i_n})$. We start by choosing one of the n items and putting it first. Once the first position is fixed, the second position chooses among $n-1$ items, the third position chooses among $n-2$ items and so on. So the number of such permutations follows by taking $r = n$ in (1.1), namely

$$N = n(n-1)(n-2)\cdots 1 = n!$$

Example 1.10. Suppose a prison has n cells and $r \leq n$ criminals. We want to place the criminals in solitary confinement and record which criminal is assigned to which cell. How many settings are possible ?

Solution. This is just a sampling without replacement. Namely we consider the prisoners in succession, then the first prisoner can choose among n cells, the second among $(n-1)$ cells, and so on. Hence, $N = n(n-1)\cdots(n-r+1)$.

Example 1.11. A subway train made up of n cars is boarded by $r \leq n$ passengers, each entering a car completely at random. What is the probability that they all end up in different cars ?

Solution. Let A be the event "all passengers end up in different cars". Then $\mathbb{P}(A) = \frac{N(A)}{N}$. To find N, note that the first passenger chooses among n cars, the second also chooses among n cars and so on. So the total number of outcomes is $N = n^r$. Next, for A to happen, the first passenger chooses among n cars, the second among $n-1$ cars and so on. So $N(A) = n(n-1)\cdots(n-r+1)$. Thus, $\mathbb{P}(A) = \frac{n(n-1)\cdots(n-r+1)}{n^r}$.

Definition 1.12. Any set of r elements chosen from a population of n elements *without regard to order* is called a *subpopulation of size r* of the original population.

Theorem 1.13. *A population of n elements has exactly*

$$C_r^n = \frac{n!}{r!\,(n-r)!}$$

subpopulations of size $r \leq n$.

Equivalently, the number of ways of selecting r items from a batch of n items is C_r^n if we disregard the order among the r items.

Proof. Let X be the number of ways of selecting r ordered elements from the population and Y the number of ways of choosing a subpopulation of size r, i.e. while disregarding the order among the r elements. Finally let Z be the number of ways to arrange the r elements. Then $X = YZ$. But $Z = r!$ Also, selecting the ordered sample in succession, we have $X = n(n-1)\cdots(n-r+1)$. Hence, $Y = \frac{X}{Z} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!\,(n-r)!}$. □

Remark 1.14. The number C_r^n is called a *binomial coefficient*. It is also denoted by $\binom{n}{r}$.

Example 1.15. What is the number of ways of choosing r cells in a prison having n cells ?

Solution. This is C_r^n by Theorem 1.13. Compare this with Example 1.10. The difference here is that we don't care about ordering the cells, the question just

asks how to pick r cells. In Example 1.10 however, not only do we pick r cells for the criminals, we also order the cells according to the occupying prisoners. This is why the results are different.

Theorem 1.16. *Given a population of n elements, let $n_1, n_2, \ldots, n_k$ be positive integers such that $n_1 + \cdots + n_k = n$. Then there are precisely*

$$N = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

ways of partitioning the population into k subpopulations of sizes $n_1, n_2, \ldots, n_k$ respectively.

Note that if $k = 2$ this reduces to Theorem 1.13 with $n_1 = r$ and $n_2 = n - r$.

Proof. Step 1 : choose a subpopulation of size n_1. This can be done in $C_{n_1}^n$ ways. Step 2 : choose a subpopulation of size n_2 from the remaining $n - n_1$ elements. This can be done in $C_{n_2}^{n-n_1}$ ways. And so on. At Step $k-1$, we choose a subpopulation of size n_{k-1} from the remaining $n - n_1 - \cdots - n_{k-2} = n_{k-1} + n_k$ elements. This can be done in $C_{n_{k-1}}^{n_{k-1}+n_k}$ ways. After that we are left with exactly n_k elements, which form the last subpopulation.

The total number of ways to partition the population is thus

$$\begin{aligned} N &= C_{n_1}^n C_{n_2}^{n-n_1} \cdots C_{n_{k-2}}^{n-n_1-\cdots-n_{k-3}} C_{n_{k-1}}^{n_{k-1}+n_k} \\ &= \frac{n!}{n_1!\,(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)} \cdots \frac{(n-n_1-\cdots-n_{k-3})!}{n_{k-2}!\,(n_{k-1}+n_k)!} \cdot \frac{(n_{k-1}+n_k)!}{n_{k-1}!\,n_k!} \\ &= \frac{n!}{n_1!\, n_2! \cdots n_k!}. \end{aligned}$$

□

Example 1.17. A batch of 100 manufactured items is checked by an inspector, who examines 10 items selected at random. If none of them is defective, he accepts the whole batch. Otherwise further inspection is performed. What is the probability that a batch containing 10 defective items will be accepted ?

Solution. Let A be the event "the batch is accepted", so $\mathbb{P}(A) = \frac{N(A)}{N}$. To find N note that there are C_{10}^{100} ways to choose 10 items from the batch, hence $N = C_{10}^{100}$. For A to occur, the inspector must choose among the 90 non-defective items. This is done in C_{10}^{90} ways. Thus,

$$\mathbb{P}(A) = \frac{C_{10}^{90}}{C_{10}^{100}} = \frac{90!}{10!\,80!} \cdot \frac{10!\,90!}{100!} = \frac{81 \cdot 82 \cdots 90}{91 \cdot 92 \cdots 100} \approx \left(1 - \frac{1}{10}\right)^{10} \approx \frac{1}{\mathrm{e}},$$

where $\mathrm{e} = 2.718\ldots$.

Example 1.18. What is the probability that two playing cards picked at random from a full deck are both aces ?

Note : in this course we always consider decks of 52 cards (no jokers).

Solution. Let A be the event "the two chosen cards are both aces". Then $\mathbb{P}(A) = \frac{N(A)}{N}$. There are $N = C_2^{52}$ ways to pick 2 cards from a deck of 52 cards. To achieve A, the cards must be chosen among the 4 aces, there are C_2^4 ways to make such a choice. Hence, $\mathbb{P}(A) = \frac{C_2^4}{C_2^{52}} = \frac{4!}{2!\,2!} \cdot \frac{2!\,50!}{52} = 12 \cdot \frac{1}{51 \cdot 52} = \frac{1}{221}$.

Example 1.19. What is the probability that each of four bridge players holds an ace ? [1]

Solution. Let A be the event "each player holds an ace". Then $\mathbb{P}(A) = \frac{N(A)}{N}$. The number of ways to distribute 13 cards to each player is $N = \frac{52!}{13!\,13!\,13!\,13!}$ by Theorem 1.16. For A to occur, we first give each player one ace. The number of ways to do this is $\frac{4!}{1!\,1!\,1!\,1!} = 24$. Next, we give each player 12 cards among the remaining 48 cards. This can be done in $\frac{48!}{12!\,12!\,12!\,12!}$ ways. Conclusion :

$$\mathbb{P}(A) = \frac{N(A)}{N} = 24\frac{48!}{(12!)^4} \cdot \frac{(13!)^4}{52!} = \frac{24(13)^4}{52 \cdot 51 \cdot 50 \cdot 49} \approx 0.105\,.$$

Remark 1.20. The following approximation for the factorial is often useful :

$$n! \sim \sqrt{2\pi n}\, n^n \mathrm{e}^{-n}\,.$$

In other words, $\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\, n^n \mathrm{e}^{-n}} = 1$. This is known as *Stirling's approximation*. The student will probably learn the proof in M232.

1.1 Exercises

1. A four-volume work is placed in random order on a bookshelf. What is the probability of the volumes being in proper order from left to right or from right to left ?
2. A wooden cube with painted faces is sawed up into 1000 little cubes, all of the same size. The little cubes are then mixed up, and one is chosen at random. What is the probability of its having just 2 painted faces ?
3. A batch of n manufactured items contains k defective items. Suppose m items are selected at random from the batch. What is the probability that ℓ of these items are defective ?

1. You only need to know that bridge is a game played by 4 players, each given 13 cards.

4. Ten books are placed in random order on a bookshelf. Find the probability of three given books being side by side.
5. One marksman has an 80% probability of hitting a target, while another has only a 70% probability of hitting the target. What is the probability of the target being hit (at least once) if both marksman fire at it simultaneously ?
6. Suppose n people sit down at random and independently of each other in an auditorium containing $n+k$ seats. What is the probability that m seats specified in advance ($m < n$) will be occupied ?
7. Three cards are drawn at random from a full deck. What is the probability of getting a three, a seven and an ace ?
8. What is the probability of being able to form a triangle from three segments chosen at random from five line segments of lengths $1, 3, 5, 7$ and 9 ?

   Hint. A triangle cannot be formed if one segment is longer than the sum of the other two.
9. Suppose a number from 1 to 1000 is selected at random. What is the probability that the last two digits of its cube are both 1 ?

   Hint. There is no need to look through a table of cubes.
10. Find the probability that a randomly selected positive integer will give a number ending in 1 if it is

    a) Squared;

    b) Raised to the fourth power;

    c) Multiplied by an arbitrary positive integer.

    Hint. It is enough to consider one-digit numbers.
11. One of the numbers $2, 4, 6, 7, 8, 11, 12$ and 13 is chosen at random as the numerator of a fraction, and then one of the remaining numbers is chosen at random as the denominator of the fraction. What is the probability of the fraction being in lowest terms ?
12. The word "drawer" is spelled with six scrabble tiles. The tiles are then randomly rearranged. What is the probability of the rearranged tiles spelling the word "reward" ?
13. In throwing $6n$ dice, what is the probability of getting each face n times ? Use Stirling's formula to estimate this probability for large n.
14. A full deck of cards is divided in half at random. Use Stirling's formula to estimate the probability that each half contains the same number of red and black cards.

15. Use Stirling's formula to estimate the probability that all 50 states are represented in a committee of 50 senators chosen at random.

    Note : each of the 50 states in the US has 2 senators.

16. Suppose $2n$ customers stand in line at a box office, n with 5-dollar bills and n with 10-dollar bills. Suppose each ticket costs 5 dollars, and the box office has no money initially. What is the probability that none of the customers has to wait for change ?

17. Prove that

$$\sum_{k=0}^{n} (C_k^n)^2 = C_n^{2n} .$$

    Hint : Use the binomial theorem to calculate the coefficient of x^n in the product $(1+x)^n(l+x)^n = (1+x)^{2n}$.

Chapter 2

Foundations

2.1 The axioms of probability theory

(1) We assume that random experiments can be modeled by a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

— Ω is a non-empty set, which we call "sample space".
Ω represents the set of all possible outcomes of the experiment.

— $\mathcal{F}$ is a family of "events".
Each event is represented by a subset $A \subseteq \Omega$.

— $\mathbb{P}$ is a "probability measure", it measures the probability of events in $\mathcal{F}$.

(2) The family $\mathcal{F}$ satisfies the following properties :

(i) Ω and $\emptyset$ are events called "the sure event" and "the impossible event", respectively. Hence, $\Omega, \emptyset \in \mathcal{F}$.

(ii) More generally, if A is an event [1], then $A^c = \Omega \setminus A$ is an event called "the complementary event of A".
Hence, $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.

(iii) If we have a sequence of events $A_1, A_2, \cdots \in \mathcal{F}$, then their union is an event, i.e. $\bigcup_n A_n \in \mathcal{F}$.

By (ii) and (iii) we also have $\bigcap_n A_n \in \mathcal{F}$. This follows from "De Morgan's law" : $\left(\bigcup_n A_n\right)^c = \bigcap_n A_n^c$.

(3) The probability measure $\mathbb{P}$ satisfies the following :

(i) $0 \leq \mathbb{P}(A) \leq 1$ for any event $A \in \mathcal{F}$,

(ii) $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$,

1. Rozanov uses the notation $\overline{A}$ instead of A^c and $A - B$ instead of $A \setminus B$. The student can use any notation.

(iii) If $A_1, A_2, \ldots$ is a finite or infinite sequence of events which are pairwise disjoint, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n)\,. \tag{2.1}$$

Example 2.1. Consider the experiment of throwing a dice. Then $\Omega = \{1, 2, \ldots, 6\}$. We take $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of Ω. Recall that $\mathcal{P}(\Omega)$ is the set of all subsets of Ω. As an example of events, "the dice gives an odd number" is modeled by the set $A = \{1, 3, 5\} \in \mathcal{F}$. Finally we may take $\mathbb{P}(A) = \frac{N(A)}{N}$, the uniform probability measure on Ω. This models the fact that the dice is unbiased, i.e. does not favor one result over the other. If on the contrary the dice favors for example the number 1 over the number 2, then this will be modeled by a different probability measure, say $\mathbf{P}$, which satisfies $\mathbf{P}(\{1\}) > \mathbf{P}(\{2\})$. For comparison, $\mathbb{P}(\{1\}) = \frac{1}{6} = \mathbb{P}(\{2\})$ does not represent the experiment in this case.

In general, in this course, we will usually use the rule $\mathbb{P}(A) = \frac{N(A)}{N}$ for experiments with a finite number of outcomes, we already solved some problems with this in Chapter 1. Later on in Chapter 4, we will have experiments with infinite number of outcomes, for example a complete interval $[a, b]$. In this case this rule doesn't work, we will instead use a probability measure in the form of an integral : $\mathbb{P}([a, b]) = \int_a^b p(x)\,\mathrm{d}x$ for some function $p(x)$.

Lemma 2.2. *If Ω is a finite set with N elements, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(A) = \frac{N(A)}{N}$ for any $A \subseteq \Omega$, then $(\Omega, \mathcal{F}, \mathbb{P})$ satisfies the axioms of probability.*

Proof. Clearly $\mathcal{F} = \mathcal{P}(\Omega)$ satisfies the required properties in (2). Also, $0 \leq \frac{N(A)}{N} \leq 1$ for any $A \subseteq \Omega$. We also have $\mathbb{P}(\Omega) = \frac{N}{N} = 1$ and $\mathbb{P}(\emptyset) = \frac{0}{N} = 0$ as required. Finally if $(A_k)_k$ is a sequence of sets with $A_i \cap A_j = \emptyset$ for $i \neq j$, then $N(A_i \cup A_j) = N(A_i) + N(A_j)$. More generally, $N(\cup_n A_n) = \sum_n N(A_n)$ in this case. So we get $\mathbb{P}(\cup_n A_n) = \frac{N(\cup_n A_n)}{N} = \frac{\sum_n N(A_n)}{N} = \sum_n \mathbb{P}(A_n)$, as required. This completes the proof. □

Definition 2.3. If $(\Omega, \mathcal{F}, \mathbb{P})$ satisfies the axioms of probability, we call it "a probability space".

Definition 2.4. Here is some useful vocabulary we will often use :

— A typical element of Ω is denoted by ω (instead of x for example). Do not confuse this with events. Events are *subsets* of Ω, typically denoted by A.

— We say that an event can occur if $A \neq \emptyset$, i.e. if $\exists\, \omega \in A$.

— We say that two events A_1 and A_2 are "mutually exclusive" if they cannot occur simultaneously. In other words, $A_1 \cap A_2 = \emptyset$.

— We say that the event A_1 implies the event A_2 if $A_1 \subseteq A_2$.
— Consider a sequence of events $A_1, A_2, \ldots$. The event "at least one event occurs" is modeled by the union $\bigcup_n A_n$. The event "all events occur" is modeled by the intersection $\bigcap_n A_n$.

2.2 Combination of events

Remark 2.5. Here are some basic rules from elementary set theory : consider sets A, B, C,

(i) $A \cup B = B \cup A, \quad A \cap B = B \cap A$.
(ii) $A \subseteq B \implies B^c \subseteq A^c$.
(iii) $(A \cup B)^c = A^c \cap B^c$.
(iv) $(A \cap B)^c = A^c \cup B^c$.
(v) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
(vi) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
(vii) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.
(viii) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

Lemma 2.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(ii) $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.
(iii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
(iv) *If $A \subseteq B$ then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof. (i) Since $A \cap A^c = \emptyset$, we have by the axioms of $\mathbb{P}$,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

so (i) follows.
(ii) We have $A = (A \setminus B) \cup (A \cap B)$. Moreover, $(A \setminus B) \cap (A \cap B) = \emptyset$. Thus, $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$. Item (ii) follows.
(iii) We have $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$. Moreover, these three sets on the RHS are pairwise disjoint. Hence,

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

(iv) If $A \subseteq B$ then $B = A \cup (B \setminus A)$, so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$. □

Remark 2.7. We sometimes denote

$$AB = A \cap B \qquad \text{and} \qquad A_1 A_2 \cdots A_n = A_1 \cap A_2 \cap \cdots \cap A_n \,.$$

We now come to an important generalization of item (iii) in the previous lemma. Given sets $A_1, \ldots A_n$, we ask "what is the probability that at least one event occurs ?" As we previously mentioned, this means we want to estimate $\mathbb{P}(\cup_k A_k)$. If the sets are pairwise disjoint we can just use the axiom of probability and say this is equal to $\sum_k \mathbb{P}(A_k)$. However when the sets are not necessarily pairwise disjoint, we have a more complicated formula as follows :

Theorem 2.8 (Inclusion-Exclusion identity)**.** *Given n events $A_1, A_2, \ldots A_n$ let*

$$P_1 = \sum_{i=1}^{n} \mathbb{P}(A_i)\,, \qquad\qquad P_2 = \sum_{1 \le i < j \le n} \mathbb{P}(A_i A_j)\,,$$

$$P_3 = \sum_{1 \le i < j < k \le n} \mathbb{P}(A_i A_j A_k),$$

More generally $P_r = \sum \mathbb{P}(A_{i_1} \ldots A_{i_r})$, where the sum runs over all C_r^n subsets of size r. In particular $P_n = \mathbb{P}(A_1 \ldots A_n)$. Then

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = P_1 - P_2 + P_3 - P_4 + \cdots + (-1)^{n+1} P_n \,. \tag{2.1}$$

Proof. We prove this by induction. For $n = 2$ this reduces to Lemma 2.6(iii).

Now suppose (2.1) holds for any $n - 1$ events. Then

$$\mathbb{P}\left(\bigcup_{k=2}^{n} A_k\right) = \sum_{i=2}^{n} \mathbb{P}(A_i) - \sum_{2 \le i < j \le n} \mathbb{P}(A_i A_j) + \sum_{2 \le i < j < k \le n} \mathbb{P}(A_i A_j A_k) - \ldots$$

and

$$\mathbb{P}\left(\bigcup_{k=2}^{n} A_1 A_k\right) = \sum_{i=2}^{n} \mathbb{P}(A_1 A_i) - \sum_{2 \le i < j \le n} \mathbb{P}(A_1 A_i A_j) + \sum_{2 \le i < j < k \le n} \mathbb{P}(A_1 A_i A_j A_k) - \ldots$$

So using Lemma 2.6(iii) and Remark 2.5(v) we get

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) &= \mathbb{P}(A_1) + \mathbb{P}\left(\bigcup_{k=2}^{n} A_k\right) - \mathbb{P}\left(\bigcup_{k=2}^{n} A_1 A_k\right) \\
&= \left(\mathbb{P}(A_1) + \sum_{i=2}^{n} \mathbb{P}(A_i)\right) - \left(\sum_{i=2}^{n} \mathbb{P}(A_1 A_i) + \sum_{2 \le i < j \le n} \mathbb{P}(A_i A_j)\right) \\
&\quad + \left(\sum_{2 \le j < k \le n} \mathbb{P}(A_1 A_j A_k) + \sum_{2 \le i < j < k \le n} \mathbb{P}(A_i A_j A_k)\right) - \ldots
\end{aligned}$$

completing the proof, by definition of P_j. □

Example 2.9. Suppose n students have n identical raincoats. To attend class, the students hang their coats on the same coat rack without paying attention. After class, the students are unable to distinguish the coats as they all look the same, so each student picks a raincoat at random. What is the probability that at least one raincoat ends up with its original owner ?

Solution. Let us number the students from 1 to n. Let A_k be the event "student k gets his original raincoat". Then we seek to estimate $\mathbb{P}(\cup_{k=1}^n A_k)$.

We first examine if the sets are pairwise disjoint or not. Here $A_i \cap A_j$ is the event "student i and student j both receive their original raincoats". This event can happen, it is not impossible, so it is not $\emptyset$. Since $A_i \cap A_j \neq \emptyset$, we cannot just use axiom (2.1), we have to use instead the inclusion-exclusion principle.

As usual, first calculate the denominator N. We have n students picking n coats at random. Student 1 picks one of n coats, then student 2 picks one of the remaining $n-1$ coats, etc. This can be done in $n(n-1)(n-2)\cdots 2\cdot 1 = n!$ ways. Thus, $N = n!$

Next, consider the event $A_{i_1}A_{i_2}\cdots A_{i_r}$ where $i_1 < i_2 < \cdots < i_r$. This event means "students $i_1, i_2, \ldots, i_r$ each gets his own raincoat". This can happen in many ways. Namely, after students i_1 to i_r have taken their own coats, the remaining $n-r$ students can choose at random among the $n-r$ remaining coats. The number of ways to do this is $(n-r)!$ as in the calculation of N. Thus, $\mathbb{P}(A_{i_1}\cdots A_{i_r}) = \frac{(n-r)!}{n!}$.

Now recall that $P_r = \sum_{i_1<\cdots<i_r} \mathbb{P}(A_{i_1}\cdots A_{i_r})$. The general term of this sum has value $\frac{(n-r)!}{n!}$ independently of i_j. So $P_r =$ (number of terms in the sum) $\times$ $\frac{(n-r)!}{n!}$. As mentioned in Theorem 2.8, the number of terms is C_r^n. Hence, $P_r = C_r^n \frac{(n-r)!}{n!} = \frac{n!}{r!(n-r)!}\frac{(n-r)!}{n!} = \frac{1}{r!}$.

Conclusion : by the inclusion exclusion principle,

$$\mathbb{P}(\cup_{k=1}^n A_k) = P_1 - P_2 + \cdots + (-1)^{n+1}P_n = 1 - \frac{1}{2!} + \cdots + (-1)^{n+1}\frac{1}{n!}$$

To get an idea of how big this is, we use Taylor's expansion to get [2]

$$\mathbb{P}(\cup_{k=1}^n A_k) \approx 1 - \mathrm{e}^{-1} \approx 0.632 \tag{2.2}$$

We now take a few more theorems.

Definition 2.10. A sequence of sets $A_1, A_2, \ldots$ is said to be "increasing" if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$. Similarly, the sequence is "decreasing" if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$.

2. Recalling that $\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \ldots$, we have $\mathrm{e}^{-1} = \frac{1}{2!} + \cdots + \frac{(-1)^n}{n!} + \ldots$. Thus, $1 - \mathrm{e}^{-1} = \mathbb{P}(\cup_{k=1}^n A_k) + \frac{(-1)^{n+2}}{(n+1)!} + \frac{(-1)^{n+3}}{(n+2)!} + \ldots$. The remainder is very small if n is large enough.

Theorem 2.11. *If $A_1, A_2, \ldots$ is an increasing sequence of events, then*

$$\mathbb{P}\left(\bigcup_k A_k\right) = \lim_{n\to\infty} \mathbb{P}(A_n)\,.$$

Proof. Let $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n > 1$. Then

(2.3)

$$A_n = (A_n \setminus A_{n-1}) \cup A_{n-1} = (A_n \setminus A_{n-1}) \cup (A_{n-1} \setminus A_{n-2}) \cup A_{n-2} = \cdots = \cup_{k=1}^n B_k\,.$$

Moreover, the sets B_j are pairwise disjoint. This is clear by making a picture. In equations, let $i \neq j$, say $i < j$. Then $B_j = A_j \setminus A_{j-1}$. But $A_{j-1} \supseteq A_i \supseteq B_i$. So $B_j \cap B_i \subseteq (A_j \setminus B_i) \cap B_i = \emptyset$.

From the above we get

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \mathbb{P}(\cup_{k=1}^n B_k) = \lim_{n\to\infty} \sum_{k=1}^n \mathbb{P}(B_k) = \sum_{k=1}^\infty \mathbb{P}(B_k) = \mathbb{P}(\cup_{k=1}^\infty B_k) = \mathbb{P}(\cup_{k=1}^\infty A_k)\,.$$

In the last step we used that $\cup_{k=1}^\infty B_k = \cup_{k=1}^\infty A_k$. In fact, using (2.3), we have $\omega \in \cup_{k=1}^\infty A_k \iff \omega \in A_k$ for some $k \geq 1 \iff \omega \in \cup_{r=1}^k B_r$ for some $k \geq 1 \iff \omega \in B_r$ for some $r \geq 1 \iff \omega \in \cup_{r=1}^\infty B_r$. □

There is an obvious counterpart for this :

Theorem 2.12. *If $A_1, A_2, \ldots$ is a decreasing sequence of events, then*

$$\mathbb{P}\left(\bigcap_k A_k\right) = \lim_{n\to\infty} \mathbb{P}(A_n)\,.$$

Proof. Using De Morgan and Lemma 2.6(i), we have $\mathbb{P}(\cap_k A_k) = 1 - \mathbb{P}(\cup_k A_k^c)$.

But using Remark 2.5(ii), the sequence $A_1^c, A_2^c, \ldots$ is increasing.

Hence, using Theorem 2.11 we conclude that

$$\mathbb{P}(\cap_k A_k) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = 1 - \lim_{n\to\infty} (1 - \mathbb{P}(A_n)) = \lim_{n\to\infty} \mathbb{P}(A_n)\,.$$ □

As we insisted during this chapter, the rule (2.1) is only valid if the sets are pairwise disjoint. On the other hand, the inclusion-exclusion principle is useful for the general case, but the result is complicated. If we only care about an upper bound (not an exact value), then the following theorem suffices :

Theorem 2.13. *For any sequence of events $A_1, A_2, \ldots$, we have*

$$\mathbb{P}\left(\bigcup_k A_k\right) \leq \sum_k \mathbb{P}(A_k)\,.$$

Proof. Define $C_n = \cup_{k=1}^n A_k$. Then $C_1, C_2, \ldots$ is increasing.

Define $B_1 = C_1$ and $B_n = C_n \setminus C_{n-1}$ for $n > 1$. Then as shown in the proof of Theorem 2.11, we have

$$\mathbb{P}(\cup_n C_n) = \mathbb{P}(\cup_n B_n) = \sum_n \mathbb{P}(B_n)\,.$$

Clearly $\cup_n C_n = \cup_n A_n$. Also,

$$B_n = C_n \setminus C_{n-1} = [A_n \cup (\cup_{k=1}^{n-1} A_k)] \setminus (\cup_{k=1}^{n-1} A_k) = A_n \setminus (\cup_{k=1}^{n-1} A_k) \subseteq A_n\,.$$

So by Lemma 2.6(iv) we have $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$. This completes the proof. □

We may now conclude the chapter with a fundamental result :

Theorem 2.14 (First Borel-Cantelli lemma)**.** *Given an infinite sequence of events $A_1, A_2, \ldots$, suppose that $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$. Then with probability one, only finitely many of the events $A_1, A_2, \ldots$ occurs.*

Proof. Let B be the event "infinitely many events occur". By Lemma 2.6(i), it suffices to prove $\mathbb{P}(B) = 0$.

We claim that $B = \cap_n \cup_{k \geq n} A_k$. Indeed, $\omega \in \cap_n \cup_{k \geq n} A_k \iff \omega \in \cup_{k \geq n} A_k$ for all $n \iff \omega \in A_k$ for some $k \geq n$, for all $n \iff \forall n\, \exists k \geq n : \omega \in A_k$. Taking $n = 1, 2, \ldots$, we find corresponding $k_1, k_2, \ldots$. So this is equivalent to $\omega \in A_{k_1}$ and A_{k_2} and A_{k_3}, ... an infinite sequence of events. By definition this is equivalent to $\omega \in B$.

Let $B_n = \cup_{k \geq n} A_k$. We showed that $B = \cap_n B_n$. Moreover, $B_1 \supseteq B_2 \supseteq \ldots$ is decreasing. So by Theorem 2.12 we have $\mathbb{P}(B) = \lim_{n \to \infty} \mathbb{P}(B_n)$. On the other hand, using Theorem 2.13, we have $\mathbb{P}(B_n) \leq \sum_{k \geq n} \mathbb{P}(A_k)$. To summarize, we have proved that $\mathbb{P}(B) \leq \lim_{n \to \infty} \sum_{k \geq n} \mathbb{P}(A_k)$. But by hypothesis, $\sum_{k \geq 1} \mathbb{P}(A_k) < \infty$. This implies the remainder $\sum_{k \geq n} \mathbb{P}(A_k)$ has limit zero as $n \to \infty$. So $\mathbb{P}(B) = 0$ as required. □

Example 2.15. A boy can play an infinite sequence of games. In game n, he loses 2^n\$ with probability $\frac{1}{2^n+1}$ and wins 1\$ with probability $\frac{2^n}{2^n+1}$. Is it to the boy's advantage to continue to play infinitely many times ?

Solution. Let A_n be the event "the boy loses game n.

We notice that $\sum_n \mathbb{P}(A_n) = \sum_n \frac{1}{2^n+1} \leq \sum_n \frac{1}{2^n} = \frac{1}{1-\frac{1}{2}} = 2 < \infty$. So by the first Borel-Cantelli lemma, we know that with probability one, only finitely many A_n can occur. In other words, we are *sure* that after some n_0, the boy will always win. So the boy should definitely continue to play, as he will get infinitely rich as time goes to infinity.

2.3 Exercises

1. Interpret the following relations involving events A, B and C :

$$a)\ AB = A; \qquad b)\ ABC = A; \qquad c)\ A \cup B \cup C = A.$$

2. When do the following relations involving the events A and B hold :

$$a)\ A \cup B = A^c; \qquad b)\ AB = A^c; \qquad c)\ A \cup B = AB\ ?$$

3. Simplify the following expressions involving events A, B and C :

$$a)\ (A \cup B)(B \cup C); \qquad b)(A \cup B)(A \cup B^c); \qquad c)(A \cup B)(A \cup B^c)(A^c \cup B).$$

4. Given two events A and B, find the event X such that

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

5. Let A be the event that at least one of three inspected items is defective, and B the event that all three items are of acceptable quality. What are the events $A \cup B$ and AB ?
6. A whole number from 1 to 1000 is chosen at random. Let A be the event that the number is divisible by 5, and B the event that the number ends in a zero. What is the event AB^c ?
7. A target is made up of 10 circular disks bounded by 10 concentric circles of radii $r_1, r_2, \ldots, r_{10}$ where $r_1 < r_2 < \cdots < r_{10}$. Let A_k be the event consisting of the disk of radius r_k being hit ($k = 1, 2, \ldots, 10$). What are the events

$$B = \bigcup_{k=1}^{6} A_k, \qquad C = \bigcap_{k=5}^{10} A_k\ ?$$

8. Given any event A prove that

$$\mathbb{P}(A) = 1 - P(A^c), \qquad \mathbb{P}(A^c) = 1 - \mathbb{P}(A)\,.$$

9. A marksman fires at a target made up of a central circular disk and two concentric rings. The probabilities of hitting the disk and the rings are 0.35, 0.30 and 0.25, respectively. What is the probability of missing the target ?
10. Five items are chosen at random from a batch of 100 items and then inspected. The whole batch is rejected if any of the items is found to be defective. What is the probability of the batch being rejected if it contains 5 defective items ?

11. A secretary forgets the last digit of a telephone number, and dials the last digit at random. What is the probability of calling no more than three wrong numbers ? How is this probability changed if she recalls that the last digit is even ?

12. Given any n events $A_1, A_2, \ldots, A_n$, prove that

$$\mathbb{P}\left(\bigcap_{k=1}^{n} A_k\right) = 1 - \mathbb{P}\left(\bigcup_{k=1}^{n} A_k^c\right).$$

13. A batch of 100 manufactured items contains 5 defective items. Fifty items are chosen at random and then inspected. Suppose the whole batch is accepted if no more than one of the 50 inspected items is defective. What is the probability of accepting the whole batch ?

14. Write an expression for the probability $p(r)$ that among r randomly selected people, at least two have a common birthday.
Comment. Rather surprisingly, it turns out that $p(r) > \frac{1}{2}$ if $r = 23$.

15. Test approximation (2.2) for $n = 3, 4, 5$ and 6.

16. Use Theorem 2.8 and Stirling's formula to find the probability that some player is dealt a complete suit in a game of bridge.

17. Given any n events $A_1, A_2, \ldots, A_n$, prove that the probability of exactly m ($m \leq n$) of the events occurring is

$$P_m - \binom{m+1}{m} P_{m+1} + \binom{m+2}{m} P_{m+2} - \cdots \pm \binom{n}{m} P_n\,,$$

where $P_m, P_{m+1}, \ldots$ are the same as in Theorem 2.8.

18. Let $n = 10$ in Example 2.9. What is the probability that exactly 5 raincoats end up with their original owners ?

19. A whole number from 1 to 1000 is chosen at random. What is the probability of its being a power (higher than the first) of another whole number ?
Hint. $31^2 < 1000 < 32^2$.

Chapter 3

Dependence

In this chapter we would like to formalize our intuition about dependence and independence among events.

3.1 Conditional probability

If A, B are two events, several situations may occur. It may be that the two events have nothing to do with each other; knowing that A occurred does not help us at all to decide if B occurred. At the other extreme, it may be that the occurrence of A implies the occurrence of B. Then there are all kinds of situations in between.

To measure how the events A, B influence each other we introduce :

Definition 3.1. Let A, B be two events with $\mathbb{P}(B) > 0$. We define the *conditional probability of A given B* by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}. \tag{3.1}$$

To clarify the meaning of (3.1), consider the case where the experiment has finitely many equiprobable outcomes, so that we use the rule $\mathbb{P}(A) = \frac{N(A)}{N}$. Then

$$\mathbb{P}(A \mid B) = \frac{N(AB)}{N} \cdot \frac{N}{N(B)} = \frac{N(AB)}{N(B)}.$$

This means that B has become the new sample space : instead of estimating the relative size of A in the full set Ω, we are only estimating the relative size of A in B. In other words, $\mathbb{P}(A \mid B)$ asks "what is the probability that A occurs, if you already know that B has occurred". Indeed, if you know that B has occurred and ask about A, then you should ignore all outcomes which are not in B.

Lemma 3.2. (1) $\mathbb{P}(\cdot \mid B)$ *is a probability measure on* $(\Omega, \mathcal{F})$.

(2) *If* A *and* B *are incompatible, i.e. if* $AB = \emptyset$, *then* $\mathbb{P}(A \mid B) = 0$.

(3) *If* B *implies* A, *i.e.* $B \subseteq A$, *then* $\mathbb{P}(A \mid B) = 1$.

Items (2) and (3) are in harmony with our intuition : if A and B are incompatible, i.e. mutually exclusive, and if we know that B has occurred, then it is impossible for A to occur, so $\mathbb{P}(A \mid B) = 0$. Similarly, if we know that B implies A and that B has occurred, then A must certainly occur, so $\mathbb{P}(A \mid B) = 1$.

Proof. (1) Since $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$, then $\mathbb{P}(A \mid B) \geq 0$ and $\mathbb{P}(A \mid B) \leq 1$ (because $\mathbb{P}(AB) \leq \mathbb{P}(B)$).

Next, $\mathbb{P}(\emptyset \mid B) = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0$ and $\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$.

Finally if $A_1, A_2, \ldots$ are pairwise disjoint, then $A_1B, A_2B, \ldots$ are also pairwise disjoint, since $A_iB \cap A_jB \subseteq A_i \cap A_j = \emptyset$ for $i \neq j$. Consequently

$$\mathbb{P}(\cup_k A_k \mid B) = \frac{\mathbb{P}(\cup_k A_k B)}{\mathbb{P}(B)} = \frac{\sum_k \mathbb{P}(A_k B)}{\mathbb{P}(B)} = \sum_k \mathbb{P}(A_k \mid B)\,.$$

This proves that $\mathbb{P}(\cdot \mid B)$ is a probability measure.

(2) If $AB = \emptyset$ then $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0$.

(3) If $B \subseteq A$ then $AB = B$ so $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$. □

We now give an important theorem which allows us to compute the probability of A by partitioning the sample space into convenient sets.

Definition 3.3. We say that $B_1, B_2, \ldots$ form a *full set of mutually exclusive events* if $B_i \cap B_j = \emptyset$ for $i \neq j$ and $\cup_k B_k = \Omega$.

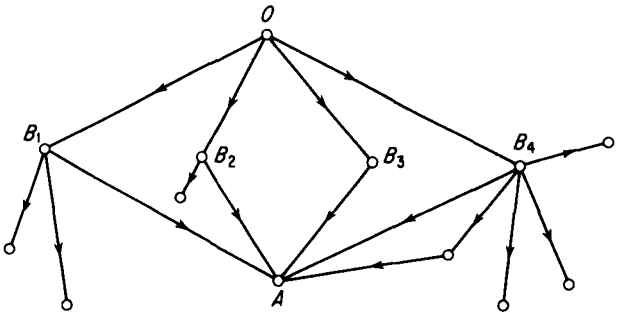
Theorem 3.4 (Total probability formula)**.** *If* $B_1, B_2, \ldots$ *form a full set of mutually exclusive events,* $\mathbb{P}(B_k) > 0$, *then*

$$\mathbb{P}(A) = \sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)\,.$$

Proof. We have $A = A \cap \Omega = A \cap (\cup_k B_k) = \cup_k (A \cap B_k) = \cup_k AB_k$. Moreover, since B_k are pairwise disjoint then AB_k are also pairwise disjoint as we saw before. Hence,

$$\mathbb{P}(A) = \mathbb{P}(\cup_k AB_k) = \sum_k \mathbb{P}(AB_k) = \sum_k \frac{\mathbb{P}(AB_k)}{\mathbb{P}(B_k)}\,\mathbb{P}(B_k) = \sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)\,. \quad \square$$

Example 3.5. A hiker leaves point O shown in the figure, choosing one of the roads OB_1, OB_2, OB_3, OB_4 at random. At each subsequent crossroads he again chooses a road at random. What is the probability that the hiker arrives at A ?



Solution. Let B_k be the event " the hiker passes by point B_k". Then $B_1, \dots, B_4$ form a full set of mutually exclusive events, and $\mathbb{P}(B_k) = \frac{1}{4}$ for each k since the hiker chooses uniformly at random. So by total probability,

$$\mathbb{P}(A) = \sum_{k=1}^{4} \mathbb{P}(A \mid B_k)\, \mathbb{P}(B_k) = \frac{1}{4} \sum_{k=1}^{4} \mathbb{P}(A \mid B_k)\,.$$

Next, if the hiker is at B_1, then he may reach A with probability $\frac{1}{3}$, so $\mathbb{P}(A \mid B_1) = \frac{1}{3}$. Similarly, we find that $\mathbb{P}(A \mid B_2) = \frac{1}{2}$, $\mathbb{P}(A \mid B_3) = 1$, $\mathbb{P}(A \mid B_4) = \frac{2}{5}$. Conclusion :

$$\mathbb{P}(A) = \frac{1}{4}\Big(\frac{1}{3} + \frac{1}{2} + 1 + \frac{2}{5}\Big) = \frac{67}{120}\,.$$

Example 3.6 (Optimal choice)**.** A set of m suitors propose in succession to a fussy young lady. She wants the best possible partner but she doesn't know in advance who it is, she must examine the suitors one by one. We assume the following :

(1) The lady can accept the first suitor, or reject him in the hope of finding a better partner. Similarly for the next marriage proposals. Note that she doesn't know in advance the "quality" of the suitors that will come later.

(2) A rejected suitor will not propose again, so she loses him forever.

(3) The lady never selects a suitor inferior to those previously rejected.

Note that rule (3) may cause the lady to never marry. This happens for example if the first suitor was in fact the best of them all, and she rejected him.

Suppose however that the lady gets married by accepting the i-th suitor. What is the probability that he is indeed the best of all m suitors ?

Solution. By rule (3), if she accepted suitor i, then he must be better than all those before him. So there is an implicit information in the problem. More precisely, if B is the event that the i-th suitor is the best among the first i suitors, and A is the event that the i-th suitor is the best of all m suitors, then the problem asks us to estimate $\mathbb{P}(A \mid B)$.

Clearly $A \subseteq B$. Consequently $AB = A$, so $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$.

To estimate B, we arrange the suitors as $(S_1, S_2, \ldots, S_i)$. The best suitor of these i gentlemen is at a random position. These suitors can be arranged in a total of $i!$ ways. For B to occur, we must fix S_i to be the best, but the first $i-1$ suitors can take any arrangement, this is done in $(i-1)!$ ways. Thus, $\mathbb{P}(B) = \frac{(i-1)!}{i!} = \frac{1}{i}$.

Similarly, for A we arrange the m suitors as $(S_1, \ldots, S_i, \ldots, S_m)$. Again the best suitor is at a random position. We have $m!$ possible arrangements for these suitors. For A to occur, we must fix the best to take the i-th position, but the remaining $m-1$ suitors can arrange themselves in any manner, there are $(m-1)!$ such arrangements. Hence, $\mathbb{P}(A) = \frac{(m-1)!}{m!} = \frac{1}{m}$.

Conclusion : $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{i}{m}$.

Example 3.7 (Gambler's ruin)**.** Adam plays a game of "head or tails" in which a coin is tossed, he wins 1 dollar if he successfully guesses the outcome and loses 1 dollar otherwise. He starts with x dollars and intends to play until he has $m \geq x$ dollars. What is the probability that Adam will be ruined, i.e. lose all his money without reaching his aim ?

Solution. Let A be the event of ruin and $p(x) = \mathbb{P}(A)$ be the probability of ruin. If Adam wins the first game, his new capital becomes $x+1$, so his new probability of ruin becomes $p(x+1)$. Similarly, if he loses the first game, his new capital becomes $x-1$, so his ruin probability becomes $p(x-1)$. In other words, if B_1 is the event "Adam wins the first game" and B_2 "Adam loses the first game", then $\mathbb{P}(A \mid B_1) = p(x+1)$ and $\mathbb{P}(A \mid B_2) = p(x-1)$. But $\mathbb{P}(B_1) = \mathbb{P}(B_2) = \frac{1}{2}$. Since B_1, B_2 clearly form a full set of mutually exclusive events, we get by total probability

$$p(x) = \frac{1}{2}\left(p(x+1) + p(x-1)\right). \tag{3.2}$$

This is a recurrence relation. The student will learn how to solve this in linear algebra. In this course we will just guess a solution and check if it is correct or not. So let us guess that the solution to this relation is

$$p(x) = C_1 + C_2 x\,. \tag{3.3}$$

To check if (3.2) is valid, we compute $\frac{1}{2}(C_1 + C_2(x+1) + C_1 + C_2(x-1)) = \frac{1}{2}(2C_1 + 2C_2 x) = C_1 + C_2 x = p(x)$. So (3.3) is indeed a solution to the problem.

To find C_1 and C_2 we use the "initial conditions". In fact, we clearly have

$$p(0) = 1 \qquad \text{and} \qquad p(m) = 0\,. \tag{3.4}$$

This just says that if Adam has no money to begin with then he is certainly ruined, and if he already has m dollars, then he will certainly not get ruined since he won't play. Substituting (3.4) into (3.3), we get $1 = C_1$ and $0 = C_1 + C_2 m$, so $C_2 = \frac{-1}{m}$. Conclusion :

$$p(x) = 1 - \frac{x}{m}.$$

Let us now go back to some theory. It often happens that $\mathbb{P}(B \mid A)$ is easier to compute than $\mathbb{P}(A \mid B)$. Is there a relation between them ? This is the content of the famous :

Theorem 3.8 (Bayes' law)**.** *Suppose* $\mathbb{P}(A), \mathbb{P}(B) > 0$*. Then*

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

If $B_1, \ldots, B_n$ *form a total set of mutually exclusive events, then*

$$\mathbb{P}(B_k \mid A) = \frac{\mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}{\sum_{j=1}^{n} \mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}. \tag{3.5}$$

Proof. We have $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(BA)}{\mathbb{P}(A)}$ and $\frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\frac{\mathbb{P}(AB)}{\mathbb{P}(B)}\,\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}$ so both sides are equal. Claim (3.5) is by the total probability formula : $\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)$. □

Example 3.9. In answering a question of multiple choice, a student either knows the answer or guesses. Let p be the probability that he knows the answer and $1 - p$ the probability that he guesses. In case of a guess, the student answers correctly with probability $\frac{1}{m}$, where m is the number of choices in the question.

What is the probability that the student knows the answer, given that he answered correctly ?

Solution. Let C be the event "the student answers correctly" and K the event "the student knows the answer". We seek $\mathbb{P}(K \mid C)$, which is not entirely clear. However, we clearly have $\mathbb{P}(C \mid K) = 1$. So we use Bayes' formula to conclude that

$$\mathbb{P}(K \mid C) = \frac{\mathbb{P}(C \mid K)\,\mathbb{P}(K)}{\mathbb{P}(C \mid K)\,\mathbb{P}(K) + \mathbb{P}(C \mid K^c)\,\mathbb{P}(K^c)} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{m}(1-p)} = \frac{mp}{1 + (m-1)p}.$$

This completes the solution.

For example, if $m = 4$ as in bubble sheets and $p = \frac{1}{2}$, then $\mathbb{P}(K \mid C) = \frac{4/2}{1+3/2} = \frac{4}{5}$.

3.2 Independence

So far we have studied how the knowledge of the occurrence of some event B can affect our prediction of an event A. We now want to introduce a mathematical condition which guarantees that two events A and B have no influence on each other, that they are independent in the intuitive sense.

Definition 3.10. We say that the events A and B are *independent* if

$$\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B)\,. \tag{3.1}$$

The meaning of (3.1) becomes more clear with the following :

Lemma 3.11. *The events A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.*

The fact that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ means that the information that B has occurred is completely useless. It has not affected at all our estimation of the probability of A. In other words, B has nothing to do with A. This is indeed what we have in mind when we say that two events are independent.

Proof. $\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B) \iff \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \mathbb{P}(A) \iff \mathbb{P}(A \mid B) = \mathbb{P}(A)$. □

Example 3.12. Let A be the event that a card picked at random from a full deck is a spade and B the event that it is a queen. Are A and B independent ?

Solution. The answer may not be clear based on intuition alone. Let us calculate. We have $\mathbb{P}(A) = \frac{13}{52} = \frac{1}{4}$, $\mathbb{P}(B) = \frac{4}{52} = \frac{1}{13}$. Next, AB is the event that we pick a queen of spades. There is only one such card so $\mathbb{P}(AB) = \frac{1}{52}$. But $\mathbb{P}(A)\,\mathbb{P}(B) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}$. Hence, $\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B)$ and the events are independent.

Example 3.13. Throw a pair of dice, let A_1 be the event "the first die turns up odd", A_2 the event "the second die turns up odd" and A_3 the event "the total number of spots is odd". Show that A_1, A_2 and A_3 are pairwise independent.

Solution. $\Omega = \{(x,y) : x,y \in \{1,\dots,6\}\}$, $A_1 = \{(x,y) : x \in \{1,3,5\}, y \in \{1,\dots,6\}\}$, $A_2 = \{(x,y) : x \in \{1,\dots,6\}, y \in \{1,3,5\}\}$. So $\mathbb{P}(A_1) = \frac{3\times 6}{6^2} = \frac{3}{6} = \frac{1}{2}$ and $\mathbb{P}(A_2) = \frac{6\times 3}{6^2} = \frac{1}{2}$. On the other hand $A_1A_2 = \{(x,y) : x,y \in \{1,3,5\}\}$ so $\mathbb{P}(A_1A_2) = \frac{3^2}{6^2} = \frac{1}{4} = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$. Thus, A_1 and A_2 are independent.

Next, let B_1 be the event "the first die is odd and the second is even" and B_2 the event "the first die is even and the second is odd". Then clearly $A_3 = B_1 \cup B_2$ and $B_1 \cap B_2 = \emptyset$, so $\mathbb{P}(A_3) = \mathbb{P}(B_1) + \mathbb{P}(B_2) = \frac{3\times 3}{6^2} + \frac{3\times 3}{6^2} = \frac{18}{36} =$

$\frac{1}{2}$. Finally $A_1A_3 = A_1(B_1 \cup B_2) = A_1B_1 = B_1$ since $A_1 \cap B_2 = \emptyset$ and $B_1 \subseteq A_1$. So $\mathbb{P}(A_1A_3) = \mathbb{P}(B_1) = \frac{3\times 3}{6^2} = \frac{1}{4} = \mathbb{P}(A_1)\,\mathbb{P}(A_3)$. So A_1 and A_3 are independent. Similarly, $\mathbb{P}(A_2A_3) = \mathbb{P}(B_2) = \frac{1}{4} = \mathbb{P}(A_2)\,\mathbb{P}(A_3)$ so A_2 and A_3 are independent. [1]

So far we only defined independence for pairs of events. How about triples ? does the above example imply that A_1, A_2, A_3 are independent ?

Definition 3.14. We say that $A_1, A_2, \dots A_n$ are independent if $\mathbb{P}(\cap_{j\in J} A_j) = \prod_{j\in J} \mathbb{P}(A_j)$ for any $J \subseteq \{1, \dots, n\}$. In more detail, this means we ask that

$$\begin{aligned}
\mathbb{P}(A_iA_j) &= \mathbb{P}(A_i)\,\mathbb{P}(A_j) && \forall\ 1 \le i < j \le n\,,\\
\mathbb{P}(A_iA_jA_k) &= \mathbb{P}(A_i)\,\mathbb{P}(A_j)\,\mathbb{P}(A_k) && \forall\ 1 \le i < j < k \le n\,,\\
&\cdots\cdots\cdots &&\\
\mathbb{P}(A_1A_2\cdots A_n) &= \mathbb{P}(A_1)\,\mathbb{P}(A_2)\cdots\mathbb{P}(A_n)\,. &&
\end{aligned}$$

Example 3.15. Back to Example 3.13, we have verified that A_1, A_2, A_3 are pairwise independent. However, $A_1A_2A_3 = \emptyset$ since if both dice turn up odd then the sum cannot be odd. So $\mathbb{P}(A_1A_2A_3) = 0$. On the other hand, $\mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\mathbb{P}(A_3) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$. So $\mathbb{P}(A_1A_2A_3) \neq \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\mathbb{P}(A_3)$. Thus, when looking at all events at the same time, A_1, A_2, A_3 are not independent.

Before concluding this chapter with an important result, we need two lemmas.

Lemma 3.16. *If $A_1, \dots A_n$ are independent then $A_1^c, \dots, A_n^c$ are independent.*

Proof. The case $n = 2$ is Exercise 7. The general case requires the inclusion-exclusion principle and a similar expansion for $\prod_{r=1}^{k}(1 - p_r)$, applied to $p_r = \mathbb{P}(A_r)$. We omit the details. □

Lemma 3.17. *We have $1 - x \le \mathrm{e}^{-x}$ for any $x \ge 0$.*

Proof. Let $f(x) = \mathrm{e}^{-x} + x - 1$. Then $f(0) = 0$ and $f'(x) = -\mathrm{e}^{-x} + 1 \ge 0$. So f is increasing. So $f(x) \ge f(0) = 0$ for any $x \ge 0$. This completes the proof. □

Definition 3.18. We say that an infinite sequence $A_1, A_2, \dots$ is independent if $A_1, \dots, A_n$ are independent for all n.

1. We presented a detailed answer for pedagogic purposes. The solution in the book [3] is also correct but may not be sufficiently detailed for students.

Theorem 3.19 (Second Borel-Cantelli Lemma)**.** *Given an infinite sequence of independent events $A_1, A_2, \ldots$, suppose that $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \infty$. Then with probability one, infinitely many of the events $A_1, A_2, \ldots$ occur.*

Proof. As we proved in the First Borel-Cantelli Lemma, $\cap_n(\cup_{k\geq n} A_k)$ represents the situation that infinitely many events occur. So if $B_n = \cup_{k\geq n} A_k$ and $B = \cap_n B_n$, we should show that $\mathbb{P}(B) = 1$. It suffices to show that $\mathbb{P}(B^c) = 0$ because $\mathbb{P}(B) = 1 - \mathbb{P}(B^c)$.

By De-Morgan, $B^c = \cup_n B_n^c$ and $B_n^c = \cap_{k\geq n} A_k^c$.

Fix any $m \geq 0$. Then $B_n^c \subseteq \cap_{k=n}^{n+m} A_k^c$. Consequently,

$$\mathbb{P}(B_n^c) \leq \mathbb{P}\Big(\bigcap_{k=n}^{n+m} A_k^c\Big) = \mathbb{P}(A_1^c)\cdots\mathbb{P}(A_{n+m}^c) = (1-\mathbb{P}(A_1))\cdots(1-\mathbb{P}(A_{n+m}))$$
$$\leq \mathrm{e}^{-\mathbb{P}(A_1)}\cdots\mathrm{e}^{-\mathbb{P}(A_{n+m})} = \exp\Big(-\sum_{k=n}^{n+m}\mathbb{P}(A_k)\Big),$$

where we used Lemmas 3.16 and 3.17 above. Since m is arbitrary, we may take $m \to \infty$ above. The left-hand side $\mathbb{P}(B_n^c)$ is independent of m and stays as it is. As for the right-hand side, since the exponential function is continuous, it tends to $\exp\Big(-\sum_{k=n}^{\infty}\mathbb{P}(A_k)\Big)$. But by hypothesis $\sum_{k=1}^{\infty}\mathbb{P}(A_k) = \infty$, so $\sum_{k=n}^{\infty}\mathbb{P}(A_k) = \infty$ for any n.[2]

Summarizing we have shown that

$$\mathbb{P}(B_n^c) \leq \exp\Big(-\sum_{k=n}^{\infty}\mathbb{P}(A_k)\Big) = \exp(-\infty) = 0$$

for any n. It follows that

$$\mathbb{P}(B^c) = \mathbb{P}\Big(\bigcup_n B_n^c\Big) \leq \sum_n \mathbb{P}(B_n^c) = 0\,.$$

Thus, $\mathbb{P}(B^c) = 0$. This completes the proof of the theorem. □

3.3 Exercises

1. Given any events A and B, prove that the events A, A^cB and $(A\cup B)^c$ form a full set of mutually exclusive events.

2. If $\sum_{k=n}^{\infty}\mathbb{P}(A_k)$ was finite then $\sum_{k=1}^{\infty}\mathbb{P}(A_k) = \sum_{k=1}^{n-1}\mathbb{P}(A_k) + \sum_{k=n}^{\infty}\mathbb{P}(A_k)$ would also be finite, a contradiction.

2. In a game of chess, let A be the event that White wins and B the event that Black wins. What is the event C such that A, B and C form a full set of mutually exclusive events ?

3. Prove that if $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, then $\mathbb{P}(B \mid A) > \mathbb{P}(B)$.

4. Prove that if $\mathbb{P}(A) = \mathbb{P}(B) = \frac{2}{3}$ then $\mathbb{P}(A \mid B) \geq \frac{1}{2}$.

5. Given any three events A, B and C, prove that

$$\mathbb{P}(ABC) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid AB)\,.$$

   Generalize this formula to the case of any n events.

6. Verify that

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

   if

$$a)\ A = \emptyset; \qquad b)\ B = \emptyset; \qquad c)\ B = \Omega; \qquad d)\ B = A; \qquad e)\ B = A^c\,.$$

7. Prove that if the events A and B are independent, then so are their complements.

8. Two events A and B with positive probabilities are incompatible. Are they dependent ?

9. Consider n urns, each containing w white balls and b black balls. A ball is drawn at random from the first urn and put into the second urn, then a ball is drawn at random from the second urn and put into the third urn, and so on, until finally a ball is drawn from the last urn and examined. What is the probability of this ball being white ?

10. In Example 3.5 find the probability of the hiker arriving at each of the 6 destinations other than A. Verify that the sum of the probabilities of arriving at all possible destinations is 1.

11. Prove that the probability of ruin in Example 3.7 does not change if the stakes are changed.

12. Prove that the events A and B are independent if $\mathbb{P}(B \mid A) = \mathbb{P}(B \mid A^c)$.

13. One urn contains w_1 white balls and b_1 black balls, while another urn contains w_2 white balls and b_2 black balls. A ball is drawn at random from each urn, and then one of the two balls so obtained is chosen at random. What is the probability of this ball being white ?

14. Nine out of 10 urns contain 2 white balls and 2 black balls each, while the other urn contains 5 white balls and 1 black ball. A ball drawn from a randomly chosen urn turns out to be white. What is the probability that the ball came from the urn containing 5 white balls ?

15. One urn contains only white balls, while another urn contains 30 white balls and 10 black balls. An urn is selected at random, and then a ball is drawn (at random) from the urn. The ball turns out to be white, and is then put back into the urn. What is the probability that another ball drawn from the same urn will be black ?

16. Two balls are drawn from an urn containing n balls numbered from 1 to n. The first ball is kept if it is numbered 1, and returned to the urn otherwise. What is the probability of the second ball being numbered 2 ?

17. A regular tetrahedron is made into an unbiased die, by labeling the four faces a, b, c and abc, respectively. Let A be the event that the die falls on either of the two faces bearing the letter a, B the event that it falls on either of the two faces bearing the letter b, and C the event that it falls on either of the two faces bearing the letter c. Prove that the events A, B and C are "pairwise independent" but not independent.

18. An urn contains w white balls, b black balls and r red balls. Find the probability of a white ball being drawn before a black ball if

    a) Each ball is replaced after being drawn;

    b) No balls are replaced.

19. (The infinite monkey theorem). Suppose a monkey hits the keys of a typewriter uniformly at random, infinitely many times. Prove that with probability one, the complete works of Shakespeare will appear in his output.

Chapter 4

Random variables

4.1 Basic definitions

Definition 4.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that $X : \Omega \to \mathbb{R}$ is a *random variable* if $\{X \leq t\} := \{\omega : X(\omega) \leq t\} \in \mathcal{F}$ for any $t \in \mathbb{R}$. In other words, $\{X \leq t\}$ should be an event, so that $\mathbb{P}(X \leq t)$ is well-defined.

Definition 4.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. Then the *distribution* of X is the probability measure P_X on $\mathbb{R}$ defined by $P_X(I) = \mathbb{P}(X \in I)$ for any interval $I \subseteq \mathbb{R}$.

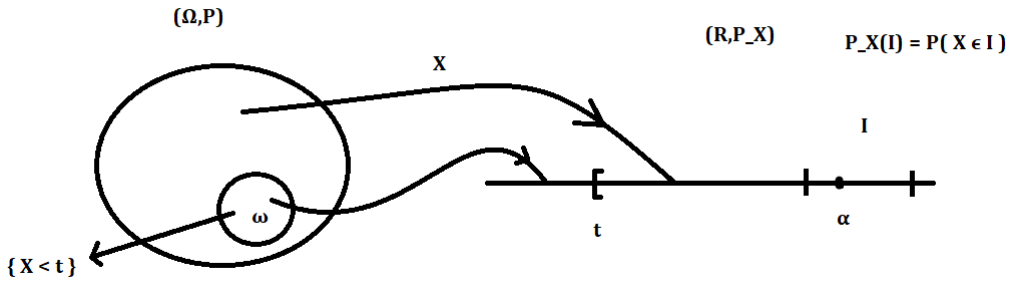


Figure 4.1 – A croquis of the various definitions.

Lemma 4.3. *P_X is a probability measure on $\mathbb{R}$.*

Proof. We have $P_X(\emptyset) = \mathbb{P}(\{\omega : X(\omega) \in \emptyset\}) = \mathbb{P}(\emptyset) = 0$.
$P_X(\mathbb{R}) = \mathbb{P}(X(\omega) \in \mathbb{R}) = 1$.
$0 \leq P_X(I) = \mathbb{P}(X(\omega) \in I) \leq 1$.

If $I_1, I_2, \dots$ are pairwise disjoint then

$$P_X(\cup_k I_k) = \mathbb{P}(X \in \cup_k I_k) = \mathbb{P}(\{X \in I_1\} \cup \{X \in I_2\} \cup \dots) \\ = \sum_k \mathbb{P}(\{X \in I_k\}) = \sum_k P_X(I_k)\,.$$

Here we used that if I_k are pairwise disjoint then the sets $\{X \in I_k\}$ are also pairwise disjoint. This completes the proof. □

Definition 4.4. We say that $X : \Omega \to \mathbb{R}$ is a *discrete random variable* if its range can be written as a finite or infinite sequence, $\operatorname{Ran} X = \{\alpha_1, \alpha_2, \dots\}$.

Example 4.5. Let $\Omega = [0,1]$ and look at the graphs :

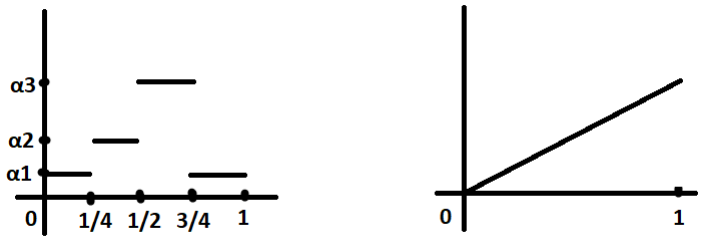


Figure 4.2 – Graph of $X(\omega)$ for two random variables. The one on the left is discrete, the one on the right is not discrete.

Suppose X is a discrete random variable with range $\{\alpha_1, \alpha_2, \dots\}$. Let $I \subseteq \mathbb{R}$. If I contains exactly one α_j, then $P_X(I) = \mathbb{P}(X \in I) = \mathbb{P}(X = \alpha_j)$. In general,

$$P_X(I) = \sum_{\alpha_j \in I} P_X(\{\alpha_j\})\,.$$

Note that if we sum over the whole range, $\sum_j P_X(\{\alpha_j\}) = P_X(\mathbb{R}) = 1$.

Example 4.6. Throw 2 dices and let X be the the total number of spots. Then X is a discrete random variable with range $\{2, 3, \dots, 12\}$. Moreover,

$$P_X([\tfrac{5}{2}, \tfrac{9}{2}]) = P_X(\{3\}) + P_X(\{4\}) = \mathbb{P}(\{(1,2),(2,1)\} + \mathbb{P}(\{(1,3),(3,1),(2,2)\}) \\ = \tfrac{2}{36} + \tfrac{3}{36} = \tfrac{5}{36}\,.$$

Definition 4.7. We say that $X : \Omega \to \mathbb{R}$ is a *continuous random variable*[1] if there exists an integrable nonnegative function p_X such that for any interval $I \subseteq \mathbb{R}$,

$$P_X(I) = \int_I p_X(t)\,\mathrm{d}t = \mathbb{P}(X \in I)\,.$$

In this case, we call p_X the *probability density of* X.

1. The correct terminology is "a random variable with an absolutely continuous distribution". However here we follow the book [3] and summarize this by just saying "continuous random variable".

Remark 4.8. Note that

$$\int_{-\infty}^{\infty} p_X(t)\,\mathrm{d}t = P_X(\mathbb{R}) = 1\,.$$

Also note that if X is a continuous random variable, then

$$P_X(\{\alpha\}) = \int_{\alpha}^{\alpha} p_X(t)\,\mathrm{d}t = 0 \qquad \forall\ \alpha \in \mathbb{R}\,.$$

Definition 4.9. Let X be a random variable. The *distribution function of* X is the function $\Phi_X : \mathbb{R} \to [0,1]$ defined by

$$\Phi_X(t) = \mathbb{P}(X \leq t) = P_X((-\infty, t])\,.$$

If X is a discrete random variable we thus have

$$\Phi_X(t) = \sum_{\alpha_j \leq t} P_X(\{\alpha_j\})\,.$$

If X is a continuous random variable, we have

$$\Phi_X(t) = \int_{-\infty}^{t} p_X(s)\,\mathrm{d}s\,.$$

Note that $P_X(]a,b]) = \Phi_X(b) - \Phi_X(a)$.

Definition 4.10. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then the *joint distribution* of X and Y is the probability measure $P_{X,Y}$ on $\mathbb{R}^2$ defined by

$$P_{X,Y}(I \times J) = \mathbb{P}(X \in I \text{ and } Y \in J)$$

for any intervals $I, J \subseteq \mathbb{R}$.

As before, if X, Y are discrete, $\operatorname{Ran} X = \{\alpha_i\}$ and $\operatorname{Ran} Y = \{\beta_j\}$, then

$$P_{X,Y}(B) = \sum_{(\alpha_i,\beta_j)\in B} P_{X,Y}(\{(\alpha_i,\beta_j)\}) = \sum_{(\alpha_i,\beta_j)\in B} \mathbb{P}(X = \alpha_i,\ Y = \beta_j)$$

for $B \subseteq \mathbb{R}^2$. Similarly, if X, Y are continuous, then

$$P_{X,Y}(B) = \iint_B p_{X,Y}(s,t)\,\mathrm{d}s\mathrm{d}t\,.$$

In this case, $p_{X,Y}$ is a nonnegative integrable function on $\mathbb{R}^2$ which is called the *joint density* of X and Y.

Definition 4.11. We say the random variables X, Y are *independent* if for any intervals I, J, the events $\{X \in I\}$ and $\{Y \in J\}$ are independent.

More generally, $X_1, \ldots, X_n$ are independent if for any intervals $I_1, , \ldots, I_n$, the events $\{X_1 \in I_1\}, \ldots, \{X_n \in I_n\}$ are independent.

An infinite sequence of random variables is called independent if for all n, the sequence $X_1, \ldots, X_n$ is independent.

Lemma 4.12. *If X, Y are independent then $P_{X,Y} = P_X \otimes P_Y$. In particular, if X has density p_X and Y has density p_Y, then (X, Y) have the joint density $p_{X,Y}(s,t) = p_X(s)p_Y(t)$.*

Proof. Let I, J be intervals. Then $P_{X,Y}(I \times J) = \mathbb{P}(X \in I \text{ and } Y \in J) = \mathbb{P}(X \in I)\,\mathbb{P}(Y \in J) = P_X(I)P_Y(J)$. Since this holds for any I, J, we get $P_{X,Y} = P_X \otimes P_Y$ (admitted). □

Definition 4.13. For any set $A \in \mathcal{F}$, we define the *indicator function* $\mathbf{1}_A$ by

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

Definition 4.14. We say that X is *uniformly distributed on* $[a, b]$ if X is a continuous random variable with probability density

$$p_X(t) = \frac{\mathbf{1}_{[a,b]}(t)}{b-a} = \begin{cases} \frac{1}{b-a} & \text{if } a \le t \le b, \\ 0 & \text{if } t < a \text{ or } t > b. \end{cases}$$

Example 4.15. Suppose we choose two points uniformly at random on $[0, L]$, and suppose we choose them independently. What is the probability that the distance between the two points does not exceed ℓ ?

Solution. Choosing a point uniformly at random on $[0, L]$ amounts to considering a uniformly distributed random variable X over $[0, L]$. The value $X(t)$ then represents the chosen point.

By hypothesis the procedure is independent, so if X, Y are the two random variables representing the choices, then the joint density is given by the product

$$p_{X,Y}(s,t) = \frac{\mathbf{1}_{[0,L]}(s)}{L} \cdot \frac{\mathbf{1}_{[0,L]}(t)}{L}.$$

In particular, if $I, J \subseteq [0, L]$ are two intervals then

$$P_{X,Y}(I \times J) = \iint_{I \times J} p_{X,Y}(s,t)\,\mathrm{d}s\mathrm{d}t = \frac{1}{L^2}\int_I 1\,\mathrm{d}s \int_J 1\,\mathrm{d}t = \frac{|I| \cdot |J|}{L^2} = \frac{\text{area } (I \times J)}{L^2}$$

Thus, $P_{X,Y}(B)$ is just the area of B divided by L^2.

Now the event "the distance between the two points does not exceed ℓ" means we must have $(X, Y) \in B$, where $B = \{(x, y) \in [0, L]^2 : |x - y| \le \ell\}$. We thus want to calculate $\mathbb{P}((X, Y) \in B) = P_{X,Y}(B)$.

We draw a figure. Note that the line $y = x$ is in B. More generally, the line $y = x + s$ is in B iff $|s| \le \ell$. So we obtain the following figure :

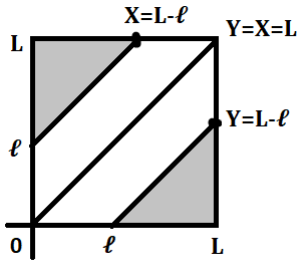


Figure 4.3 – The set B is the white region above.

We thus conclude that

$$P_{X,Y}(B) = \frac{\text{area } (B)}{L^2} = \frac{2L\ell - \ell^2}{L^2},$$

where we calculated the area of B to be the square L^2 minus the two triangles $2\frac{1}{2}(L-\ell)^2$, so the total is $L^2 - (L-\ell)^2 = 2L\ell - \ell^2$.

Example 4.16 (Buffon's needle problem)**.** Suppose a needle of length ℓ is tossed uniformly at random on a floor made of parallel strips of width $L \geq \ell$. What is the probability that the needle will intersect 2 strips ?

Solution. We choose our coordinates so that the strips are parallel to the x axis and the closest parallel line lies above the needle.

Let X be the angle between the needle and the x axis and Y be the vertical distance from the horizontal line to the end of the needle below the line.

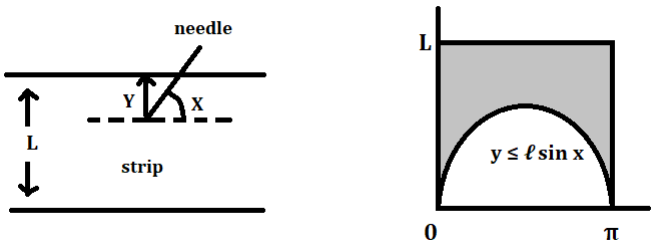


Figure 4.4 – The set B is the white region on the right.

Then in general X can take any value in $[0, \pi]$ and Y any value in $[0, L]$. Moreover, by elementary geometry, an intersection will occur iff $Y \leq \ell \sin X$. So we consider the event $B = \{Y \leq \ell \sin X\}$ and try to find $\mathbb{P}((X,Y) \in B) = P_{X,Y}(B)$.

Since the needle is tossed uniformly at random, and the coordinates X, Y are independent, then the joint distribution over $\mathbb{R}^2$ is given by $p_{X,Y}(x,y) = \frac{\mathbf{1}_{[0,\pi]}(x)}{\pi} \cdot \frac{\mathbf{1}_{[0,L]}(y)}{L}$.

As in the previous example, this implies

$$P_{X,Y}(B) = \frac{\text{area } (B)}{\pi L} = \frac{\int_0^\pi \ell \sin x \, \mathrm{d}x}{\pi L} = \frac{\ell(-\cos\pi + \cos 0)}{\pi L} = \frac{2\ell}{\pi L}\,.$$

Theorem 4.17. *Suppose X, Y are independent continuous random variables. Then the probability density of $Z = X + Y$ is given by the convolution of p_X and p_Y. In other words,*

$$p_Z(x) = \int_{-\infty}^{\infty} p_X(x-t)p_Y(t)\,\mathrm{d}t\,.$$

Proof. Since X, Y are independent, we have $p_{X,Y}(s,t) = p_X(s)p_Y(t)$. Hence, if $I = [a,b]$ is any interval and $B = \{(s,t) \in \mathbb{R}^2 : a \le s+t \le b\}$, then

$$P_Z(I) = \mathbb{P}(a \le X+Y \le b) = \iint_B p_X(s)p_Y(t)\,\mathrm{d}s\mathrm{d}t\,.$$

Now notice that $(s,t) \in B \iff s = x - t$, with $a \le x \le b$. This holds for any $-\infty < t < \infty$. Consequently,

$$P_Z(I) = \int_a^b \int_{-\infty}^{\infty} p_X(x-t)p_Y(t)\,\mathrm{d}t\,\mathrm{d}x$$

But $P_Z(I) = \int_a^b p_Z(x)\,\mathrm{d}x$. We deduce that $p_Z(x) = \int_{-\infty}^{\infty} p_X(x-t)p_Y(t)\,\mathrm{d}t$. □

Example 4.18. Suppose X, Y are two independent uniformly distributed random variables over $[0,1]$. Find the probability density of $Z = X + Y$.

Solution. Since each random variable takes values in $[0,1]$, then Z takes values in $[0,2]$. So we know without working that $p_Z(x) = 0$ if $x \notin [0,2]$.

Now suppose $x \in [0,2]$. We apply the convolution theorem. Here $p_X = \mathbf{1}_{[0,1]} = p_Y$. Consequently

$$\begin{aligned} p_Z(x) = \int_{-\infty}^{\infty} \mathbf{1}_{[0,1]}(x-t)\mathbf{1}_{[0,1]}(t)\,\mathrm{d}t &= \int_0^1 \mathbf{1}_{[0,1]}(x-t)\,\mathrm{d}t \\ &= \int_{x-1}^{x} \mathbf{1}_{[0,1]}(s)\,\mathrm{d}s = \int_{-\infty}^{\infty} \mathbf{1}_{[0,1]\cap[x-1,x]}(s)\,\mathrm{d}s\,. \end{aligned}$$

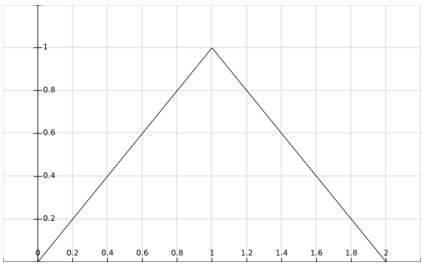
If $x \in [0,1]$ then $[0,1] \cap [x-1,x] = [0,x]$. So $p_Z(x) = \int_{-\infty}^{\infty} \mathbf{1}_{[0,x]}(s)\,\mathrm{d}s = \int_0^x \mathrm{d}s = x$.

If $x \in [1,2]$, then $[0,1] \cap [x-1,x] = [x-1,1]$.

So $p_Z(x) = \int_{-\infty}^{\infty} \mathbf{1}_{[x-1,1]}(s)\,\mathrm{d}s = \int_{x-1}^{1} \mathrm{d}s = 2 - x$. Thus,

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

The graph of p_Z is shown below :



4.2 Mathematical Expectation

Definition 4.19. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. Suppose $\int_\Omega |X(\omega)| \, \mathrm{d}\,\mathbb{P}(\omega) < \infty$. Then we define the *expectation* or *mean* of X by

$$\mathbb{E}(X) = \int_\Omega X(\omega) \, \mathrm{d}\,\mathbb{P}(\omega) \, .$$

Theorem 4.20 (Change of variables formula)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable. Then for any "nice" $\varphi : \mathbb{R} \to \mathbb{R}$, we have*[2]

$$\mathbb{E}(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(t) \, \mathrm{d}P_X(t) = \begin{cases} \sum_j \varphi(\alpha_j) P_X(\{\alpha_j\}) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} \varphi(t) p_X(t) \, \mathrm{d}t & \text{if } X \text{ is continuous.} \end{cases}$$

Similarly, if X, Y are random variables with joint distribution $P_{X,Y}$, then for any "nice" $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$, we have

$$\begin{aligned} \mathbb{E}(\varphi(X,Y)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(s,t) \, \mathrm{d}P_{X,Y}(s,t) \\ &= \begin{cases} \sum_{j,k} \varphi(\alpha_j, \beta_k) P_{X,Y}(\{\alpha_j, \beta_k\}) & \text{if } X,Y \text{ are discrete,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(s,t) p_{X,Y}(s,t) \, \mathrm{d}s\mathrm{d}t & \text{if } X,Y \text{ are continuous.} \end{cases} \end{aligned}$$

Proof. Admitted. □

Example 4.21. If X is discrete with range $\{\alpha_j\}$, then $\mathbb{E}(X^n) = \sum_j \alpha_j^n P_X(\{\alpha_j\})$.

If X is continuous, then $\mathbb{E}(X^n) = \int_{-\infty}^{\infty} t^n p_X(t) \, \mathrm{d}t$.

Example 4.22. Throw 2 dices at random and let X be the random variable giving the total number of spots on both dices. Compute $\mathbb{E}(X)$.

Solution. We have $\Omega = \{(x,y) : x, y \in \{1, \ldots, 6\}\}$. So $\mathbb{P}(\{(x,y)\}) = \frac{1}{36}$. So

$$\begin{aligned} \mathbb{E}(X) = \int_\Omega X(\omega) \, \mathrm{d}\,\mathbb{P}(\omega) &= \sum_{\omega \in \Omega} X(\omega) \, \mathbb{P}(\omega) = \frac{1}{36} \sum_{(x,y) \in \Omega} (x+y) \\ &= \frac{1}{36}((2+3+\cdots+7) + (3+4+\cdots+8) + \cdots + (7+8+\cdots+12)) \\ &= \frac{27+33+39+45+51+57}{36} = \frac{252}{36} = 7 \, . \end{aligned}$$

2. For example any piecewise continuous φ is nice, but it can be more general.

Solution 2. We use the change of variables formula. Here $\operatorname{Ran} X = \{2, \dots, 12\}$. $P_X(\{2\}) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36}$, $P_X(\{3\}) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{2}{36}$, $P_X(\{4\}) = \mathbb{P}(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36}$, $P_X(\{5\}) = \mathbb{P}(\{(1,4),(2,3),(3,2),(4,1)\} = \frac{4}{36}$, $P_X(\{6\}) = \mathbb{P}((1,5),(2,4),(3,3),(4,2),(5,1)\} = \frac{5}{36}$. Continuing, $P_X(\{7\}) = \mathbb{P}((1,6),(2,5),(3,4),(4,3),(5,2),(6,1)) = \frac{6}{36}$. In the same way, $P_X(\{8\}) = \mathbb{P}((2,6),(3,5),(4,4),(5,3),(6,2)) = \frac{5}{36}$, $P_X(\{9\}) = \mathbb{P}((3,6),(4,5),(5,4),(6,3)) = \frac{4}{36}$, $P_X(\{10\}) = \mathbb{P}((4,6),(5,5),(6,4)) = \frac{3}{36}$, $P_X(\{11\}) = \mathbb{P}((5,6),(6,5)) = \frac{2}{36}$, $P_X(\{12\}) = \mathbb{P}((6,6)) = \frac{1}{36}$.

Consequently,

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=2}^{12} k P_X(\{k\}) \\ &= \frac{2\cdot 1 + 3\cdot 2 + 4\cdot 3 + 5\cdot 4 + 6\cdot 5 + 7\cdot 6 + 8\cdot 5 + 9\cdot 4 + 10\cdot 3 + 11\cdot 2 + 12\cdot 1}{36} \\ &= \frac{2+6+12+20+30+42+40+36+30+22+12}{36} = 7\,.\end{aligned}$$

Of course the second solution is more complicated here. The change of variables method is useful when we know the distribution P_X in advance.

Example 4.23. Suppose X is uniformly distributed over $[a, b]$. Find $\mathbb{E}(X)$.

Solution. Using the change of variables formula,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t\, p_X(t)\, \mathrm{d}t = \frac{1}{b-a}\int_a^b t\, \mathrm{d}t = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}\,.$$

Lemma 4.24. *The following properties hold :*

a) $\mathbb{E}(1) = 1$,

b) $\mathbb{E}(cX) = c\,\mathbb{E}(X)$ *for any constant* c,

c) $|\mathbb{E}(X)| \le \mathbb{E}(|X|)$,

d) $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$,

e) *If* $X \ge 0$, *then* $\mathbb{E}(X) \ge 0$. *Similarly, if* $X \le Y$, *then* $\mathbb{E}(X) \le \mathbb{E}(Y)$,

f) *If* X, Y *are independent random variables, then* $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$,

g) $\mathbb{P}(X \in I) = \mathbb{E}(\mathbf{1}_{\{X\in I\}})$

Proof. All these properties follow from the construction of the integral, except e), which follows from the change of variables formula. We omit the proofs. □

Remark 4.25. In all the above, we defined a random variable as a map $X : \Omega \to \mathbb{R}$. We can generalize this and define the concepts of "complex random variable" as a map $X : \Omega \to \mathbb{C}$, and the concept of a "random vector" as a map $X : \Omega \to \mathbb{R}^n$ and so on.

4.3 Chebyshev's inequality, the variance and the correlation coefficient

Theorem 4.26 (Chebyshev's inequality)**.** *Let X be a random variable. Then*

$$\mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon^2}\,\mathbb{E}(X^2)\,.$$

Proof. We have $|X| > \varepsilon \iff X^2 > \varepsilon^2$. Hence,

$$\mathbb{P}(|X| > \varepsilon) = \mathbb{E}\left(\mathbf{1}_{\{|X|>\varepsilon\}}\right) = \mathbb{E}\left(\mathbf{1}_{\{\frac{X^2}{\varepsilon^2}>1\}}\right) \leq \mathbb{E}\left(\frac{X^2}{\varepsilon^2}\mathbf{1}_{\{\frac{X^2}{\varepsilon^2}>1\}}\right) \leq \frac{1}{\varepsilon^2}\,\mathbb{E}(X^2)\,. \quad \square$$

It is clear from the proof that we also have $\mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon^\alpha}\,\mathbb{E}(|X|^\alpha)$ for any $\alpha > 0$.

Corollary 4.27. *If $\mathbb{E}(X^2) = 0$, then with probability one, $X = 0$.*

Proof. We have $\{X = 0\} = \cap_n \{|X| \leq \frac{1}{n}\}$ and the sets are decreasing. By a theorem, we deduce that $\mathbb{P}(X = 0) = \lim_{n\to\infty} \mathbb{P}(|X| \leq \frac{1}{n})$. But $\mathbb{P}(|X| > \frac{1}{n}) \leq n^2\,\mathbb{E}(X^2) = 0$ for any n, by Chebyshev's inequality. Hence, $\mathbb{P}(X = 0) = 1$ as required. $\square$

Definition 4.28. Let X be a random variable with mean a, i.e. $\mathbb{E}(X) = a$. We define the *variance* of X by

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

Lemma 4.29. *The following properties hold :*

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

b) $\mathrm{Var}(1) = 0$,

c) $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$,

d) *If X, Y are independent, then*

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

Proof. a) Let $\mathbb{E}(X) = a$. We have

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}((X - a)^2) = \mathbb{E}(X^2 - 2aX + a^2) = \mathbb{E}(X^2) - 2a\,\mathbb{E}(X) + a^2\,\mathbb{E}(1)\\ &= \mathbb{E}(X^2) - 2a^2 + a^2 = \mathbb{E}(X^2) - a^2\,.\end{aligned}$$

b) We have $\mathbb{E}(1) = 1$. So $\mathrm{Var}(1) = \mathbb{E}((1 - 1)^2) = 0$.

c) Let $\mathbb{E}(X) = a$. We have $\mathbb{E}(cX) = c\,\mathbb{E}(X) = ca$.
So $\mathrm{Var}(cX) = \mathbb{E}((cX - ca)^2) = c^2\,\mathbb{E}((X-a)^2) = c^2\,\mathrm{Var}(X)$.
d) Let $a = \mathbb{E}(X)$, $b = \mathbb{E}(Y)$. Then $\mathbb{E}(X+Y) = a + b$. So

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}((X+Y-a-b)^2) = \mathbb{E}((X-a)^2 + (Y-b)^2 + 2(X-a)(Y-b)) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathbb{E}((X-a)(Y-b))\end{aligned}$$

Now X, Y are independent, so

$$\mathbb{E}((X-a)(Y-b)) = \mathbb{E}(X)\,\mathbb{E}(Y) - a\,\mathbb{E}(Y) - b\,\mathbb{E}(X) + ab = 0\,. \tag{4.1}$$

So d) follows. □

Definition 4.30. We say that a sequence of random variables $X_1, X_2, \ldots$ are i.i.d. if they are independent and identically distributed, i.e. $P_{X_j} = P_{X_1}$ for all j.

Lemma 4.31. *If X, Y are i.i.d. then $\mathbb{E}(X) = \mathbb{E}(Y)$ and $\mathrm{Var}(X) = \mathrm{Var}(Y)$.*

Proof. This follows from the change of variables formula :

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t\,\mathrm{d}P_X(t) = \int_{-\infty}^{\infty} t\,\mathrm{d}P_Y(t) = \mathbb{E}(Y)\,,$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \int_{-\infty}^{\infty} t^2\,\mathrm{d}P_X(t) - (\mathbb{E}(Y))^2 = \int_{-\infty}^{\infty} t^2\,\mathrm{d}P_Y(t) - (\mathbb{E}(Y))^2 \\ &= \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \mathrm{Var}(Y)\,.\end{aligned}$$ □

Definition 4.32. Let X_1, X_2 be random variables. Let $a_1 = \mathbb{E}(X_1)$, $a_2 = \mathbb{E}(X_2)$, $\sigma_1 = \sqrt{\mathrm{Var}(X_1)}$, $\sigma_2 = \sqrt{\mathrm{Var}(X_2)}$. We define the *correlation coefficient* of X_1 and X_2 by

$$r = \frac{\mathbb{E}((X_1 - a_1)(X_2 - a_2))}{\sigma_1 \sigma_2}\,.$$

Lemma 4.33. *The following properties hold :*

a) *If X_1, X_2 are independent, then their correlation coefficient $r = 0$.*

b) *The converse of* a) *is not true.*

c) *We always have $|r| \leq 1$, so $r \in [-1, 1]$.*

Proof. a) See the calculation in (4.1).
b) This is Exercise 19.
c) This follows from the *Cauchy-Schwarz inequality*, we omit the proof. □

Remark 4.34 (Outside curriculum)**.** Let X, Y be random variables. Consider the problem of having the *best mean square approximation of* X using a combination $cY + d$. In other words, we want to find c, d such that

$$\mathbb{E}((X - cY - d)^2) = \min_{e,f \in \mathbb{R}} \mathbb{E}((X - eY - f)^2)\,.$$

Then it can be shown that we should take

$$cY + d = a_1 + r\frac{\sigma_1}{\sigma_2}(Y - a_2)\,.$$

The student can find a proof in the book [3, p.49-50].

4.4 Exercises

1. A motorist encounters four consecutive traffic lights, each equally likely to be red or green. Let X be the number of green lights passed by the motorist before being stopped by a red light. What is the probability distribution of X ?
2. Give an example of two distinct random variables with the same distribution function.
3. Find the distribution function of a random variable X uniformly distributed over $[a, b]$.
4. A random variable X has probability density

$$p_X(x) = \frac{a}{x^2 + 1} \qquad (-\infty < x < \infty)\,.$$

   Find

   a) The constant a;

   b) The distribution function of X;

   c) The probability $\mathbb{P}(-1 \leq X \leq 1)$.

5. A random variable X has probability density

$$p_X(x) = \begin{cases} ax^2 \mathrm{e}^{-kx} & \text{if } 0 \leq x < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

   ($k > 0$). Find

   a) The constant a;

b) The distribution function of X;

c) The probability $\mathbb{P}(0 \leq X \leq 1/k)$.

6. A random variable X has distribution function

$$\Phi_X(x) = a + b \arctan \frac{x}{2} \qquad (-\infty < x < \infty)\,.$$

Find

a) The constants a and b;

b) The probability density of X.

7. Two nonoverlapping circular disks of radius r are painted on a circular table of radius R. A point is then "tossed at random" onto the table. What is the probability of the point falling in one of the disks ?

8. What is the probability that two randomly chosen numbers between 0 and 1 will have a sum no greater than 1 and a product no greater than $\frac{2}{9}$?

9. Given two independent random variables X and Y with probability densities

$$p_X(x) = \begin{cases} \frac{1}{2}\mathrm{e}^{-x/2} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \qquad p_Y(x) = \begin{cases} \frac{1}{3}\mathrm{e}^{-x/3} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

find the probability density of the random variable $Z = X + Y$.

10. Given three independent random variables X_1, X_2 and X_3, each uniformly distributed in the interval $[0, 1]$, find the probability density of the random variable $X_1 + X_2 + X_3$.

Hint : see Example 4.18.

11. A random variable X takes the values $1, 2, \ldots$ with probabilities

$$\frac{1}{3}, \frac{1}{3^2}, \cdots, \frac{1}{3^k}, \cdots$$

and the value 0 with probability $\frac{1}{2}$. Find $\mathbb{E}(X)$.

12. Balls are drawn from an urn containing w white balls and b black balls until a white ball appears. Find the mean value m and variance σ^2 of the number of black balls drawn, assuming that each ball is replaced after being drawn.

13. Find the mean and variance of the random variable X with probability density

$$p_X(x) = \frac{1}{2}\mathrm{e}^{-|x|} \qquad (-\infty < x < \infty)\,.$$

14. Find the mean and variance of the random variable X with probability density

$$p_X(x) = \begin{cases} \frac{1}{2b} & \text{if } |x-a| \le b, \\ 0 & \text{otherwise.} \end{cases}$$

15. The distribution function of a random variable X is

$$\Phi_X(x) = \begin{cases} 0 & \text{if } x \le -1, \\ a + b\arcsin(x) & \text{if } -1 \le x \le 1, \\ 1 & \text{if } x \ge 1. \end{cases}$$

Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

16. Let X be the number of spots obtained in throwing an unbiased die. Find the mean and variance of X.

17. In the preceding problem, what is the probability P of X deviating from $\mathbb{E}(X)$ by more than $\frac{5}{2}$? Show that Chebyshev's inequality gives only a very crude estimate of P.

18. Prove that if X is a random variable such that $\mathbb{E}(e^{aX})$ exists, where a is a positive constant, then

$$\mathbb{P}(X > \varepsilon) \le \frac{\mathbb{E}(e^{aX})}{e^{a\varepsilon}}.$$

Hint : Apply Chebyshev's inequality to the random variable $Y = e^{aX/2}$.

19. Let X be a random variable taking each of the values $-2, -1, 1$ and 2 with probability $\frac{1}{4}$, and let $Y = X^2$. Prove that X and Y (although obviously dependent) have correlation coefficient 0.

20. Find the means and variances of two random variables X and Y with joint probability density

$$p_{X,Y}(x,y) = \begin{cases} \sin(x)\sin(y) & \text{if } 0 \le x \le \frac{\pi}{2},\ 0 \le y \le \frac{\pi}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

What is the correlation coefficient of X and Y ?

21. Find the correlation coefficient r of two random variables X and Y with joint probability density

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{2}\sin(x+y) & \text{if } 0 \le x \le \frac{\pi}{2},\ 0 \le y \le \frac{\pi}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

22. Given a random variable X, let $\varphi(t)$ be a nondecreasing positive function such that $\mathbb{E}(\varphi(X))$ exists. Prove that

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(\varphi(X))}{\varphi(t)} \tag{4.1}$$

23. Deduce Chebyshev's inequality as a special case of (4.1).

24. Let X be a random variable with probability density

$$p_X(x) = \frac{1}{\pi(1+x^2)} \qquad (-\infty < x < \infty)\,.$$

Show that $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ fail to exist.

Chapter 5

Important distributions

Definition 5.1. A sequence of Bernoulli trials is a sequence of identical independent experiments in which an event A occurs with probability $\mathbb{P}(A) =: p$ and fails to occur with probability $1 - p =: q$. The occurrence of A is called "success", the non-occurrence is called a "failure".

Example 5.2. Toss a coin n times and consider the event $A = \{\text{The result is a head}\}$. Then this gives a sequence of Bernoulli trials with $p = q = \frac{1}{2}$.

Question : In tossing 100 coins, what is the probability of getting 37 heads ?
Answer. Let B be the event "37 head appeared".

1. Each sequence $\omega = HTHHT\cdots T$ of length 100 containing 37 heads has probability $\frac{1}{2^{37}} \cdot \frac{1}{2^{63}} = \frac{1}{2^{100}}$ to occur.
2. The number of sequences of length 100 containing 37 heads is the number of ways to choose 37 digits in a 100-tuple, put H on them, regardless of order wihin these 37 digits. Hence the number of ways is C_{37}^{100}.
3. Conclusion $\mathbb{P}(B) = C_{37}^{100} \cdot \frac{1}{2^{100}}$.

In general, if X is the total number of successes in a sequence of n Bernoulli trials, then the same argument shows that

$$\mathbb{P}(X = k) = C_k^n p^k q^{n-k}$$

for $k = 0, 1, \ldots, n$. In other words, X has the *binomial distribution* with parameter n :

$$P_X(k) = \begin{cases} C_k^n p^k q^{n-k} & \text{if } k = 0, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\sum_{k=0}^{n} P_X(k) = 1$ as required. Indeed, $\sum_{k=0}^{n} C_k^n p^k q^{n-k} = (p+q)^n = 1^n = 1$ by the binomial theorem.

Lemma 5.3. *If X has the binomial distribution with parameter n, then*

$$\mathbb{E}(X) = np \quad \textit{and} \quad \mathrm{Var}(X) = npq\,.$$

Proof. Define

$$X_k = \begin{cases} 1 & \text{if k-th trial is a success,} \\ 0 & \text{if k-th trial is a failure.} \end{cases}$$

Then $X = \sum_{k=1}^{n} X_k$. Indeed, $\sum_{k=1}^{n} X_k$ is precisely the number of successes in the n trials. Consequently $\mathbb{E}(X) = \sum_{k=1}^{n} \mathbb{E}(X_k)$. Since the trials are independent, the X_k are independent, so $\mathrm{Var}(X) = \sum_{k=1}^{n} \mathrm{Var}(X_k)$.

But

$$\mathbb{E}(X_k) = 0 \cdot \mathbb{P}(X_k = 0) + 1 \cdot \mathbb{P}(X_k = 1) = p\,.$$

Also,

$$\begin{aligned} \mathrm{Var}(X_k) &= \mathbb{E}(X_k^2) - (\mathbb{E}(X_k))^2 = 0^2 \cdot \mathbb{P}(X_k = 0) + 1^2 \cdot \mathbb{P}(X_k = 1) - p^2 \\ &= p - p^2 = p(1-p) = pq \end{aligned}$$

Thus, the result follows. □

We now come to our second discrete distribution :

Definition 5.4. We say that X has the *Poisson distribution* with parameter a if

$$P_X(k) = \frac{a^k}{k!}\mathrm{e}^{-a}\,, \qquad k = 0, 1, 2, \ldots$$

Note that $\sum\limits_{k=0}^{\infty} P_X(k) = 1$ as required. In fact, $\sum\limits_{k=0}^{\infty} \frac{a^k}{k!}\mathrm{e}^{-a} = \mathrm{e}^{-a} \sum\limits_{k=0}^{\infty} \frac{a^k}{k!} = \mathrm{e}^{-a}\mathrm{e}^{a} =$ 1 by the Taylor expansion of e^x.

Lemma 5.5. *If X has the Poisson distribution with parameter a, then*

$$\mathbb{E}(X) = a\,.$$

Proof. By the change of variables formula,

$$\begin{aligned} \mathbb{E}(X) = \sum_{k=0}^{\infty} k P_X(k) = \sum_{k=0}^{\infty} k \frac{a^k}{k!}\mathrm{e}^{-a} = \mathrm{e}^{-a} \sum_{k=1}^{\infty} \frac{a^k}{(k-1)!} = \mathrm{e}^{-a}\Big(\frac{a}{0!} + \frac{a^2}{1!} + \ldots\Big) \\ = \mathrm{e}^{-a} \sum_{k=0}^{\infty} \frac{a^{k+1}}{k!} = a\mathrm{e}^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} = a\mathrm{e}^{-a}\mathrm{e}^{a} = a\,. \quad \square \end{aligned}$$

The student will calculate the variance for the Poisson distribution in Exercise 9.

Theorem 5.6 (Poisson limit theorem)**.** *The binomial distribution approaches the Poisson distribution when the number of trials gets large and the success probability gets small. More precisely, if* $p_n = \dfrac{a}{n}$ *and* $q_n = 1 - p_n$ *then*

$$\lim_{n\to\infty} C_k^n p_n^k q_n^{n-k} = \mathrm{e}^{-a}\frac{a^k}{k!}\,.$$

Proof. Let $a_k(n) = C_k^n p_n^k q_n^{n-k}$. Then

$$\lim_{n\to\infty} a_0(n) = \lim_{n\to\infty} q_n^n = \lim_{n\to\infty}(1-p_n)^n = \lim_{n\to\infty}\left(1-\frac{a}{n}\right)^n = \mathrm{e}^{-a}\,.$$

Also, for $k \geq 1$,

$$\frac{a_k(n)}{a_{k-1}(n)} = \frac{(k-1)!(n-k+1)!}{k!(n-k)!}\cdot\frac{p_n}{q_n} = \frac{n-k+1}{k}\cdot\frac{p_n}{q_n} = \frac{n-k+1}{n-a}\cdot\frac{a}{k}$$

So

$$\lim_{n\to\infty}\frac{a_k(n)}{a_{k-1}(n)} = \frac{a}{k}\,.$$

It follows that $\lim\limits_{n\to\infty} a_1(n) = \lim\limits_{n\to\infty}\frac{a_1(n)}{a_0(n)}a_0(n) = \frac{a}{1}\mathrm{e}^{-a}$.
Similarly $\lim\limits_{n\to\infty} a_2(n) = \lim\limits_{n\to\infty}\frac{a_2(n)}{a_1(n)}a_1(n) = \frac{a}{2}\cdot\frac{a}{1}\mathrm{e}^{-a} = \frac{a^2}{2!}\mathrm{e}^{-a}$.
By induction we see that $\lim\limits_{n\to\infty} a_k(n) = \frac{a^k}{k!}\mathrm{e}^{-a}$. □

Remark 5.7. 1. The theorem remains true if we just assume $\lim\limits_{n\to\infty} np_n = a$ instead of $p_n = \frac{a}{n}$. The proof requires a little more manipulation with the limit.

2. An important consequence of Theorem 5.6 is that we may approximate the Binomial distribution $C_k^n p^k q^{n-k}$ by the Poisson distribution $\frac{a^k}{k!}\mathrm{e}^{-a}$ by choosing $a = np$. The approximation will be accurate if n is large enough.

Example 5.8. Suppose N raisin buns of equal size are baked from a batch of dough into which n raisins have been carefully mixed. What is the probability that any given bun contains at least one raisin ?

Solution. Fix the given bun and consider the n raisins successively. Then we may view this as a series of Bernoulli trials, where "success" in the k-th trial means the k-th raisin ends up in the given bun. The success probability is then $p = 1/N$, since each raisin chooses a bun uniformly at random.

Let A be the event "the fixed bun has at least one raisin". Then A^c is the event "the given bun has no raisin", so $\mathbb{P}(A^c) = P_X(0)$.

The exact probability is thus given by the Binomial distribution :

$$\mathbb{P}(A) = 1 - P_X(0) = 1 - C_0^n p^0 q^{n-0} = 1 - \left(1-\frac{1}{N}\right)^n.$$

For example if $N = 5$ and $n = 100$ then $(1 - \frac{1}{N})^n \approx 2.037e^{-10}$, so $\mathbb{P}(A) \approx 1$.

If we make the Poisson approximation here with $a = np = \frac{n}{N}$ we get

$$1 - P_X(0) = 1 - \frac{a^0}{0!}e^{-a} = 1 - e^{-n/N}.$$

If $N = 5$ and $n = 100$ then $e^{-n/N} = e^{-20}$, so we still have $1 - e^{-n/N} \approx 1$. This shows the Poisson approximation is valid for this problem.

So far we only discussed discrete distributions in this chapter. We now introduce :

Definition 5.9. We say that a random variable X has the *normal* or *Gaussian* distribution if it is continuous with probability density

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

Note that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = 1$ as required. This is a famous integral usually proved using polar coordinates.

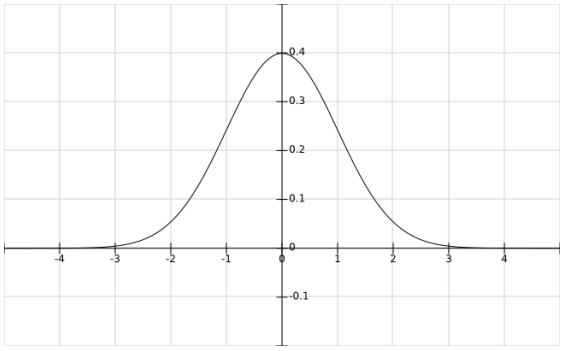


Figure 5.1 – The famous bell shape of the density $p_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

Lemma 5.10. *If X has the normal distribution then*

$$\mathbb{E}(X) = 0 \qquad \textit{and} \qquad \mathrm{Var}(X) = 1.$$

Proof. We have $\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}\,dx = 0$ because we integrate an odd function.

Next, integrating by parts,

$$\begin{aligned}\mathbb{E}(X^2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2e^{-x^2/2}\,dx &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} -x\mathrm{d}(e^{-x^2/2}) \\ &= \frac{1}{\sqrt{2\pi}}\Big[(-xe^{-x^2/2}|_{-\infty}^{\infty}) + \int_{-\infty}^{\infty} e^{-x^2/2}\,dx\Big] = 1,\end{aligned}$$

where we used $\lim_{x\to\pm\infty} xe^{-x^2/2} = 0$ and $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = 1$.

It follows that $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1 - 0 = 1$. □

The normal distribution is important in statistics and social science, partly because of the *central limit theorem* which we discuss in more detail in Chapter 6. For now we present the following special case :

Theorem 5.11 (De Moivre-Laplace)**.** *Consider n i.i.d random variables $X_1, \dots, X_n$ where $X_k = 1$ with probability p and $X_k = 0$ with probability $q = 1 - p$. Define*

$$S_n = \sum_{k=1}^{n} X_k \qquad \text{and} \qquad S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - np}{\sqrt{npq}}\,.$$

Then

$$\lim_{n\to\infty} \mathbb{P}(a \le S_n^* \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x\,.$$

Proof. Admitted (special case of the theorem in Chapter 6). □

To give an application of this, we should first clarify how to estimate the right-hand side, namely $\frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x$. Unfortunately we cannot compute this integral exactly. However, note that if X is a normal random variable, then

$$\frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x = P_X(]a,b]) = \Phi_X(b) - \Phi_X(a)\,,$$

where $\Phi_X(x)$ is the distribution function. So it suffices to be able to compute Φ_X for the normal distribution. And we have a table for this; see the last page of this chapter.

Note that the table does not cover $\Phi(x)$ when x is negative. For this we use the rule

$$\Phi(-x) = 1 - \Phi(x)\,, \qquad x \ge 0\,.$$

Indeed, since $p(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ is even, we have for $x \ge 0$,

$$\Phi(-x) = \int_{-\infty}^{-x} p(t)\,\mathrm{d}t = \int_x^{\infty} p(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} p(t)\,\mathrm{d}t - \int_{-\infty}^{x} p(t)\,\mathrm{d}t = 1 - \Phi(x)\,.$$

We may now give the following application of the De Moivre-Laplace theorem :

Corollary 5.12. *The distribution of the relative frequency is approximately normal. More precisely, if we perform a sequence of n independent identical experiments and let $n(A)$ be the number of times that some event A occurred, then if n is large,*

$$\mathbb{P}\left(a \le \frac{n(A)}{n} \le b\right) \approx \Phi\left(\frac{nb - np}{\sqrt{npq}}\right) - \Phi\left(\frac{na - np}{\sqrt{npq}}\right),$$

where $p = \mathbb{P}(A)$ and $q = 1 - p$.

Proof. Consider a sequence of i.i.d. random variables $X_1, \ldots, X_n$ such that $X_k = 1$ if A occurs and $X_k = 0$ otherwise. Then $\sum_{k=1}^n X_k$ is exactly the number of occurrences of A in the n trials, i.e. $S_n = \sum_{k=1}^n X_k = n(A)$. Thus,

$$\mathbb{P}\left(a \leq \frac{n(A)}{n} \leq b\right) = \mathbb{P}\left(na \leq S_n \leq nb\right) = \mathbb{P}\left(\frac{na-np}{\sqrt{npq}} \leq \frac{S_n - np}{\sqrt{npq}} \leq \frac{nb-np}{\sqrt{npq}}\right)$$
$$= \mathbb{P}\left(\frac{na-np}{\sqrt{npq}} \leq S_n^* \leq \frac{nb-np}{\sqrt{npq}}\right) \approx \Phi\left(\frac{nb-np}{\sqrt{npq}}\right) - \Phi\left(\frac{na-np}{\sqrt{npq}}\right)$$

where we used the De Moivre-Laplace theorem in the last step. □

Remark 5.13. More generally, a random variable X with probability density

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}} \tag{5.1}$$

is also called a normal random variable. In this case $\mathbb{E}(X) = a$ and $\mathrm{Var}(X) = \sigma^2$ (prove this using a change of variables and Lemma 5.10).

5.1 Exercises

1. Consider the game of "heads or tails," as in Example 3.7. Show that the probability of correctly calling the side of the coin landing upward is always $\frac{1}{2}$ regardless of the call, provided the coin is unbiased. However, show that if the coin is biased, then "heads" should be called all the time if heads are more likely, while "tails" should be called all the time if tails are more likely.
2. There are 10 children in a given family. Assuming that a boy is as likely to be born as a girl, what is the probability of the family having
   a) 5 boys and 5 girls; b) From 3 to 7 boys ?
3. Suppose the probability of hitting a target with a single shot is 0.001. What is the probability P of hitting the target 2 or more times in 5000 shots ?
4. The page proof of a 500-page book contains 500 misprints. What is the probability P of 2 or more misprints appearing on the same page ?
5. Let p be the probability of success in a series of Bernoulli trials. What is the probability P_n of an even number of successes in n trials ?
6. What is the probability of the pattern SFS appearing infinitely often in an infinite series of Bernoulli trials, if S denotes "success" and F "failure" ?
7. An electronic computer contains 1000 transistors. Suppose each transistor has probability 0.001 of failing in the course of a year of operation. What is the probability of at least 3 transistors failing in a year ?

8. A school has 730 students. What is the probability that exactly 4 students were born on January 1 ? Neglect leap years.

9. Let X be a random variable with the Poisson distribution with parameter a. Find
$$\sigma^2 = \mathrm{Var}(X) \qquad \text{and} \qquad \frac{\mathbb{E}(X-a)^3}{\sigma^3}.$$

10. Read the proof of the De Moivre-Laplace Theorem in [3, Theorem 5.1]. Where is the uniform convergence used in the proof ? [1]

11. The probability of occurrence of an event A in one trial is 0.3. What is the probability P that the relative frequency of A in 100 independent trials will lie between 0.2 and 0.4 ?

12. Suppose an event A has probability 0.4. How many trials must be performed to assert with probability 0.9 that the relative frequency of A differs from 0.4 by no more than 0.1 ?

13. The probability of occurrence of an event A in one trial is 0.6. What is the probability P that A occurs in the majority of 60 trials ?

14. Two continuous random variables X_1 and X_2 are said to have a *bivariate normal distribution* if their joint probability density is
$$\begin{aligned} &p_{X_1,X_2}(x_1,x_2) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left\{\frac{-1}{2(1-r^2)}\left[\frac{(x_1-a)^2}{\sigma_1^2} - 2r\frac{(x_1-a)(x_2-b)}{\sigma_1\sigma_2} + \frac{(x_2-b)^2}{\sigma_2^2}\right]\right\} \end{aligned}$$
where $\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < r < 1$. Prove that each of the random variables X_1 and X_2 has a univariate (i.e., one-dimensional) normal distribution of the form (5.1), where $\mathbb{E}\, X_1 = a$, $\mathrm{Var}\, X_1 = \sigma_1^2$, $\mathbb{E}\, X_2 = b$, $\mathrm{Var}\, X_2 = \sigma_2^2$.

15. Prove that the number r in Problem 14 is the correlation coefficient of the random variables X_1 and X_2. Prove that X_1 and X_2 are independent if and only if $r = 0$.

    Comment. This is a situation in which r is a satisfactory measure of the extent to which the random variables X_1 and X_2 are dependent (the larger $|r|$, the "more dependent" X_1 and X_2).

16. Let X_1 and X_2 be the same as in Problem 14. Find the probability distribution of $Y = X_1 + X_2$.

1. This is not in the curriculum.

Table 2. Values of the normal distribution function $\Phi(x)$ given by formula (5.12).

x	$\Phi(x)$
0.0	0.5000
0.1	0.5398
0.2	0.5793
0.3	0.6179
0.4	0.6554
0.5	0.6915
0.6	0.7257
0.7	0.7580
0.8	0.7881
0.9	0.8159
1.0	0.8413
1.1	0.8643
1.2	0.8849
1.3	0.9032
1.4	0.9192
1.5	0.9332
1.6	0.9452
1.7	0.9554
1.8	0.9641
1.9	0.9713
2.0	0.9773
2.1	0.9821
2.2	0.9861
2.3	0.9893
2.4	0.9918
2.5	0.9938
2.6	0.9953
2.7	0.9965
2.8	0.9974
2.9	0.9981
3.0	0.9986

Chapter 6

More Limit theorems

Theorem 6.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_1^2) < \infty$. Then for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \mathbb{E}(X_1) \right| > \varepsilon \right) = 0 \,.$$

This result means that the "time average" $\frac{1}{n} \sum_{k=1}^{n} X_k$ is essentially equal to the "space average" $\mathbb{E}(X_1)$. We call $\frac{1}{n} \sum_{k=1}^{n} X_k$ a time average because we consider the mean over n observations $X_1, \dots, X_n$ at times $k = 1, \dots, n$. In contrast, $\mathbb{E}(X_1)$ is a mean over the sample space.

Proof. By Chebyshev, we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \mathbb{E}(X_1) \right| > \varepsilon \right) \le \frac{1}{\varepsilon^2} \mathbb{E}\left\{ \left(\frac{1}{n} \sum_{k=1}^{n} X_k - \mathbb{E}(X_1) \right)^2 \right\}.$$

On the other hand, $\mathbb{E}(\frac{1}{n} \sum_{k=1}^{n} X_k) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}(X_k) = \mathbb{E}(X_1)$. So we may replace the right-hand side by

$$\frac{1}{\varepsilon^2} \operatorname{Var}\left(\frac{1}{n} \sum_{k=1}^{n} X_k \right) = \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \sum_{k=1}^{n} \operatorname{Var}(X_k) = \frac{\operatorname{Var}(X_1)}{\varepsilon^2 n} \,,$$

where we used the X_j are i.i.d. As $n \to \infty$, this goes to zero, proving the theorem. □

Remark 6.2. We also have a "Strong Law of Large Numbers". This says that if $X_1, X_2, \dots$ is a sequence of i.i.d. random variables with $\mathbb{E}(|X_1|) < \infty$ then with probability one,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mathbb{E}(X_1) \,.$$

We do not prove this result in this course. It implies the above weak law.

Corollary 6.3. *For any $\varepsilon > 0$, the relative frequency of an event A satisfies*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{n(A)}{n} - \mathbb{P}(A)\right| > \varepsilon\right) = 0\,.$$

Proof. Call the occurrence of A a "success" and non-occurrence of A a "failure". Let

$$X_k = \begin{cases} 1 & \text{if k-th trial is a success,} \\ 0 & \text{if k-th trial is a failure,} \end{cases} \tag{6.1}$$

Then $\sum_{k=1}^{n} X_k$ is the number of times that A occurs in the n experiments, i.e. it is $n(A)$. Also, $\mathbb{E}(X_1) = 0\,\mathbb{P}(X_1 = 0) + 1\,\mathbb{P}(X_1 = 1) = \mathbb{P}(A)$. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{n(A)}{n} - \mathbb{P}(A)\right| > \varepsilon\right) = \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}(X_1)\right| > \varepsilon\right) = 0\,. \qquad \square$$

Definition 6.4. Let X be a discrete random variable with values in $\{0, 1, 2, \ldots\}$.

We define the *generating function* of X by

$$G_X(z) = \sum_{k\in \operatorname{Ran} X} P_X(k) z^k \qquad \text{for } z \in \mathbb{C}\,, |z| \le 1\,.$$

Example 6.5. If X has the Poisson distribution with parameter a then

$$G_X(z) = \sum_{k=0}^{\infty} \frac{a^k}{k!} \mathrm{e}^{-a} z^k = \mathrm{e}^{-a} \sum_{k=0}^{\infty} \frac{(az)^k}{k!} = \mathrm{e}^{a(z-1)}\,.$$

Lemma 6.6. (1) *X is uniquely determined by G_X. More precisely,*

$$P_X(k) = \frac{1}{k!} G_X^{(k)}(0)\,.$$

(2)

$$G_X(z) = \mathbb{E}(z^X)\,.$$

(3)

$$\mathbb{E}(X) = G_X'(1) \qquad \textit{and} \qquad \operatorname{Var}(X) = G_X''(1) + G_X'(1) - G_X'(1)^2\,.$$

(4) *If $X_1, \ldots, X_n$ are independent and $X = \sum_{k=1}^{n} X_k$ then*

$$G_X(z) = G_{X_1}(z) \cdots G_{X_n}(z)\,.$$

Proof. (1) We have $G_X(z) = P_X(0) + P_X(1)z + P_X(2)z^2 + \ldots$.

So $G_X'(z) = P_X(1) + 2P_X(2)z + 3P_X(3)z^2 + \ldots$.

In general $G_X^{(k)}(z) = k!P_X(k) + (k+1)!P_X(k+1)z + \frac{(k+2)!}{2}P_X(k+2)z^2 + \ldots$.

Hence, $G_X^{(k)}(0) = k!P_X(k)$.

(2) Apply the change of variables formula with $\varphi(t) = z^t$.

(3) $G'_X(z) = \sum_{k\geq 1} kP_X(k)z^{k-1}$ so $G'_X(1) = \sum_{k\geq 1} kP_X(k) = \mathbb{E}(X)$ as required.

$G''_X(z) = \sum_{k\geq 2} k(k-1)P_X(k)z^{k-2}$, so $G''_X(1) = \sum_{k\geq 0} k(k-1)P_X(k) = \mathbb{E}(X^2) - \mathbb{E}(X)$.

Thus, $G''_X(1) + G'_X(1) - G'_X(1)^2 = \mathbb{E}(X^2) - \mathbb{E}(X) + \mathbb{E}(X) - \mathbb{E}(X)^2 = \mathrm{Var}(X)$.

(4) Since $X_1, \ldots, X_n$ are independent then $z^{X_1}, \ldots, z^{X_n}$ are independent (admitted)[1]. Using (2) we get $G_X(z) = \mathbb{E}(z^{X_1}z^{X_2}\cdots z^{X_n}) = \mathbb{E}(z^{X_1})\,\mathbb{E}(z^{X_2})\cdots\mathbb{E}(z^{X_n}) = G_{X_1}(z)\cdots G_{X_n}(z)$. □

Example 6.7. If X has the binomial distribution with parameter n, then

$$G_X(z) = (pz + q)^n\,.$$

Indeed, $X = \sum_{k=1}^{n} X_k$ where X_k are defined in (6.1). As the trials are independent, the X_k are independent, so $G_X(z) = G_{X_1}(z)\cdots G_{X_n}(z)$ by the previous lemma. Finally, $G_{X_j}(z) = \sum_k P_{X_j}(k)z^k = P_{X_j}(0) + P_{X_j}(1)z = q + pz$ for any j. Thus, $G_Z(z) = (q + pz)^n$.

Definition 6.8. We define the *characteristic function of* X by

$$\varphi_X(t) = \mathbb{E}(\mathrm{e}^{\mathrm{i}Xt}) = \begin{cases} \sum_k P_X(k)\mathrm{e}^{\mathrm{i}kt} & \text{if X is discrete with values in } \{0,1,\ldots\} \\ \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}xt}p_X(x)\,\mathrm{d}x & \text{if X is continuous.} \end{cases}$$

Recall that if X is discrete, $G_X(z) = \mathbb{E}(z^X)$. In particular $G_X(\mathrm{e}^{\mathrm{i}t}) = \mathbb{E}(\mathrm{e}^{\mathrm{i}tX}) = \varphi_X(t)$.

Example 6.9. If X has the Poisson distribution then $\varphi_X(t) = G_X(\mathrm{e}^{\mathrm{i}t}) = \mathrm{e}^{a(\mathrm{e}^{\mathrm{i}t}-1)}$.

If X has the binomial distribution then $\varphi_X(t) = G_X(\mathrm{e}^{\mathrm{i}t}) = (q + p\mathrm{e}^{\mathrm{i}t})^n$.

Remark 6.10. Note that the characteristic function is essentially the *Fourier transform* of the density. As such, we can recover p_X from φ_X by the formula

$$p_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx}\varphi_X(t)\,\mathrm{d}t$$

as the student will learn in another course. We also have an inversion for discrete random variables :

$$P_X(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}tk}\varphi_X(t)\,\mathrm{d}t\,.$$

Indeed, we have $\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}tk}\varphi_X(t)\mathrm{d}t = \sum_n P_X(n)\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}(n-k)t}\,\mathrm{d}t = 2\pi P_X(k)$ because $\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}(n-k)t}\,\mathrm{d}t = \frac{\mathrm{e}^{\mathrm{i}(n-k)t}}{\mathrm{i}(n-k)}\big|_{-\pi}^{\pi} = \frac{2\sin(n-k)\pi}{n-k} = 0$ if $n \neq k$ and $\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}(n-k)t}\,\mathrm{d}t = 2\pi$ if $n = k$.

1. In general if X, Y are independent then $f(X)$ and $g(Y)$ are independent. The proof is actually easy but requires some mathematical language so we omit it.

Lemma 6.11. (1) *If X has the normal distribution $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ then*

$$\varphi_X(t) = e^{-t^2/2}.$$

(2)

$$\mathbb{E}(X) = -i\varphi_X'(0) \qquad \text{and} \qquad \mathrm{Var}(X) = -\varphi_X''(0) + \varphi_X'(0)^2.$$

(3) *If $X_1, \dots, X_n$ are independent and $X = \sum_{k=1}^n X_k$ then*

$$\varphi_X(t) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t).$$

Proof. (1) We have $\varphi_X'(t) = \int_{-\infty}^{\infty} ixe^{itx} e^{-x^2/2}\, dx$. Take $u = ie^{itx}$ and $dv = xe^{-x^2/2}$. Then

$$\varphi_X'(t) = -ie^{itx} e^{-x^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}(i^2 t e^{itx})\, dx = -t\varphi_X(t).$$

Moreover, $\varphi_X(0) = \mathbb{E}(1) = 1$. The unique solution to the differential equation $\varphi_X'(t) = -t\varphi_X(t)$ satisfying $\varphi_X(0) = 1$ is $\varphi_X(t) = e^{-t^2/2}$.
(2) $\varphi_X(t) = \mathbb{E}\big(\sum_{n=0}^{\infty} \frac{(iXt)^n}{n!}\big) = \mathbb{E}\left(1 + iXt + \frac{(iXt)^2}{2} + \sum_{k=3}^{\infty} \frac{(iXt)^n}{n!}\right)$.

Hence, $\varphi_X'(0) = \mathbb{E}(iX)$ and $\varphi_X''(0) = \mathbb{E}(-X^2)$. We easily deduce (2).
(3) Since $e^{itX_1}, \dots, e^{itX_n}$ are independent, $\varphi_X(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(e^{itX_1} \cdots e^{itX_n}) = \mathbb{E}(e^{itX_1}) \cdots \mathbb{E}(e^{itX_n}) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t)$. □

Theorem 6.12 (The Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with finite mean and variance σ^2. Define $S_n = \sum_{k=1}^n X_k$ and $S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{n\sigma^2}}$. Then*

$$\lim_{n\to\infty} \mathbb{P}(a \le S_n^* \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx. \tag{6.2}$$

Note that by the Law of Large Numbers, $\lim_{n\to\infty} \frac{S_n - \mathbb{E}(S_n)}{n} = 0$. The Central Limit Theorem tells us more precisely that $S_n - \mathbb{E}(S_n) \sim \sqrt{n\sigma^2}$ in distribution, which gives us the speed of convergence.

The important point is that it gives a "universal" result : whatever the distributions of X_k, the fluctuations $\frac{S_n - \mathbb{E}(S_n)}{\sqrt{n\sigma^2}}$ converge to the normal distribution. This implies that the statistical methods that work for the normal distribution can also be applied to problems with a non-normal data, which is useful to practitioners. The De Moivre-Laplace theorem is the special case $X_k = 0$ or 1.

Proof of Theorem 6.12. Let $Y_j = X_j - \mathbb{E}(X_j)$. Then $\mathbb{E}(Y_j) = 0$ and $\mathbb{E}(Y_j^2) = \mathrm{Var}(X_j) = \sigma^2$. Hence,

$$\varphi_{Y_j}(t) = \mathbb{E}(e^{iY_j t}) = \mathbb{E}\left(1 + iY_j t + \frac{(iY_j t)^2}{2}\right) + t^2\varepsilon(t) = 1 - \frac{t^2\sigma^2}{2} + t^2\varepsilon(t),$$

where $\varepsilon(t)$ goes to zero [2] as $t \to 0$. Now $\varphi_{S_n^*}(t) = \mathbb{E}(e^{iS_n^* t}) = \mathbb{E}(e^{i(Y_1+\cdots+Y_n)\frac{t}{\sqrt{n\sigma^2}}}) = \varphi_{Y_1}(\frac{t}{\sqrt{n\sigma^2}}) \cdots \varphi_{Y_n}(\frac{t}{\sqrt{n\sigma^2}})$. Since the Y_j are identically distributed they have the same characteristic function (apply the change of variables formula). We thus showed that

$$\varphi_{S_n^*}(t) = \Big(\varphi_{Y_1}\Big(\frac{t}{\sqrt{n\sigma^2}}\Big)\Big)^n = \Big(1 - \frac{t^2}{2n} + \frac{t^2}{n\sigma^2}\varepsilon\Big(\frac{t}{\sqrt{n\sigma^2}}\Big)\Big)^n .$$

Apply $|u^n - v^n| = |\sum_{k=1}^n u^{n-k}(u-v)v^{k-1}| \leq n|u-v| \cdot \max(|u|,|v|)^{n-1}$ to $u = \varphi_{Y_1}(\frac{t}{\sqrt{n\sigma^2}})$ and $v = 1 - \frac{t^2}{2n}$. Note that here $|v| \leq 1$ and $|u| \leq \mathbb{E}\left(\left|e^{iY_1\frac{t}{\sqrt{n\sigma^2}}}\right|\right) = 1$. We thus get

$$\Big|\varphi_{S_n^*}(t) - \Big(1 - \frac{t^2}{2n}\Big)^n\Big| \leq n\left|\frac{t^2}{n\sigma^2}\varepsilon\Big(\frac{t}{\sqrt{n\sigma^2}}\Big)\right| \to 0$$

as $n \to \infty$. But $|(1 - \frac{t^2}{2n})^n - e^{-t^2/2}| \to 0$ as $n \to \infty$. So by the triangle inequality we deduce that $|\varphi_{S_n^*}(t) - e^{-t^2/2}| \to 0$. In other words, $\lim_{n\to\infty} \varphi_{S_n^*}(t) = e^{-t^2/2}$. Using Lemma 6.11(1), we conclude that for any t,

$$\lim_{n\to\infty} \varphi_{S_n^*}(t) = \varphi_X(t)\,, \tag{6.3}$$

where X is a random variable with the normal distribution.

Using *Levy's continuity theorem* (admitted), (6.3) implies (6.2). □

Example 6.13. A teacher has 90 exams to correct. The times required to correct the exams are independent with the same distribution, having mean 10 minutes and variance 4 minutes. Using the central limit theorem, approximate the probability that the teacher will grade at least 40 exams in the first 380 minutes of work.

Solution. Let X_k be the time (in minutes) to grade exam k. Then $S = \sum_{k=1}^{40} X_k$ is the time needed to grade the first 40 exams. So we seek $\mathbb{P}(S \leq 380)$. Here $\mathbb{E}(S) = 400$ and $\sqrt{n\sigma^2} = \sqrt{160} \approx 12.65$. By the Central Limit Theorem, $\mathbb{P}(S \leq 380) = \mathbb{P}(\frac{S-400}{12.65} \leq \frac{-20}{12.65}) \approx \mathbb{P}(-\infty < \frac{S-400}{12.65} \leq -1.58) \approx \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-1.58} e^{-x^2/2}\,dx = \Phi(-1.58) = 1 - \Phi(1.58) \approx 1 - 0.945 = 0.055$.

2. This follows from the Taylor-Lagrange expansion of the exponential near zero. The student who doesn't know it can instead apply l'Hôpital twice to see that $\varepsilon(t) = \mathbb{E}(\frac{e^{iY_j t} - 1 - iY_j t - \frac{(iY_j t)^2}{2}}{t^2}) \to 0$ as $t \to 0$.

6.1 Exercises

1. Show that the conclusion of the Law of Large Numbers can be written as follows : for any $\delta > 0$ and $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$,

$$a - \varepsilon \leq \frac{1}{n}(X_1 + \cdots + X_n) \leq a + \varepsilon$$

with probability greater than $1 - \delta$.

2. Let $X_1, \ldots, X_n$ be n i.i.d. random variables, with common mean $a = \mathbb{E}\, X_1$ and variance $\sigma^2 = \operatorname{Var} X_1$. Suppose a is known. Can the quantity

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - a)^2$$

be used to estimate σ^2 ?

3. A random variable X has probability density

$$p_X(x) = \begin{cases} \frac{x^m}{m!}\mathrm{e}^{-x} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where m is a positive integer. Prove that

$$\mathbb{P}(0 \leq X \leq 2(m+1)) > \frac{m}{m+1}$$

Hint. Use Chebyshev's inequality.

4. The probability of an event A occurring in one trial is $\frac{1}{2}$. Is it true that the probability of A occurring between 400 and 600 times in 1000 independent trials exceeds 0.97 ?

5. Let X be the number of spots obtained in throwing an unbiased die. What is the generating function of X ?

6. Use Lemma 6.6(3) and the result of the preceding problem to solve Problem 16 of Chapter 4

7. Let X be a random variable with the Poisson distribution with parameter a. Use Lemma 6.6(3) to show that $\mathbb{E}(X) = \operatorname{Var}(X) = a$.

8. Find the generating function of the random variable X with distribution

$$\mathbb{P}(X = k) = \frac{a^k}{(1+a)^{k+1}} \qquad (a > 0)\,.$$

Use Lemma 6.6(3) to find $\mathbb{E}(X)$ and $\operatorname{Var}(X)$.

9. Let Y be the sum of two independent Poisson random variables X_1 and X_2 with parameters a and a', respectively. Show that Y also has a Poisson distribution, with parameter $a + a'$.

10. Let S_n be the number of successes in a series of n independent trials, where the probability of success at the k-th trial is p_k. Suppose $p_1, \dots, p_n$ depend on n in such a way that

$$p_1 + \cdots + p_n = \lambda$$

while

$$\max\{p_1, \dots, p_n\} \to 0$$

as $n \to \infty$. Using [3, Theorem 6.2], prove that S_n has a Poisson distribution with parameter λ in the limit as $n \to \infty$.

11. Find the characteristic function $\varphi_X(t)$ of the random variable with probability density

$$p_X(x) = \frac{1}{2}\mathrm{e}^{-|x|} \qquad (-\infty < x < \infty)\,.$$

12. Use Lemma 6.11(2) and the result of the preceding problem to solve Problem 13 of Chapter 4.

13. Find the characteristic function of a random variable uniformly distributed in the interval $[a, b]$.

14. A continuous random variable X has characteristic function

$$\varphi_X(t) = \mathrm{e}^{-a|t|} \qquad (a > 0).$$

Find the probability density of X.

15. The derivatives $\varphi'_X(0)$ and $\varphi''_X(0)$ do not exist in the preceding problem. Why does this make sense ?

Hint. See Problem 24 of Chapter 4.

16. Let ν be the total number of spots which are obtained in 1000 independent throws of an unbiased die. Then $\mathbb{E}\,\nu = 3500$, because of Problem 16 of Chapter 4. Estimate the probability that ν is a number between 3450 and 3550.

17. Let S_n be the same as in Problem 10, and suppose $\sum_{k=1}^{\infty} p_k q_k = \infty$. Prove that

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - \sum_{k=1}^{n} p_k}{\sqrt{\sum_{k=1}^{n} p_k q_k}} \le b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x$$

using [3, Theorem 6.3] (which generalizes Theorem 6.12 to random variables with distinct distributions. It is called the Lyapunov CLT).

Epilogue : Missing proofs and further reading

Throughout the course we have tried to maintain a good level of rigor without getting lost in abstraction (as there is no time). We hope this will help the student when studying more advanced topics in probability.

1. A Fun reading. As a parallel and more leisurely approach, we recommend the book [1] for the student to experience more fun with probabilities. This makes a good vacation reading.

2. More rigor ? We now discuss some missing proofs and further reading. Everything below is beyond the level of 2nd year students, but this discussion may be useful if they read it after becoming more mathematically mature.

Concerning Chapter 2, the proper language for probability spaces is *measure theory*. The set of events $\mathcal{F}$ forms what we call a *σ-algebra*. We have discussed the easy but fundamental case where Ω is a finite set and $\mathcal{F} = \mathcal{P}(\Omega)$ is the power set of Ω. We have also discussed $\Omega = \mathbb{R}$ while introducing distributions P_X in Chapter 4. In that chapter however we did not mention what is the family of events $\mathcal{F}$ which should correspond to $\mathbb{R}$. It turns out that we cannot take the power set of $\mathbb{R}$ anymore, the family $\mathcal{P}(\mathbb{R})$ contains too many sets so we have to discard some bizarre sets to which we cannot assign a probability. This said, any set the student can imagine will probably lie in the good family of events, so while we do not take all sets, i.e. do not take $\mathcal{F} = \mathcal{P}(\mathbb{R})$, we do take "essentially all subsets". In particular, the family $\mathcal{F}$ contains all intervals, and it is actually sufficient to understand the probability measure P_X on such intervals.

We have briefly studied Bayes' law in Chapter 3. This actually opens up a whole theory of Bayesian Statistics which is interesting from both a theoretical and applied point of view. The interested student can perhaps start by reading the Wikipedia entry on this topic and proceed from the references there.

In Chapter 4 we emphasized the interplay between two probability spaces : $(\Omega, \mathbb{P})$ and $(\mathbb{R}, P_X)$. The main events which interested us were of the form $\{X \in I\}$, which we defined by $\{\omega \in \Omega : X(\omega) \in I\}$. This is sometimes denoted by $X^{-1}(I)$ and called "the inverse image of I by X", to emphasize that $I \subseteq \mathbb{R}$ but $\{X \in I\} \subseteq \Omega$.

We call

$$m_k := \mathbb{E}(X^k)$$

the *k-th moment of X*. We also call

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

the *covariance of X and Y*. Note that the correlation coefficient may then be summarized as $r = \frac{\mathrm{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$. Clearly $\mathrm{cov}(X, X) = \mathrm{Var}(X)$ and if X, Y are independent then $\mathrm{cov}(X, Y) = 0$.

We admitted some results in Chapter 4 without proof : the change of variables formula, some basic properties of expectation, the Cauchy-Schwarz inequality. None of these results is difficult to prove, some are actually true "by definition", but the language itself requires measure theory. The student will learn this in another course, but can already check [4, Appendix A] for an excellent but condensed treatment. See in particular Section A.5 and Section A.7 in that book for the properties of $\mathbb{E}$ and the change of variables, respectively. The student will learn Cauchy-Schwarz's inequality in the course of *Functional Analysis*, but this requires first to understand $L^2(\mathbb{R})$ and show that it is an inner product space. For this, see Section 0.3 and Section 0.6 in [4].

We omitted the proof of the De Moivre-Laplace theorem in Chapter 5, as it is a special case of the Central Limit Theorem. However, the book [3] presents an independent proof in Theorem 5.1 p. 59. The advantage perhaps is that it does not use characteristic functions and Lévy's continuity. The argument may seem a bit obscure so let us explain the main idea. The theorem wants to assess the probability that $S_n^* \in [a, b]$. Since the X_k are discrete, so is S_n^*, so this probability can be expanded as a sum $\sum_{\alpha_k(n) \in [a,b]} \mathbb{P}(S_n^* = \alpha_k(n))$. Moreover, the probabilities here are governed by a "shifted" Binomial distribution by definition of X_k and S_n. More precisely, as the X_k take the values $0, 1$, the S_n take the values $\{k\}_{k=0}^n$ and so S_n^* takes the values $\{\frac{k-np}{\sqrt{npq}}\}_{k=0}^n$, with probability $C_k^n p^k q^{n-k}$. So let $\alpha_k(n) = \frac{k-np}{\sqrt{npq}}$. The proof then goes by first establishing a *local result*, namely $\mathbb{P}(S_n^* = \alpha_k(n)) \approx \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\alpha_k(n)^2/2}}{\sqrt{npq}}$. Since $\alpha_k(n) - \alpha_{k-1}(n) = \frac{1}{\sqrt{npq}}$, this means

that $\sum_{\alpha_k(n)\in[a,b]} \mathbb{P}(S_n^* = \alpha_k(n)) \approx \sum_{\alpha_k(n)\in[a,b]} \frac{e^{-\alpha_k(n)^2/2}}{\sqrt{2\pi}}(\alpha_k(n) - \alpha_{k-1}(n))$, which is the Riemann sum for the Gaussian integral $\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$, so we are done by taking $n \to \infty$. So the major part of the proof is to check the local result, and this is done by using Stirling's approximation in the binomial expression.

Still in Chapter 5, we omitted a pleasant introduction to *Brownian motion* which Rozanov gives on p.63-65. The main question is the following : suppose a particle makes a "random walk" on the real line by jumping at times $n\Delta t$, $n = 0, 1, 2 \ldots$. Under some natural assumptions on this motion, let X_t be the position of the particle at time t. Then we ask : what is the distribution of X_t as $n \to \infty$? The De Moivre-Laplace theorem is used to show that this limiting distribution is the normal distribution. This random particle motion is a mathematical idealization of a particle suspended in a fluid, undergoing random collisions (this is certainly not one-dimensional). This first result opens the way to the mathematical study of Brownian motion, which is beyond undergraduate level.

Concerning Chapter 6, we proved the weak Law of Large Numbers (LLN) and admitted the strong one. A proof of the strong law can be found e.g. in [2, Section 5.3].

A random variable having the normal distribution is often called a normal random variable, and denoted by $\mathcal{N}(0,1)$. A random variable with the more general probability density $\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-a)^2}{2\sigma^2}}$ is denoted by $\mathcal{N}(a, \sigma^2)$.

The Central Limit Theorem (CLT) can be summarized as follows : the S_n^* *converge in distribution* to $\mathcal{N}(0,1)$.

Next come generating functions and characteristic functions. A result we skipped is that if X_n, X are discrete, then the X_n converge in distribution to X iff $G_{X_n}(z) \to G_X(z)$ uniformly in every disk $|z| \leq r < 1$. This is not difficult and can be found in [3, Theorem 6.2]. This can be used for example to give another proof of the Poisson Limit Theorem.

An analogous result holds for characteristic functions but the proof is much longer - this is known as *Lévy's continuity theorem*. This theorem says that if X is continuous at 0, then X_n converges to X in distribution iff for all t, $\varphi_{X_n}(t) \to \varphi_X(t)$. The fact that convergence in distribution implies convergence of φ_{X_n} is easy. The other direction is the hard part, and it is this hard direction that we use in the proof of the Central Limit Theorem. The argument is essentially the following : first of all, the family of functions $\{x \mapsto e^{ixt}\}_{t\in\mathbb{R}}$ is

sufficiently big to approximate any continuous function (just like polynomials $\{x \mapsto P_n(x)\}_{n\geq 0}$, which is Weierstrass's approximation theorem). The second argument is that the convergence of φ_{X_n} to φ_X guarantees that all X_n somehow live in the same region, i.e. their range does not escape to infinity. These two arguments are used to show that if $\mathbb{E}(e^{iX_n t}) = \varphi_{X_n}(t) \to \varphi_X(t) = \mathbb{E}(e^{iXt})$ for all $t \in \mathbb{R}$, then $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all continuous functions f, and this implies convergence in distribution. The detailed proof of Lévy's continuity theorem can be found in [2, Section 15.3].

Let us mention in passing that there are several generalizations of the CLT. For example, Lyapunov's version in [3] allows for random variables with different distributions.

3. What's next ? The two previous points give some sketches and references as to how to make this course more fun or more detailed. From the present basics of Probability Theory, the student can either follow a course on Statistics (this has concrete applications in real-life) or a course on Stochastic Processes (this requires more theoretical foundations, and also has important applications).

Bibliography

[1] Richard Isaac. *The pleasures of probability*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1995. Readings in Mathematics.

[2] Achim Klenke. *Probability theory. A comprehensive course*. Universitext. Springer, London, second edition, 2014.

[3] Y. A. Rozanov. *Probability theory: A concise course*. Dover Publications, Inc., New York, english edition, 1977. Translated from the Russian and edited by Richard A. Silverman.

[4] Gerald Teschl. *Mathematical methods in quantum mechanics*, volume 157 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2014. With applications to Schrödinger operators.